Assessing the strength of non-contemporaneous forensic speech evidence

Richard William Rhodes

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

The University of York

Department of Language and Linguistic Science

Submitted December 2012

Abstract

The aim of this thesis is to assess the impact of long term non-contemporaneity on the strength of forensic speech evidence. Speakers experience age-related changes to the voice over long delays and this time also presents the opportunity for social factors to vary. These changes are shown to impact on speech parameters used in forensic analyses.

Using longitudinal data from the *Up* documentary series, this thesis analyses the effects of aging on forensically useful acoustic parameters in eight speakers at five seven-year intervals between ages 21 and 49. The investigation reveals significant age-related changes in real-time across adulthood. Frequencies of the first three formants in monophthongs /i: ɪ e a ɑ: ʌ ɒ ʊ & u:/ and diphthongs /eɪ & aɪ/ show comprehensive reduction. For monophthongs, F1 exhibits mean change of 8.5%, greater than F2 and F3 at 3.7% and 2.2% respectively. Vowel quality also impacts on magnitude of change in each formant. Estimations based on this data suggest that vocal tract extension and restricted articulator movement are probable drivers for acoustic change, operating on different timelines. Counter-examples to this aging pattern can generally be explained by social factors, as a result of mobility or in accordance with mainstream changes in a variety.

Strength of evidence estimates for these non-contemporaneous data are calculated using a numerical likelihood ratio (LR) approach. Age-related changes result in weaker and fewer correct LRs with greater length delays. Cubic coefficients of diphthong formants are investigated in line with a formant dynamic approach. These LR tests show promising results and resilience to aging, especially in F1; tentatively suggesting that, for these speakers, some speaker-specific behaviour pervades in spite of physiological changes. This analysis raises several questions with regards to applying an overtly numerical LR approach where there is apparent mismatch between forensic samples.

The effect of aging on an ASR system (*BATVOX*) is also tested for six male subjects. The system measures Mel-frequency cepstral coefficient (MFCC) parameters that reflect the physical properties of the vocal tract. Predicted degradation of the system's performance with increasing age is apparent. The reduction in performance is significant, varies between speakers, and is striking in longer delays for all speakers.

The degradation in strength of evidence for acoustic data from monophthongs and formant dynamic coefficients, as well as that for the ASR system, demonstrates that aging presents a real problem for forensic analysis in non-contemporaneous cases. Furthermore, aging also presents issues for speech databases for the purpose of assessing strength of evidence, where further research into distributions of parameters in different age groups is warranted.

Contents

List of Figures

List of Tables

List of Abbreviations

AR	Articulation rate
ASR	Automatic speaker recognition
AT	Apparent time
C_{llr}	Log-likelihood-ratio cost
EER	Equal error rate
F0	Fundamental frequency
F1-4	Formants (i.e. F1 = first formant)
FD	Formant dynamics
FSC	Forensic speaker comparison
GMM-UBM	Gaussian mixture model – universal background model
Log_{10}	Logarithmic scale with base 10
LR	Likelihood ratio
LLR	Log likelihood ratio (base 10 in this thesis)
LTAS	Long-term average spectra/formant
LT(A)F	Long-term (average) formant
MFCC	Mel-frequency cepstral coefficient(s)
MVKD	Multivariate kernel density
NHSCR	National Health Service Central Register
RP	Received pronunciation
RT	Real time
SES	Socio-economic status
SNR	Signal-to-noise ratio
SSBE	Standard southern British English
StFi	Standard Finnish
VSA(e)	Vowel space area (estimate)
VTL(e)	Vocal tract length (estimate)
VQ	Voice quality

Acknowledgements

Thanks to Paul Foulkes for providing the opportunity to do this, his supervision and for wading through a sea of semi-colons.

Dominic Watt and Peter French gave me valuable thought and criticism throughout. I am grateful to Phil Harrison for scripts and technical help, and also to Frantz Clermont for assistance with estimation formulae. Other students and staff in the department and at J P French Associates make York such an active and interesting place to study forensic phonetics. Special thanks to Colleen Kavanagh for day-to-day desk sharing support.

Thanks to my family for pushing me to get this far, especially to Catie for encouragement, giving me something to aim for and introducing me to coffee.

Declaration

This is to certify that this thesis comprises original work toward the degree of doctor of philosophy, that due acknowledgement has been made to all other materials used in the text, and that it is less than 80,000 words in total (excluding references).

Richard William Rhodes

1 Introduction

There are three main aims of the current study. The first is to investigate the effects of aging on vocal parameters. This analysis relates principally to forensic situations where there is a long delay between recordings and to any population database where aging may cause overall differences in frequency patterns. Secondly, it investigates dynamic measures of diphthong formants to ascertain their performance, relative to other measures commonly employed in FSC and whether they have any advantage in cases of non-contemporaneity. Thirdly, the study examines how experts express strength of evidence in court: specifically the effects of age mismatch on likelihood ratio (LR) calculations.

This chapter places the present project into the context of forensic speech science and discusses some of the practical requirements of casework using speech evidence.

1.1 Forensic Speech Science

The primary motivation for this project is to extend our understanding of vocal aging which may have implications for how we make decisions in forensic speech analysis. To understand the context of this research, it is useful to present a short summary of the field and its demands, and some of the restrictions and desirable qualities that are relevant to the types of speech materials that are regularly used in a forensic exercise.

1.1.1 Forensic speaker comparison

As the most prevalent application of forensic speech science, this thesis focuses on forensic speaker comparison (henceforth FSC), as defined in works such as Nolan (1983), Rose (2002) and Jessen (2008). This is not to say that the current research project is only relevant to FSC, as a better understanding of non-contemporaneous speech is vital for voice profiling, voice line-ups and reference database selection.

This thesis principally addresses forensic speaker comparison in relation to the combined (Rose, 2002; French, Nolan, Foulkes, Harrison, & McDougall, 2010) or interactive (Nolan & Grigoras, 2005) auditory and acoustic approach (also termed a 'two-vector' approach (Künzel, 2009)). Other analysis methods, such as automatic speaker recognition (ASR) systems, are used globally in a forensic setting and have been shown to perform well in experimental settings (Alexander, 2007; Künzel, 2009; Harrison & French, 2010). This

study secondarily examines the performance of ASR in the face of aging in comparison with acoustic data.

In general in FSC there are two samples of speech, the first being drawn from one or more evidential recordings: often referred to as a disputed, questioned or trace samples. In this thesis the term 'evidential' sample is preferred. This sample is most frequently taken from telephone recordings or (in some cases covert) surveillance of the suspect(s) (Nolan, 1983; Rose, 2002). The second sample is extracted from a recording of a suspect, and is normally referred to as a known or reference sample. The term 'suspect' sample is used throughout the thesis. In the UK these samples are normally taken from recorded police (PACE) interviews (Nolan, 1983; Rose, 2002), although it is not unheard of for bespoke recordings to be collected, or analysis to be made of archive recordings (especially in civil cases).

Within the combined auditory-acoustic model, the analyst uses both auditory and acoustic analyses to make decisions on how consistent the samples are and to what extent the phonetic-acoustic and linguistic features analysed are distinctive enough to separate a speaker from the broader population. There are different methods for expressing the strength of this evidence and although the two processes naturally inform and shape each other, this is discussed separately from speech analysis (see §2.3). Generally within any conclusion framework, FSC is principally two comparisons: one between two samples to determine consistency or similarity, and secondly with respect to the wider population to determine distinctiveness or typicality.

1.1.2 **The speaker, the voice and speech evidence**

It is worth making a distinction between some concepts at this point, especially when they are central to the concept of forensic comparison. In this thesis, a 'speaker' refers to a person who produces a speech sample, 'voice' refers to the vocal equipment with which they produce speech, and 'speech' itself refers to linguistic sounds produced by speakers through the voice. So in this sense, the speaker is an individual and speech is an artefact of a process that goes on within that speaker involving mental and physical procedures. To clarify the distinction it is useful to extend a metaphor in Robertson and Vignaux (1995) to speech, where a gun represents the speaker and the barrel (the voice) leaves individual marks on a bullet (speech):

> From the moment it is manufactured each barrel leads a different life. The number of times it is fired, its exposure to the elements, how it is cleaned and any accidental damage will all affect the marks that the barrel leaves on bullets in a unique and unreproducable way (Robertson & Vignaux, 1995, p. 152)

In the case of speech articulation, this metaphor is relatively simplistic as there are many more dimensions of variance as a result of the inherent flexibility of the articulators and their movement than found in a gun barrel.

In any forensic exercise, and particularly with respect to speech, it is not possible to make 100% confidence statements about identification, so terms like speaker 'identification' or 'individualisation' are now generally avoided throughout the field. It might also be sensible (if a little pedantic) to prefer the term 'speech comparison' or 'voice comparison' (Rose & Morrison, 2009) to 'speaker comparison', when in fact analysis of speech sounds and systematic study of voices is what actually takes place, rather than analysis of a speaker or person. French et al. (2010), however, suggest that speaker comparison (or speech comparison) is preferable to 'voice comparison' as some features are not products of the voice. In any case, forensic speaker comparison is a well-established term which is used interchangeably (FSC) with forensic speech comparison.

1.1.3 **The forensic condition**

The types of speech features that prove effective in FSC often depend on the conditions in which the recordings are made. It is important to consider these conditions before examining auditory and acoustic speech features for forensic purposes. In most cases, the quality of recordings is far from optimal, for a variety of reasons (examples are given below). Moreover, there is generally a significant technical mismatch between the evidential and suspect samples. This may be due to recording or channel mismatches, which are summarised below. To combat this, the analyst must have an understanding of the factors that vary between samples and be able to predict their effects on the evidential recordings. These technical and channel considerations also influence which features an analyst can utilise.

1.1.3.1 Speaker effects

There are many possible speaker effects which might impact on a forensic recording. Speaker effects in this context are changes in articulation which are made by the speaker, whether consciously or not.

Disguising the voice is one way a speaker might directly affect forensic parameters. However, disguise is reported in fewer than 5% of UK cases (French, personal communication, reported in Clark & Foulkes (2007)) and up to 25% of cases in Germany (Künzel, 2000) (probably due to a higher number of kidnap and defamation cases). Although Nolan (1983) reports that accent disguises are difficult to maintain, there are other forms of disguise which may present problems, such as oral obstruction with a pen/pencil or other object or speaking with a different voice quality or laryngeal setting. New technologies also present easy methods for electronic disguise (cf. Clark and Foulkes (2007)) which may be easily detectable, but hard to reverse in some cases. One of the most common factors is probably speaking style (Jessen, Köster, & Gfroerer, 2005), which can vary massively according to a number of social, environmental and individual factors. Other speaker factors can include stress (Jessen, Köster, & Gfroerer, 2005; Kirchhübel, Howard, & Stedmon, 2011) and extreme distress (Roberts, 2011), intoxication (Chin & Pisoni, 1997; Barfüsser & Schiel, 2010; Baumeister & Schiel, 2010) including drug addiction (Papp, 2009), multilingualism/L2 speakers or comparison between languages (Sullivan & Schlichting, 2000), to name a few. The analyst must use all available information about such factors to make predictions about the case, even to exclude certain features in light of these predictions.

In general any combination of these factors may be present, and the analyst is expected to be able to make sensible predictions about the potential for a voice to behave in certain ways in certain circumstances. For example someone speaking using raised voice in one sample and modal voicing in another is likely to exhibit very different F0 measures, however, this may not lead the analyst to conclude that the two are inconsistent.

1.1.3.2 Channel effects

A widespread problem in a majority of cases, both in the UK and abroad, is recording and transmission of speech. Telephone bandwidth limitations generally prevent the analysis of higher frequency information, which has been shown to perform well in discriminating between speakers (Nolan, 1983; Loakes, 2004). Furthermore, bandwidth-based attenuation of lower formant frequencies creates a false weighting 'telephone effect' which has been shown to increase estimates of F1 values by an average of 13-14% for [German] landlines, (Künzel, 2001) and 29% for mobile phones (Byrne & Foulkes, 2004). F2 and F3 were not significantly affected in these studies. Subsequently analysts have

only examined F1 with great caution, or in cases without telephone transmission. There are also a number of questions about new technologies and transmission effects, such as Voice over Internet Protocol (VOIP) transmissions (such as *Skype*).

Furthermore there is generally a significant technical mismatch in the way forensic recordings are made, with the majority of suspect samples being recorded on audio cassette (though more and more digital recorders are being introduced). In some cases samples are filtered to emulate the transmission and technical quality of the other recording, cancelling out these differences in order to carry out auditory analysis. It is not feasible to use this to great effect in acoustic analysis, however.

There are also more general concerns with recording equipment differences, incorporating changing microphone-to-source distance (Vermeulen, 2009), cases where recording devices are partly concealed and different materials are shown to have different acoustic filtering effects (Watt, Llamas, & Harrison, 2010; Fecher, 2011). In general there are a number of concerns which may lead an analyst to make a bespoke testing sample with the equipment in question, or at least make predictions about the effects on forensic data based on existing research.

1.1.3.3 Situational effects

Situation effects refer to the environment in which recording or transmission occurred, which might impact on either a speaker or a recording/transmission method.

Background noise can have two main effects on the forensic recording. First and most commonly, is the obfuscating effect on the speech data in questions from overlapping sounds. Where these sounds are predictable (such as known music) or periodic, there are limited ways of removing these from the recording. These procedures are documented for GSM interference (Harrison, 2001) and music (Alexander & Forth, 2011), for instance. However, this is never wholly reliable, especially when speech and noise occupy the same area of the frequency spectrum. The second effect is a situation-related speaker effect, where background noise (or indeed the mode of transmission, i.e. using a mobile phone) may cause a speaker to speak differently. Elevated speaking style due to background noise is known as the 'Lombard reflex' (summarised in a forensic context in French (1998) and Jessen et al. (2005)). Its most frequently observed effects include increases in

speaking rate, fundamental frequency and in some cases (as with the telephone effect) the first formant.

Non-contemporaneity is arguably neither a speaker effect nor a channel effect, but is an integral part of the forensic condition. Rose (2002) states that, logically, no evidential and suspect recordings are made contemporaneously. Naturally the magnitude of delay is different in every case, but in the UK analysts can expect there to be a typical delay of somewhere between 2-5 months (French, personal communication). This delay is most likely due to processing of evidence and the course of the investigation. Police reportage normally reveals the extent of this delay, giving the analyst an idea of the level of non-contemporaneity they should expect. There are some cases where delays are relatively short, however, and also cases of extreme long-term non-contemporaneity, which are of specific interest to the present study (discussed in later sections, see §2.1.2.9).

1.1.4 **Desirable criteria for FSC parameters**

There is an ever-expanding body of research on which features have desirable characteristics for FSC, and this is based on a set of criteria made in response to the task and those conditions outlined above. A very useful set of criteria is outlined by Nolan (1983):

i. *High inter-speaker variability*: Any parameter should show a high degree of variation between speakers
ii. *Low intra-speaker variability*: Any parameter should show consistency through the speech of one speaker
iii. *Resistance to disguise (or mimicry)*: Any parameter must withstand attempts on the part of the speaker to disguise his voice
iv. *Availability*: Any parameter should produce large amounts of test data in both reference [suspect] and evidence [evidential] samples
v. *Robustness in transmission*: Any parameter must withstand recording processes and telephone transmission
vi. *Measurability*: The extraction of the parameter must not be prohibitively difficult

The final four criteria are concerned more with practical casework requirements, while the first two capture the real crux of the forensic task: to examine features which vary between speakers but remain consistent within a speaker. For example, in terms of inter-speaker variability, mean F0 performs poorly (i.e. 60% of SSBE speakers have average F0 within a 20Hz range, cf. Hudson et al. (2007)). It is also likely to be highly variable within a speaker (due to a number of factors, such as emotional state in police interviews

compared with in evidential recordings, cf. Jessen et al. (2005)). Thus, mean F0 is not an ideal FSC parameter.

1.1.5 **Summary of features used in FSC**

This section cannot provide an extensive summary of research into forensic parameters. However, it presents statistics on which features are used commonly and summarises those that form the basis for analysis in the present study. There are numerous studies presented in the *International Journal of Speech, Language and the Law* for further reference.

Gold and French (2011) carried out a survey of 36 forensic speech practitioners from 13 countries in 5 continents in order to present statistics on the types of analysis that are widely used and their frequency, along with information on parameters, their confidence in those parameters, conclusion frameworks and a number of other details. This paper gives an excellent snapshot of the field at the time of publishing. A summary is presented in the table below.

Table 1 - Summary of features used and their frequency in FSC exercises

Type of Feature	Aspect of feature	Specific parameter	Analysts using feature
Segmental			
Vocalic			100% (81% invariably, 13% routinely)
	Auditory quality		94%
	Durations		58%
	Formant analysis		97%
		F1	87% (of those using formants)
		F2	100% (of those using formants)
		F3	87% (of those using formants)
		F4	17% (of those using formants)
Consonantal			100% (52% invariably)
	Auditory quality		88%
	Timing		82%
	Energy loci frequency		48%
Suprasegmental			
	F0		100%
		Mean F0	94%
		SD	72%
		Alternative Baseline	25%
	Voice Quality		94% (77% invariably or routinely)
	Intonation		85% (25% invariably)
	Tempo		93%
	Rhythm		73% (variably)
Linguistic			
	Discourse markers		76%
	Conversational behaviour		76%
	Lexical-grammatical		88%
non-Linguistic			
	Various (i.e. clicks)		94%

Data source: Gold & French (2011)

A proportion of respondents were also using ASR systems alongside their analysis. Participants were also asked to rate which features were the best speaker discriminants in their experience. The features, ranked along with the proportion of analysts who rated them highly, are below:

- 'Voice quality' (32%)
- 'Dialect or accent variants' & 'vowel formants' (28%)
- 'Fundamental frequency' & 'tempo' (20%)
- 'Rhythm' (16%)
- 'Lexical-grammtical choices', 'vowel/consonantal realisations', 'phonological processes' & 'fluency' (13%)

Participants and the authors noted that although individual features can be highly discriminatory, it is overwhelmingly the case that the overall interpretation of a number of features that leads to a conclusion.

Although voice quality is stated as being one of the most utilised and best performing parameters, there is very little research into population distributions and speaker discrimination methods (although cf. Stevens and French (2012)). There is however, a plethora of studies on the usefulness of formants (including formant dynamics), fundamental frequency, cepstral coefficients and a number of other features in discriminating between speakers. The present study characterises the effects of long-term non-contemporaneity on a number of these features which are popularly used in FSC analyses.

1.1.6 **Applications**

Broad applications of this research are outlined in this section. The most obvious benefit of a model of vocal aging in a forensic context relates to FSC cases where there is a long-term delay; an example of a case like this is in the case of the Yorkshire Ripper Hoaxer (R v John Samuel Humble, 2005) where there was a 27 year delay between evidential and suspect samples. The findings of this study are also useful in other forensic exercises where there is a long delay. An example might be a speaker profiling case where there is apparent mobility or a reference recording which is not recent.

There is also an indirect benefit to investigating age-correlated changes to frequency characteristics, and that is the make-up of reference populations, which are central to the calculation of likelihood ratios for strength of evidence estimates. If there are significant patterns for change due to age, we would expect general populations to follow these patterns, and therefore an age mismatch between the reference population and the recordings in question could lead to inaccurate estimation of strength of evidence in those cases. This may also be a concern for selecting foils in a voice line-up (cf. Broeders et al., (2002); Nolan (2003)). A further consequence of these findings would relate to speaker verification systems, where the reference sample against which voices are tested may need to be re-recorded after a certain period of time to prevent inaccuracy due to age-correlated changes.

There are also non-forensic applications of this data, especially in the field of sociophonetic study into language change. Much research is carried out with simulated longitudinal data using an apparent-time (AT) method. If there are individual changes and also age-correlated population differences, this could limit the conclusions that can be drawn from AT data (see §2.1.1.2).

2 Research Review

This chapter presents background research from the three focus areas of this project, namely vocal change over the lifespan (including discussion of social factors which may influence this process), formant dynamics as parameters for forensic speaker comparison, and how the strength of evidence is expressed in casework. This process highlights initial predictions for the study and forms research questions in each area.

2.1 Longitudinal Research

This section addresses previous research into its primary area of investigation: changes in a speaker's voice across the lifetime. The research discussed in this section originates from three different perspectives: firstly from sociolinguistics and sociophonetics and secondly from a forensic perspective. Finally, evidence from the physiological study of the effects of aging on the voice is presented.

2.1.1 Sociolinguistic research

Although this section reviews sociolinguistic research that is relevant to this project, there is relatively little longitudinal data on changes to the speech patterns of individuals, due to the cost and reliability of participants in such an exercise. Those sociolinguistic studies that do exist are largely concerned with a single or small set of phonological variables. This section also presents data about the UK population which are relevant to the present study and which raise concerns about the sampling methods of sociolinguistic studies.

2.1.1.1 Longitudinal methods in sociolinguistics

Within sociolinguistics there are two main methods for researching language change over time: through real-time (RT) and apparent-time (AT) studies (Tillery & Bailey, 2003; Schilling-Estes, 2004).

Real-time studies are essentially longitudinal, following participants from a speech community over a period of time. The approach is further divided into two different methods, trend studies and panel studies. In a trend study, random samples are taken from the same speech community at different time intervals, meaning data come from different speakers at different time points. For example, Blake and Josey (2003) revisited Martha's Vineyard 40 years after Labov's original research there (1962) and found a reversal of the original pattern by sampling a new group from the same community.

Nahkola and Saanilahti (2004) point out that the logical consequence of sampling a new group is that trend studies do not tell us anything about the individual or the idiolect. This is problematic as it assumes that all members of a speech community are equally representative of that group and does not contribute to our understanding of individual variation.

In panel studies, by contrast, the same speakers are re-recorded at different time intervals. This allows for the observation of individuals' changes in speech, either by planned re-recordings or by sourcing speakers from an earlier project for recording. For example, recordings of the Queen's annual Christmas broadcast were used to research accent changes in a number of papers by Harrington, Palethorpe and Watson (2000a; 2000b; 2005). In terms of the current project, panel studies are essential and provide the most relevant data regarding individual vocal change; hence most of the studies presented in this chapter follow that particular methodology.

Apparent time studies are used widely in sociolinguistics since their development in Labov's early work in Martha's Vineyard (1962) and New York (1966). The apparent time construct is based on the widely-held assumption that speakers' speech patterns do not change significantly after adolescence (at least for those speakers who do not move or suffer 'significant life events' (Labov, 1994)). Therefore, data from speakers from different generations within a speech community can indicate either 'genuine' linguistic changes in progress or 'age-grading' (Hockett, 1950). Age-grading is the process where speakers conform to different sociolinguistic patterns at different stages of life, according to a 'group norm which recycles itself through generations'. Hockett calls this "a continuity of tradition in a community through successive generations of children" (1950, p. 452). In Labov's (1963) study of Martha's Vineyard, he described how older speakers had progressively lower nuclei of diphthongs. He attributed this to apparent time changes, but stated that each generation may have also displayed age-graded behaviour (i.e. each generation exhibiting lowering in diphthong nuclei) as a 'secondary factor'. How this distinction between real change and age-grading is reached with AT data is unclear; José comments that "only RT evidence can differentiate between an observed generational difference that is a reflection of language change and one that reflects age grading" (2010, p. 35).

Nahkola and Saanilahti summarise AT as a:

> Synchronic comparison of different age groups within the speech community. The resulting synchronic age variation is then projected into the future, thus creating a dimension of apparent diachrony (2004, p. 75).

2.1.1.2 Real vs. apparent time

There are, however, practical reasons for sociolinguists generally avoiding RT panel data in favour of apparent time. Funding for such long-term projects is rarely awarded (or even sought) and the expense of the project is hard to justify given the risks of speakers dropping out, moving or dying. These factors inevitably reduce the sample as time progresses. Tillery and Bailey (2003) also point out that the sample can become less representative of the speech community over time and that the same conditions are hard to preserve, i.e. social setting, interviewer, mode *et cetera*. Nevertheless, they state that "real-time data is valuable in sociolinguistics, the fact that it is so hard to come by makes it even more valuable" (2003, p. 354). Tillery and Bailey argue that due to these methodological problems, RT data is not better evidence for studying language change than AT data (although this appears to conflate effectiveness of evidence and ease of collection).

For sociolinguists, AT studies have dominated research as they capture changes in progress in a much more time efficient and affordable way. It may be that the perceived success of the approach has led, in part, to this dominance in sociolinguistics; Chambers speculates that AT study of change in progress might be "the most striking single achievement of contemporary linguistics" (1995, p. 147). Despite this, there is acknowledgement that apparent time is only a simulation of real change. For example Bailey (2004), a vocal supporter of AT, notes that it is only a surrogate for RT data and that it is based on certain incompletely tested assumptions. Labov suggests that the rate at which pronunciation changes may well be underestimated in apparent time studies (1994).

It is important to acknowledge that the goals of sociolinguistic work may be very different from those of this project. The aim of most sociolinguistic research is to investigate linguistic patterns that characterise a whole language variety. By contrast, the focus of this project elucidates the processes of vocal change across time *in the individual*. Importantly from a forensic perspective, this study examines those features which remain

stable as well as those which change significantly. In sociolinguistic research, when a member of a speech community leaves that group, they cease to be of interest in regard to that community. In the present context, the individual does not cease to be an interesting individual as soon as they move. In fact the effect of moving is one factor this study seeks to investigate.

Clearly, for sociolinguists, individuals are viewed as the products of a community rather than the other way round. This premise is well demonstrated in Schilling-Estes' introduction to a chapter on Time: "variationists are interested in change at all levels, as well as disentangling genuine changes in a language or language variety from age-related changes in the speech of individuals" (2004, p. 311). In this framework then, age-related, individual change is merely obfuscatory to finding true [sociolinguistic] change. This is surprising, given that the AT methodology relies on the notion, which has been assumed in a relatively uncritical manner, that language is stable post-adolescence. Laver and Trudgill (1979, p. 5) make the following comment:

> These comments about accent differences between speakers make the assumption that a speaker's accent is fixed and unchanging. It seldom is, of course, and a further area where linguistic concepts can help us to refine our analysis of social markers in speech concerns certain aspects of the notion of linguistic variability.

2.1.1.3 Sociolinguistic research in real-time

Bowie (2010, p. 1) notes that:

> Studies of linguistic production and *post*-adolescent aging generally deal with the relationship between speech processes and age-associated mental and physical pathologies. However, it is also worthwhile to look into the linguistic processes associated with what is often thought of as 'normal' aging – that is, aging marked simply by the continuing progression of time rather than by significant mental and physical pathology, or even by the occurrence of major social changes.

To this effect, there is a small body of RT research which has set out to test the effects of 'normal aging' and consequently the underlying assumption of AT that any effects represent age-grading. Bowie investigated stability in individuals' realisations of changes in progress (2005) and other less prevalent changes (2010). He found that in a series of recordings of individuals from Mormon Church conferences across several decades there was variability in the *fill-fell* merger, pre-lateral raising of /ɪ/, /æ/ raising and diphthongisation and /u/ fronting. Chao et al. (2007) also discovered significant

difference in individuals' r-lessness across different time periods within the same corpus of Mormon leaders' speeches. Change was not linear and was variable within the individual. As well as not having moved and not suffering major 'life events', these speakers were also all recorded in the same speech environment (i.e. lectures to the church conference), so would not be predicted by an AT approach to show any significant difference across recordings. These kinds of details are lost in traditional AT studies were large groups are reduced to percentages which describe gradual changes in usage. Of course, some features were also realised consistently across time. Bowie stresses that there is a mixed pattern across the homogenous group of speakers and that individuals change in different ways, which highlights the "importance of looking at individuals' behaviour as the behaviour of individuals, and not just as members of a larger group" (2010, p. 20).

To expand on these findings, Bowie (2006) explored the speech habits of American speakers from Waldorf, Maryland in both AT and RT. Both methods were used to investigate monophthongisation of /aɪ/, and showed, using logistic regression, the separate effects of RT and AT analysis across the '30s, '60s-'70s and the '90s. Although the RT and AT methods showed a similar trend in each case, the two methods showed differing patterns. In this example, although the AT study showed a linear relationship between age and monophthongisation, the RT approach was able to identify that in most speakers in was predicted first by a weakening in the glide. Bowie raises the point that researchers do not know enough about between-session variability to claim that either approach can represent real change, a point that is pertinent in forensic terms and is returned to in §2.1.3.1.

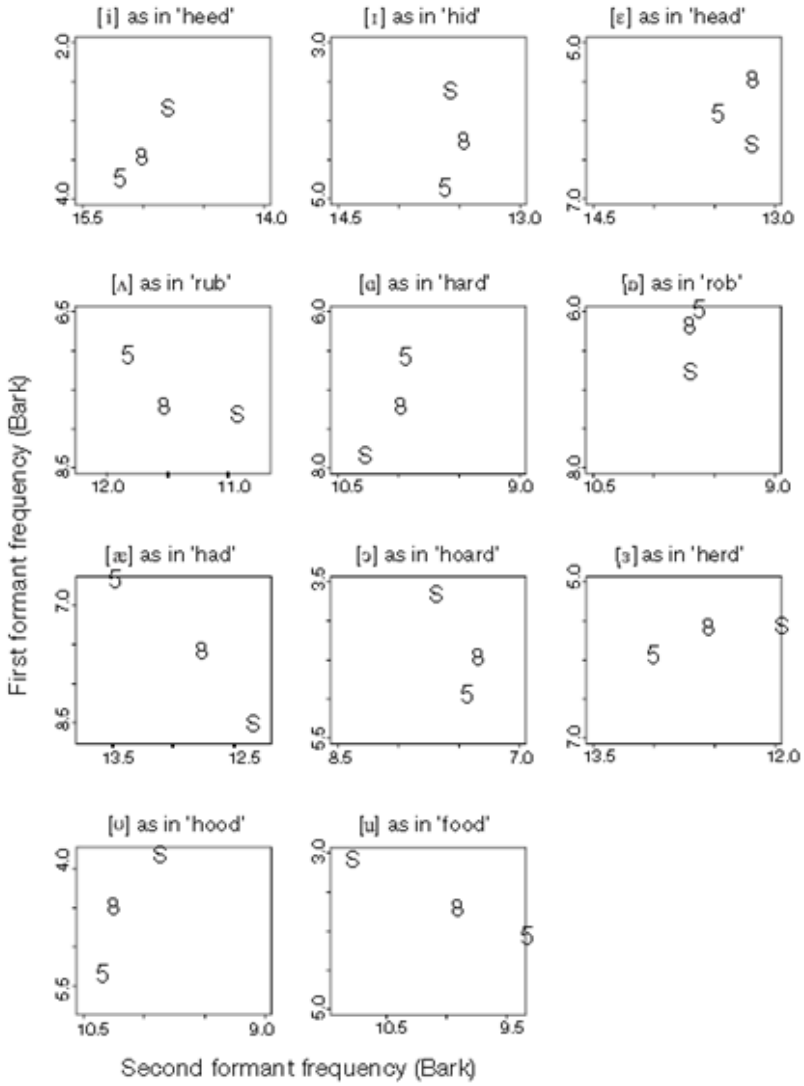
Further to this, Bowie (2010) explored intraspeaker variation in three speakers across just one year and found significant changes in vowel realisation in 42-50% of the stressed vowels analysed (n=345-605 vowels for each year). Only some of these changes conformed to ongoing changes in the language variety (i.e. *pin-pen* merger and raising of /aɪ/ before voiceless obstruents). Over such a short period these cannot all be sociolinguistically motivated and Bowie points to intraspeaker inconsistency in production as part of the reason for the discrepancy. Still, these results demonstrate how adult linguistic production cannot be treated as systemically stable. From his body of work, Bowie concludes that "individuals vary over the course of their adult lifespans, and this

intra-individual variation is both real and large" (2010, p. 21). The problem with Bowie's work, as with almost all real-time studies, is that the size of data samples is so limited due to difficulties collecting real-time data. Although the results are interesting, it is difficult to relate them to large numbers of speakers with confidence. Ideally, a study would have access to a large set of real-time data. In reality, a combined real and apparent time approach is practically the most convincing option.

To compare the performance of AT to RT in Finnish, Nahkola and Saanhilahti (2004) carried out simultaneous AT and RT research on six different age groups of nine speakers across a ten year period. They analysed 14 variables, all of which were undergoing some form of change with respect to the Standard Finnish (StFi) variable; some experiencing new non-standard forms and others returning to the StFi variant. In order to investigate the success of the AT approach, they used AT to predict the changes in each variant and compared the 'actual' results from RT data against the predictions. In 10/14 variables the AT prediction was 'fairly successful', but in 4/14 of the cases the prediction was not borne out by the data. They stated that AT was valid for 'steadily advancing change', where the pattern was consistent through the age groups (generally StFi variants advancing into the rural dialect), although "if middle aged and young [speakers] are taking the change in opposite directions, the AT prediction will be basically no more than a guess" (Nahkola & Saanilahti, 2004, p. 85). They also make interesting observations about individual speakers. Children learning categorical features (i.e. those with little variation in the community) are likely to have that feature remain categorical, meaning there is very little chance of change in adulthood. Those features which have competing variants, however, are much more likely to shift usage during adulthood. Interestingly "the more equal the proportions of the rivalling are, the more likely it is that one of the variants will gain dominancy during the speakers' lifetime" (Nahkola & Saanilahti, 2004, p. 75). Even more relevant for the present thesis is the finding that within individuals, around two fifths of variation patterns had changed over the ten year period and that this was much more prevalent in individual results than when the group results were averaged out (echoing Bowie's (2010) call for more individual attention). Crucially, this variation was "as common for the middle aged informants as for the young" (Nahkola & Saanilahti, 2004, p. 87). So although AT did successfully predict limited variation in this dialect, there are variations demonstrated by RT data that are not accounted for. Moreover, the findings show a high level of intraspeaker variation in variable features even in middle age.

Harrington et al. (2000a; 2000b; 2005) carried out acoustic analysis on a very specific variety of English to investigate exactly these kinds of linguistic changes in speakers; using archive recordings of Christmas broadcasts between 1950 and 1980, they carried out vowel formant measures on vowels from Queen Elizabeth II's English, a version of Wells' U-RP (1982). Harrington et al. demonstrated significant changes in monophthongs (2000a; 2000b) and diphthongs (2005). Data for the Queen's vowels in 1950 and 1980 are shown in comparison with widespread SSBE values in Figure 1 below:

Figure 1 – F1 and F2 of monophthongs from the Queen's broadcasts in 1950 (5), 1980 (8) and average SSBE frequency (S). All frequencies are converted to the Bark scale by the researchers



Source: Harrington et al. (2000b)

It is very clear from this data that not only are there significant changes to the production of monophthongs, but that the changes are also largely predictable from more widespread SSBE varieties (these were normalised from measurements of a sample of female SSBE speakers). Harrington et al. (2005) also found this pattern for the /aɪ, aʊ, oʊ,

eɪ/ diphthongs, where 16/20 predictions based on 1980s changes towards SSBE were borne out with significant results. They argue, from this single, quite exceptional case study, that older members of the community are unwittingly adapting their accents towards community changes, brought about by young innovative speakers (Harrington, Palethorpe, & Watson, 2005). They also contend that the extent of this change is probably more marked for average members of the community as they are not tasked with preserving a very individual, traditional form of English. While there were widespread changes in F1 and F2, F3 remained relatively stable throughout the recordings. What this study did not take into account are the kinds of physiologically caused changes we might expect in a speaker between 25-30 and 50-60 years of age (indeed F1 changes and limited F2 and F3 changes are predictable from a physiological perspective, cf. §2.1.4.10). In the vowels in the figure above, all but three are showing a reduction in F1, which is an expected formant change with age (Endres, Bambach, & Flösser, 1971; Linville & Rens, 2001; Reubold, Harrington, & Kleber, 2010). It may simply be a coincidence that this follows a general shift in SSBE, or more likely a combination of both factors.

Reubold et al. (2010) expanded on this methodology, using archive data from four further speakers (Margaret Lockwood, Roy Plomley, Margaret Thatcher, Alistair Cooke) to investigate vocal aging across a period of 29-35 years, and the relationship between F0 and F1. Again these subjects cannot be characterised as 'normal' speakers. Thatcher had well-documented vocal training (Van Buuren, 1988) and Lockwood, Plomley and Cooke are all 'professional' voice users (Linville, 2001). Despite this they still provide valuable data on RT change. All speakers demonstrated significant changes in both F0 and F1, but fewer changes in F2 and F3 (for more detailed discussion of F0 and related formant changes see §2.1.4.10). These findings, and those in their previous work, lead the researchers to warn that "apparent-time studies may under- or overestimate sound changes if [their] effect on formants is different from, or similar to, that of physiologically-based changes due to aging" (2006, p. 639).

Sankoff (2004) used a sub-set of the '*7 Up*' (Apted, 1964) dataset for the present study to investigate the BATH/START and FOOT/STRUT patterns in two mobile Northern speakers: Neil and Nick. Both were raised in the north of England where they would use /a/ for BATH and /ʊ/ for STRUT words. Sankoff shows with an auditory analysis from ages 7-35 that

speakers are able, and likely where there is a benefit to that speaker, to make 'significant phonetic and possibly phonemic alterations to their speech after adolescence' (2004, p. 18). Both speakers develop forms that are different from those they learned at childhood and beyond, as both were geographically mobile to areas with different accent types. This would seem to suggest that speakers are not inflexible in their speech after adolescence. However, Sankoff does temper this conclusion by suggesting to readers that this kind of geographical mobility is unusual. Support for this statement is unclear, and more specific statistics for these kinds of mobility are presented later in section 2.1.2.

Finally, Cukor-Avila (2002) investigated grammatical and discourse variants, and found that there was change in verbs of quotation in African American vernacular English (AAVE) across three years for 14 speakers. While there were changes throughout adulthood, this was substantial in a number of speakers and not present in others. While the present study does not examine grammar explicitly, it is still useful to have a full picture of adult stability. Bailey (2004), however, cites this research as support for stable adult vernacular, claiming that in this case, all that is demonstrated is that the threshold for change is maybe slightly later than adolescence and perhaps early adulthood and that some of these grammatical changes are simply lexical changes to conform to social norms (i.e. the use of *be like*).

2.1.1.4 Summary

General patterns and conditions of change are presented in the previous section. These are summarised below:

- There is evidence for linguistic instability after adolescence/young adulthood, in:
    - Phonological choices
    - Phonetic quality of vowels
    - F0
    - Grammar
- The type, prevalence and magnitude of post-adolescent vocal change can be very individual
- Change may be restricted to features that are variable within the accent type
- Vowel pronunciation changes may be predictable from language varieties which are more widespread and culturally accepted than the speakers' own
- Speakers' F3 seems to remain most stable and F2 more stable than F1

As the AT construct relies on adult linguistic stability, this first bullet point above must present a real issue for AT studies, and presents a first research question for this thesis:

- What is the extent of vocal instability in adulthood and how will this affect the interpretation of results from apparent-time studies?

Labov (1994) comments that:

> Apparent time studies may underestimate the actual rate of sound change, since older speakers show a limited tendency towards communal change, participating to a small extent in the change taking place around them.

What if speakers' changes are more than limited tendencies, such as in the case of Harrington et al.'s Queen data? What if there are changes in opposing directions, caused by physiological or social factors? These questions about individuals are generally ignored in the search for community patterns. Furthermore, the majority of these studies exclude those speakers who have moved or undergone a significant 'life event', as well as those who have suffered physical or mental trauma. However, these non-mobile, healthy subjects should be viewed as unrealistic or 'over-idealistic' speakers in terms of forensic casework, as the following statistics from the UK demonstrate.

2.1.2 **Social factors and individual change**

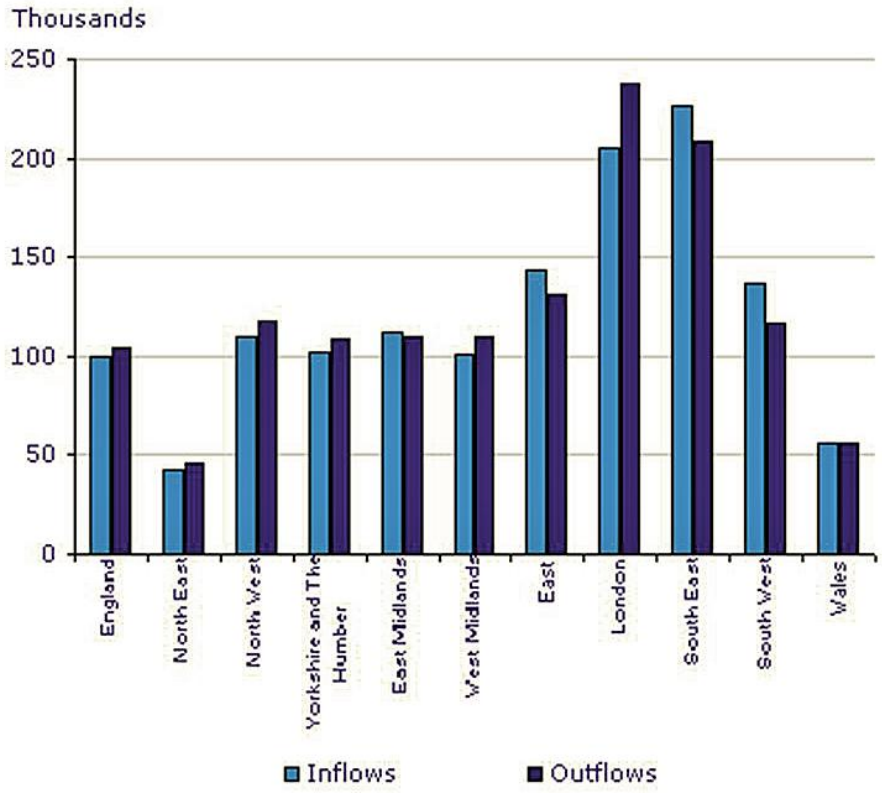
This section illustrates population data from England or the UK to place discussions of sociolinguistic and forensic approaches in context, and give an insight into the concept of a 'normal' or 'ideal' speaker and how this interacts with practical conditions. This has implications for sociolinguistic research, but more importantly on how we use data from these and other studies in forensic settings. It also presents the limited research on how mobility affects speakers' language use.

There are probably an exhaustive number of factors which could be discussed, those presented here are largely represented in the speakers in the current dataset or those factors involved in screening sociolinguistic participants, to give an idea of how representative a 'traditional' sociolinguistic subject really is. For reference, the UK population is estimated to be 62.3 million on the 30th June 2011 (Office for National Statistics, 2011).

2.1.2.1 Geographical mobility

To give an overall snapshot of geographical mobility, or 'internal migration', within England and Wales for one year (June 2008-June 2009), there were between around 100,000 to 200,000 internal migration movements in each of nine regions of England, as the figure below demonstrates.

Figure 2 - Internal migration movements between countries and regions within England and Wales for year to June 2009



Source: (Office for National Statistics, 2010)

Table 2 below provides a slightly longer view over the last eight years prior to the figure above, which shows that internal mobility across the last decade is fairly steadily around 1.3 million across the UK and within regions, though with a slight decline. The division used in this migration classification is also quite useful in linguistic terms, as most of the regions have accents which are different. This naturally underestimates moves that could result in a different origin and destination accent type, as intra-regional differences (i.e. west Yorkshire to south Yorkshire) are not displayed here. Furthermore, all these data are based on patient records from the National Health Service Central Register (NHSCR) when re-registering at new health practices. It would be sensible to assume this is a slight underestimate of actual migration, given that not everyone will register, at least straight away, and that everyone who has registered has definitely moved.

Table 2 - Internal Migration in the UK (thousands) by destination and by origin

Region of destination	2002	2003	2004	2005	2006	2007	2008	2009
England	101	98	97	98	96	92	93	92
North East	43	42	41	40	40	39	38	37
North West	109	109	105	102	100	96	96	95
Yorkshire & Humber	100	99	98	94	93	91	90	90
East Midlands	120	115	112	106	107	107	100	100
West Midlands	99	95	95	94	93	91	89	88
East	150	145	146	139	144	143	133	131
London	155	148	155	161	168	164	177	178
South East	229	221	223	217	225	221	206	206
South West	146	142	139	132	136	134	125	122
Wales	64	63	60	56	57	55	51	50
Scotland	53	60	57	59	50	56	46	46
Northern Ireland	11	12	13	12	13	12	11	10
TOTAL	**1377**	**1348**	**1339**	**1310**	**1320**	**1300**	**1254**	**1245**

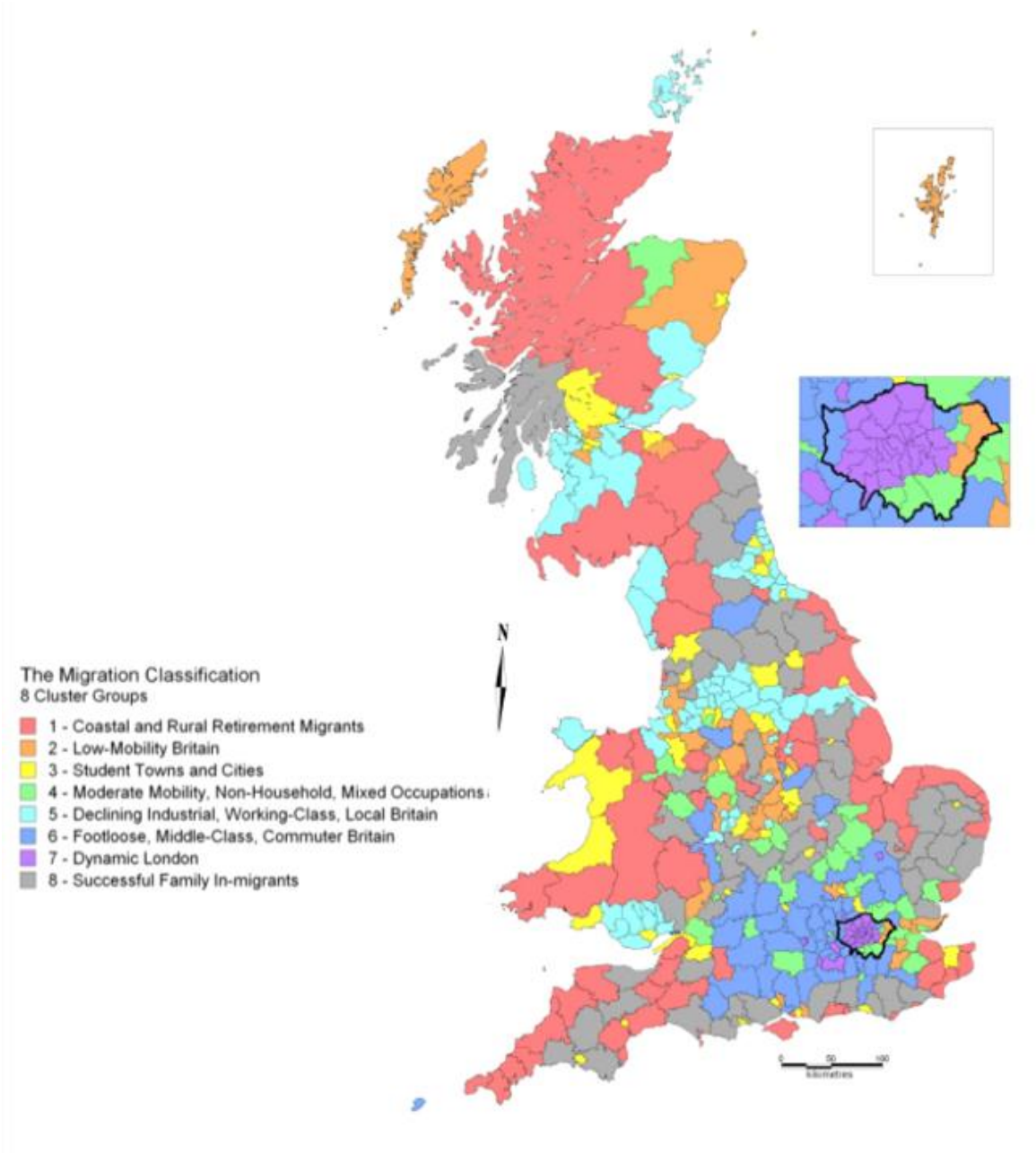
Region of origin	2002	2003	2004	2005	2006	2007	2008	2009
England	119	126	122	118	110	114	101	98
North East	41	40	39	39	39	39	40	40
North West	108	104	104	103	104	104	103	102
Yorkshire & Humber	95	93	92	93	94	95	94	95
East Midlands	97	96	97	97	99	98	95	95
West Midlands	103	102	101	99	101	100	98	98
East	130	127	128	124	127	124	118	118
London	263	263	260	243	247	246	223	216
South East	220	211	208	201	201	198	189	185
South West	111	108	108	107	108	105	103	104
Wales	50	48	49	50	49	48	48	48
Scotland	48	46	45	45	44	41	42	41
Northern Ireland	11	12	10	13	11	11	10	11
TOTAL	**1395**	**1376**	**1364**	**1330**	**1334**	**1322**	**1262**	**1251**

Note: Slight differences in origin and destination are due to adjustments from cross-border flow data which is assumed to be more accurate than NHSCR records; also differences in assumed time lag between each dataset.

Data source: (Office for National Statistics, 2003-2010)

Clearly internal migration is a constant feature of the UK, and with just over 2% of the population making significant regional moves every year, a large proportion of the population is likely to have been mobile at some point. Furthermore, migration not only differs by origin, but different regions experience different styles of internal migration, as summarised below.

Figure 3 - CIDER Migration classification summary



Source: (Denmett & Stillwell, 2011)

This might impact on how speech changes as a result, if there are age-related trends, especially between, for example, student town or dynamic urban areas compared with retirement areas. In forensic terms, the migration status of an area could have significant impact on how we assess the population of potential perpetrators (see §2.3.2.5) and relevant accent features. This would be especially different in high mobility areas such as those urban and student areas against low mobility communities.

2.1.2.2 Health issues

Another factor which is screened for in traditional sociolinguistic methodology is physical trauma (Labov, 1994). The definition itself is somewhat unclear and difficult to implement, but to give a general idea of how prevalent this might be in the UK population NHS critical care data are presented below. Categories ACDEHPQRT all involve physiological units involved in producing speech and could presumably lead to a subject being screened out of a study. This represents 61% of all critical care treatments and means that for these two years, between 100,000 and 90,000 people were admitted for critical care in these categories.

Table 3 - Patients in critical care for 2008/9-2009/10 by condition and gender

		2009-10			2008-09		
HRG 3.5 chapter		**Male**	**Female**	**Total**	**Male**	**Female**	**Total**
A	The nervous system	6,290	4,724	11,014	5,107	3,902	9,009
B	Eyes and periorbita	82	32	114	73	28	101
C	Mouth, head, neck and ears	4,878	3,190	8,068	3,778	2,391	6,169
D	Respiratory system	11,461	9,510	20,971	8,721	7,448	16,169
E	Cardiac surgery and primary cardiac conditions	25,751	11,293	37,044	20,516	9,110	29,626
F	Digestive system	15,344	13,506	28,850	11,960	10,289	22,249
G	Hepato-biliary and pancreatic system	4,736	3,588	8,324	3,516	2,760	6,276
H	Musculoskeletal system	4,974	4,484	9,458	3,647	3,270	6,917
J	Skin, breast and burns	1,645	1,765	3,410	1,186	1,411	2,597
K	Endocrine and metabolic system	995	1,353	2,348	892	1,058	1,950
L	Urinary tract and male reproductive system	6,390	4,090	10,480	5,176	3,135	8,311
M	Female reproductive system	11	1,864	1,875	8	1,566	1,574
N	Obstetrics and neonatal care	148	1,698	1,846	32	1,220	1,252
P	Diseases of childhood	802	727	1,529	524	499	1,023
Q	Vascular system	7,959	3,416	11,375	6,239	2,614	8,853
R	Spinal surgery and primary spinal conditions	1,499	1,411	2,910	932	924	1,856
S	Haematology, infectious diseases, poisoning and non-specific groupings	3,891	3,687	7,578	3,337	2,980	6,317
T	Mental health	224	99	323	154	79	233
U	Unidentified groups	902	625	1,527	857	736	1,593
Blank	Unknown	83	35	118	47	29	76
Total		**98,065**	**71,097**	**169,162**	**76,702**	**55,449**	**132,151**

Source: (NHS Information Centre, 2011a)

2.1.2.3 Mental health

Table 4 - Mental Health summary for England: 1 detainees in NHS hospitals under Mental Health Act, 1983; 2 those in supervised community treatment (SCT); 3 in NHS secondary mental health care or in CPA

		2006-07	2007-08	2008-09	2009-10	2010-11
1	Total Formal Admissions*	25,624	26,140	25,902	28,057	27,471
	via Courts and Prison	1,339	1,408	1,370	1,431	1,421
2	Supervised community treatment			2,109	4,020	3,730
3	People in NHS secondary mental health care	1,149,472	1,151,260	1,190,542	1,222,365	1,270,731
	CPA**	131,983	135,040	164,998	171,248	230,608

Note: *excludes places of safety admission

**Care program approach (illegibility includes severe mental disorder, risk of suicide or harm to others, vulnerable, drug/alcohol abuse, recently detained under Mental Health Act)

Data source: **1, 2** (NHS Information Centre, 2011d)

3 (NHS Information Centre, 2011e)

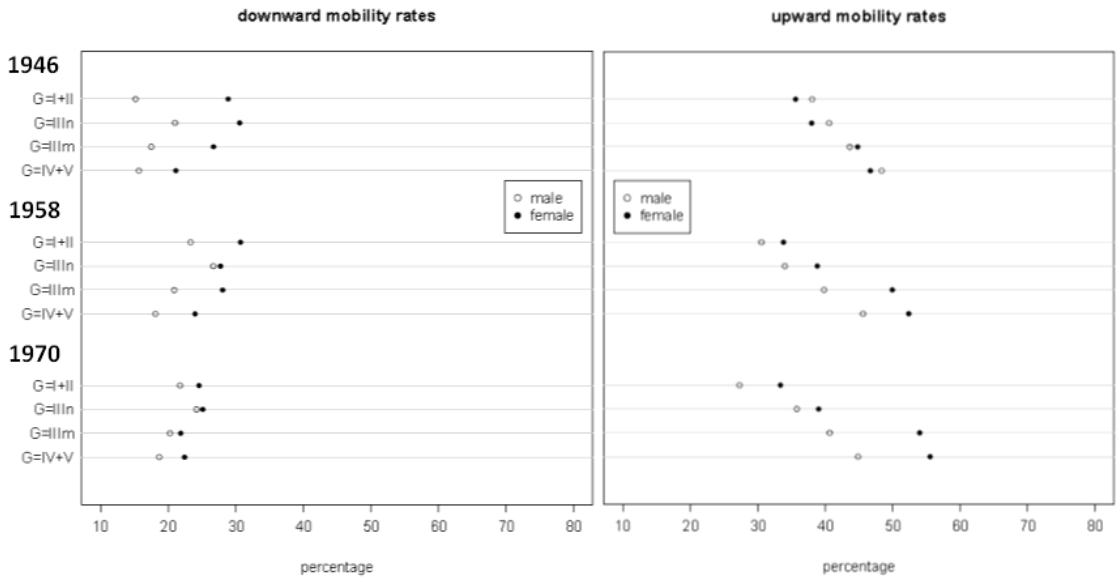
Although relatively few people are detained through the Mental Health Act (1983) and very few are detained through the courts and prison system, there is a large proportion of the UK population in secondary mental health care and a number in care programmes (note that many of those in the table above probably overlap with those detained with drug-related mental health issues in Table 6 in the following sections). This is comparable with the number of geographically mobile speakers, and like those speakers, those with mental health issues are generally excluded from sociolinguistic study.

2.1.2.4 Social Mobility

Social mobility is largely described using an intergenerational professional model, i.e. the difference between a parent's and child's eventual type of profession, to attempt to measure a notion of what Crawford et al. (2011) call "equality of opportunity". Blanden, Gregg and Machin (2005) show that for recent years (i.e. birth cohorts from 1970 onwards) mobility in Britain has been decreasing. Chan and Boliver (2011) argue that more than just parents' class counts, as grandparents' social position plays a large part, although grandparents' position is still largely determinative (i.e. grandchild is likely to follow grandparent). Figure 4 below corroborates earlier findings (Blanden, Gregg, & Machin, 2005) that intergenerational mobility is decreasing in Britain and gives us a clearer idea of mobility rates. Cohorts across different classes show that percentage of upwardly mobile people, especially males, is much lower for those born in 1970 than in 1946 or 1950. Females in the bottom two classes show fairly consistent mobility, with

slight increases. Downward mobility appears to have remained somewhat more stable (especially for females).

Figure 4 - Upward and downward intergenerational social mobility in three generations of adults (aged 34-43, birth year given) by gender and class of grandparent



Note: G = Classes: professional and managerial; skilled non-manual; skilled manual; unskilled manual
Source edited from: (Chan & Boliver, 2011)

However, this does not help describe the likely pattern for within-speaker changes, i.e. changes between the class categories within a lifetime. The question is whether it is worth drawing a distinction between parents' socio-economic status (SES) and an individual's, i.e. whether there is necessarily a blank slate or 'starting SES' for an individual except that transmitted from their family. The usefulness of this intra-generational data seems limited, given the strong correlations between parents' SES and the probable SES destination of the next generation in adulthood (Corak, 2004).

Furthermore, the reliance on profession or earnings for defining social status or class is problematic (especially when considering some of the speakers in the present study). We also need to question whether these (largely economic) metrics used in these studies reflect class, as class is a concept that is strongly linked to language in the UK (Llamas & Watt, 2010), particularly in traditional standard varieties like RP and U-RP (Wells, 1982).

2.1.2.5 Smoking and substance abuse

Smoking has been shown to affect acoustic parameters significantly, particularly F0 (see §2.1.4.7). Proportions of the English population who smoke are presented below; although frequency has dropped since 1948 from 52% to 21% in 2008, a large proportion

of people are smokers or have been smokers (47%). This should have an impact on how analysts assess F0 in casework, as comparing non-smokers to a smoking or mixed reference point and vice versa may reduce reliability.

Table 5 - Percentage of adults (over 16) who smoke in England

1	1948	1982	1990	1992	1994	1996	1998
Current smoker	52	35	29	28	26	28	27
Ex-smoker		23	26	26	26	25	26
Never/occasionally smoked		43	45	46	47	47	48

2	2000	2001	2002	2003	2004	2005	2006	2007	2008
Current smoker	27	27	26	25	25	24	22	21	21
Ex-smoker	24	24	24	24	24	24	24	25	26
Never/occasionally smoked	50	50	50	51	52	53	54	54	53

Note: 1 and 2 are calculated using a different weighting, 1948 data are for Great Britain
Data source: (NHS Information Centre, 2011b)

Despite early significant findings on the effects of drug use (Papp, Schreuder, Theunissen, & Ramaekers, 2011), and particularly drug addiction (Papp, 2009) on speech, the number of different drugs and drug categories, alongside ethical concerns, mean that relatively little research into this subject has been undertaken. Table 6 below indicates that between 10-12% of respondents to an NHS survey reported having used some sort of illegal drug in the last year, and that a large number of people are treated for substance abuse problems. Table 7 indicates that a large proportion of the UK population has used drugs at any point in their lifetime, with 34.9% of respondents having used any drug.

Table 6 - Substance abuse information for England: 1 Self-report use for UK population; 2 diagnosis of drug-related mental health issues with hospital admission; 3 long-term habit treatment (NDTMS)

		1998-99	2000-01	2002-03	2004-05	2006-07	2007-08	2008-09	2009-10
1	Self-report any drug use in last year	12.1%	11.9%	12.2%	11.3%	10.0%			
2	Primary diagnosis	9,131	8,027	7,691	7,857	6,743	6,675	5,668	5,809
	Primary/Secondary diagnosis	24,236	25,683	31,490	35,737	38,170	40,421	42,170	44,585
3	In NDTMS*						195,464**	207,580**	206,889**

Note: This includes private patients in NHS hospitals but not in private care, some data were excluded where location was not available
*National Drug Treatment Monitoring System
**72-73% male participants in NDTMS
Source: (NHS Information Centre, 2008-2010)

Table 7 - Showing the proportion of 16-59 year-olds reporting having used drugs in their lifetime by age group, data from 2005/6

Drug	Age group							
	16–19 (%)	*20–24 (%)*	*25–29 (%)*	*30–34 (%)*	*35–44 (%)*	*45–54 (%)*	*55–59 (%)*	*All ages 16–59 (%)*
Cannabis	35.1	44.4	46.7	40.1	28.5	18.8	11.1	29.8
Amphetamines	7.5	14.5	23.8	20.5	11.8	5.6	2.8	11.5
Any cocaine	6.5	14.5	15.2	10.4	6.4	2.4	1.1	7.3
Ecstasy	5.8	14.4	18.2	14.0	5.7	0.9	0.2	7.2
Opiates	0.4	1.1	1.9	1.2	0.9	0.5	0.3	0.9
Class A	11.2	21.8	26.5	21.4	13.6	6.6	3.6	13.9
Any drug	40.4	49.0	51.6	45.8	34.2	23.4	15.4	34.9

Source: (Reuter & Stevens, 2007)

Furthermore, Table 8 below shows that the estimates for opiate and crack cocaine use are much more expansive than those registered in care (above); Hay et al (2009) estimate that not only are there many more Class A drug users than reported, but this number is rising exponentially and is concentrated in 'areas of disadvantage', which are generally linked to high-crime areas.

Table 8 - National prevalence estimates for using opiates and/or crack cocaine

Drug	Estimate
Combination of opiate and crack	321,229
Opiate	262,428
Crack cocaine	188,697

Source: (Hay, Gannon, Casey, & Millar, 2009)

Even this limited snapshot of the data shows that drug use is prevalent, even in the general population. In forensic terms, the criminal population is probably of greater interest than the general population, and drug use is reported as being even higher in criminals. According to Reuter and Stevens (2007, p. 8):

> Some of the estimated 327,000 problem drug users in England commit very high numbers of offences – most commonly shoplifting – to fund their drug use. Around a fifth of arrestees appear to be dependent on heroin.

Drug use and alcohol abuse are common amongst arrestees and would be present in many forensic recordings: the table below shows reported drug use (corroborated by an

oral swab test) among arrestees. The response rate for this test was under 25%, which might indicate that the actual incidence is higher.

Table 9 - Proportions of sampled arrestees (2003-2004) reporting drug use and alcohol abuse in England and Wales

Reporting drug/alcohol use	*Percentage of sampled arrestees*
Use of cannabis in last month	46
Use of heroin in last month	18
Use of crack in last month	15
Use of cocaine in last month	10
Dependent use of heroin	18
Problematic or hazardous drinking	57

Source: (Reuter & Stevens, 2007) adapted from (Boreham, Fuller, Hills, & Pudney, 2006)

2.1.2.6 Summary

There are further factors not explicitly mentioned above that may be likely to be excluded from sociolinguistic study which impact on speech patterns, and should also be considered in a social or forensic framework. From this limited data it becomes clear that the over-idealised speakers sampled in linguistic studies may not provide an accurate picture of what real speakers in the UK are like. With respect to changes to the voice over time, it is likely that many of these factors have an effect outside of the normal scope of aging.

Clearly there is a need to consider these factors in our model of speech behaviour; not only are sociolinguistic studies tending to discount those who are geographically mobile (around 1.3 million annually) and those in care for mental health issues (around 1.4 million annually), there is a dearth of research investigating the specific effects of these factors. Furthermore there are other factors which may affect speech which are not taken into account in sociolinguistic studies; in some cases, such as smoking or drug abuse, these are known to have potential effects on speech. With regards to many of these factors, such as social mobility, the effects are under-researched but presumably could have significant effects on individuals' speech (probably in a very individual way). Detailed research into these factors is important, but outside the scope of this study (although it comments on the possible influences of these factors on sociolinguistic study and methods for assessing strength of evidence).

In particular reference to the criminal population, some of these factors may be proportionately more frequent in criminals, which makes a more detailed study of their effects on speech a concern for forensic speech analysis. In FSC it is important to be able to estimate the effects of known factors in a case, and also to have knowledge of how factors might affect population distributions.

2.1.2.7 Individual motivations for change

This section discusses ideas relating to individuals' motivations for behavioural changes in speech, which informs observations about data in the present study (factors with biological effects, such as smoking, are covered later in section 2.1.4). Although there is a plethora of studies on how neighbouring accent areas behave, or how speakers behave or innovate within a speech community, there is relatively little written on, for example, the effects of social or geographical mobility on individual speakers. What is written largely refers to the effects of mobility on language communities. This is precisely because speakers like this are routinely excluded from traditional sociolinguistic studies.

This exclusion is surprising given the mobility data presented in the previous section, but probably due to a number of factors, principally how the study of language has developed. Structuralist ideas of language were rooted in the political ideology of nationalism and ideas about shared history and culture, expressed through language (Johnstone, 2000), which partly explains the view of language use as a shared utility rather than means of individual expression. The goals of generative theory also contribute in being concerned with an "idealised speaker-hearer in an idealised homogenous community" (Johnstone, 2000, p. 411). A further problem with trying to describe individuals' linguistic behaviour rather than a group trend is the amount of data this requires: Hudson (1996, p. 147) describes the situation for social influence of language as a "highly complex multi-dimensional space". For traditional approaches, change in the individual comes second best to the study of changes in the linguistic system. There are, however, ideas of the individual's role in language use which might prove useful in interpreting data from this dataset.

Despite the complexity of these speech contexts and of human interaction in general, one thing that has long been shown to be true is that people act in a way that holds personal benefit. For Bourdieu (1977; 1991), speakers accumulate a set of linguistic resources (linguistic *habitus*) from which they can then transact with a linguistic marketplace, i.e.

speakers' use of language is motivated by what they can gain in that environment (*marketplace)*:

> Speaker's linguistic strategies are oriented...by the chances of profit for that particular speaker, occupying a particular position in the structure of distribution of capital...[this is] not the particular speaker's personal chances of profit, but those chances evaluated by him in terms of a particular habitus, which govern his perception and appreciation of individual chances. (Bourdieu, 1977, pp. 653-5)

Woolard (2009) criticises Bourdieu's description for being too centred on a single dominant market which relies on notions of uniform or consensual prestige or values, which must be "temporarily suspended when the vernacular is used" (Milroy & Milroy, 1992). However, Bordieu's emphasis on the dominant is based on discussions of how the dominant perpetuate that dominance through language and education. His theory does encapsulate the idea that general *habitus* is individual and therefore different individuals perceive value in different environments or 'markets'. For Bourdieu (1977, p. 653) "discourse is a symbolic asset which can receive different values depending on the market in which it is offered". In general terms, speakers will act where they see value or gain, dependant on their individual social surrounding. This simple hypothesis is particularly useful when considering the patterns found in the present study.

In terms of the practical process of how an individual goes about enacting change, Johnstone argues (from Biber and Finegan (1989), Ochs (1992) and Eckert (2000)) that repeated stance-taking builds linguistic styles which may make up a speaker's repertoire (2009, p. 29): "repeatable linguistic styles emerge out of stancetaking *strategies that prove repeatedly useful for particular speakers* in particular kinds of interaction" [my emphasis]. Johnstone takes the model of stance-taking across groups and focuses it onto an individual speaker, stating that repeated stancetaking moves can form a repeatable style which reflects not only social, but also personal identity (Johnstone, 2009). One further issue that Johnstone (2000) problematises is that in practice there is a blurring of the traditional distinction drawn between social variation (seen as the automatic results of facts about the speaker) and stylistic variation (defined as strategic adaptation to the situation at hand). In fact in cases where social changes may have originated from the use of strategic stylistic changes, this distinction becomes problematic. If speakers (such as those in the present study) are using style consistently for social effect, how do we define this and is it necessary to make a distinction at all?

The relevance both these approaches hold for the current study is the idea that, for individuals, utility is at the heart of linguistic change. Furthermore, that social interaction is a highly complex system which makes studying these kinds of change difficult.

2.1.2.8 Mobility as motivation to change

There are general effects of geographical and social mobility that are referred to as part of the processes of language change and enregisterment which might indicate some likely properties of a mobile speaker. Johnstone, Andrus and Danielson (2006) describe mobility (mainly geographical, but promoted by social) as a factor in making speakers more aware of the indexical meaning of their speech patterns, by coming into contact with those that are different. Mobility has the effect of raising the level of awareness for those speakers from the level of a Labovian (1972) indicator to a marker or from a Silversteinian (2003) "*n*-th-order-indexical" to an "*n+*1-th-order-indexical*".* That is, a speaker may use an indicator more frequently if it is common in the area that they are from, but a marker shows stylistic variation. This marker will be correlated with use in different contexts because it is socially meaningful (whether they are aware of it or not). Presumably greater mobility or mobility into speech areas with more difference or difference in a greater number of features creates more awareness of the indexical nature of those features. Milroy and Milroy (1992) argue that mobile individuals' social networks are more diffuse and less multiplex. This means that these speakers are less likely to come under pressure to use conservative language than non-mobile speakers, who are more likely to live within less close-knit and multiplex social networks. It must be stressed that these observations are made about communities and not individuals, and in both cases the arguments are made with anecdotal reference to trends in increasing mobility which are not substantiated with any data on either geographical or social mobility.

There are a number of studies which examine the process of acquiring a second dialect following an internal migration movement to an area with a different accent type, which are well summarised in Nycz (2011). Nycz (2011) looked at *cot/caught* and raised (aw) nucleus in Canadian speakers who moved to New York. She found that speakers developed the New York distinction, but retained Canadian raising, and that those who had a larger distinction also raised (aw) more. Chambers (1988; 1992) found that in six pre-teen and teenage Canadians who moved to Oxfordshire, acquisition of new dialect

forms (r-lessness, intrusive-r, low back vowel contrast) depended largely on age (with age inversely correlated with extent of acquisition). They also seemed to be more likely to stop producing marked Canadian forms (flapping) than to attain new forms from the 'host' area. Tagliamonte and Molfenter (2007) looked at Canadian children who moved to York under the age of five, and found that they also dropped the use of flapping very easily. Many of these studies, however, are concerned with very obviously 'non-British English' stereotype-type (Labov, 1972) features such as flapping, and present findings about mobility and assuming accents on the basis of these small set of features. What they do show, however, is the ability of mobile speakers to assume novel accent features from host areas.

In terms of adults (or those beyond the 'critical period'), Bowie (2000) found that geographically mobile speakers from Waldorf, Maryland varied from speaker to speaker in how they changed after moving in a number of features. However, he did observe that (like in Nahkola and Saanhilahti (2004)) those forms which were variable were more likely to change. Evans and Iversen (2007) present evidence from about 20 speakers of different northern varieties of English before, during and after starting University at southern institutions. They found that speakers did not suddenly produce BATH category words with [ɑ:] but did use a 'softened' (Evans & Iversen, 2007) hybrid form. They also found that northern speakers shifted STRUT class words to a more centralised production, like but not reaching [ʌ].

In summary, it is predicted that speakers are likely to use language in a self-beneficial way, according to their personal beliefs and social marketplace. There is a chance that mobile speakers are more aware of the indexical meanings of their language, and are more likely to be innovative. Furthermore, in particular reference to those cases where speakers move, evidence shows change to be individually variable and variable between features. I these cases speakers are more likely to lose non-standard features from their origin region than adopt those from a host region (Nycz, 2011), and are able to adjust current vocalic productions towards the host accent type. However, much of this evidence is based on small samples and features which are highly marked and not widespread in the UK.

2.1.2.9 Forensic implications

The forensic implications of the issues outlined in the previous sections should not be overlooked. There are two main problems to consider.

Firstly, given that speakers who are geographically mobile are routinely excluded from studies of language, we know very little about the effects of mobility on accent and speech, because of the assumption that language fossilises at adolescence. This holds true for a number of other factors summarised above, such as mental health issues and substance abuse. Forensic analysts are expected to make informed hypotheses about the speech patterns of suspect speakers. However, without the proper research in place relating to these factors it is not currently possible to do this as fully and accurately as possible.

Secondly, and perhaps more importantly, if speakers who are mobile, have physical or mental health issues or have suffered a significant life event are excluded from sociolinguistic study, should this preclude us from using data from those studies to inform forensic casework? In a total population of 62 million, if 1.3 million people are moving around the UK, and more than 1.2 million are in mental healthcare *each year* (as an example of just two factors for which we know relatively little about), our model of speech [based on results from these studies] is not satisfactorily representative. It is explained further in §2.3 that forensic evidence must be expressed with reference to data in relevant reference populations. This can take place either theoretically (i.e. by the expert applying their knowledge and general research findings) or directly, by means of using reference population data to express the strength of evidence numerically. If data from sociolinguistic studies are limited to that taken from 'ideal' speakers, then there is a logical problem in deriving statistics or general principles from these data and applying this to speech evidence from speakers who exhibit mobility or other factors mentioned. Moreover this is especially true for criminal populations, where prevalence of factors which impact on speech, such as drug or alcohol abuse, is higher. This is not so damaging for descriptions of dialects generally which may inform a case, but where strength of evidence assessments of acoustic or auditory measurements (whether through statistical testing or by an expert's judgement, see §2.3.2) are based on a non-representative sample, this will necessarily impact on the quality and reliability of the evidence given.

Certainly analysts should be cautious of the restrictions placed on sampling in these kinds of study, and give courts and jurors as full a picture of these limitations as possible.

Scientific studies must sample a small part of the wider population to make broader conclusions, but if these sampling methods are partially exclusive then there should be concern in using them to assess the strength of forensic evidence. The DNA database, for example, cannot currently hold a record of the entire population's DNA, but it does hold representative samples of groups which may express different characteristics (i.e. ethnic groups). If our models of speech production, and the data we use to assess the strength of evidence, are based on samples which routinely exclude factors which may affect speech, then these are not reliable measures. That is not to say that these studies should not inform casework (there are always limitations to the level of scrutiny we can put on reference data) just that a critical approach to using the data is needed, and the limitations should be transparent and available to triers-of-fact.

2.1.3 **Longitudinal forensic research**

Owing to the forensic situation, research in forensic speech science has to take account of non-contemporaneous speech samples. However, there are two perspectives which are, although it may not be possible to completely distinguish between them, of differing relevance to the present study. Principally, there is research which has tried to account for the longitudinal change in speakers' linguistic patterns; this research generally accounts for non-contemporaneity of greater than 1-2 years, which is more relevant to the present study. The second perspective, while it may not be of such direct pertinence to this thesis, is highly relevant to all speech science in general and especially forensics. That is, the acceptance that speech is highly variable at an individual level, and that no speaker ever produces the same utterance in exactly the same way (Nolan, 1983; Rose, 2002). To try and account for this when investigating the effectiveness of certain parameters in discriminating between speakers, forensic research has employed samples which are non-contemporaneous to the degree of a few weeks, months or even up to a year or so.

2.1.3.1 'Short-term' non-contemporaneity

Short-term non-contemporaneity of this type, between a few days, weeks or months, is still important to forensic work, so the kinds of study that have taken place are illustrated

in brief, using Rose et al. (2003) as an example. The aim, as in many of these studies, is not to investigate the effects of non-contemporaneity itself but to examine the discriminatory strength of a certain variable. In this case (Rose et al. (2003)) the authors are concerned with calculating the possible strength of evidence generated from limited vowel analysis as well as cepstral measures for a multi-speaker analysis (n=300). Non-contemporaneous recordings, in this case of around 3-4 months, are simply treated as the standard data for forensics, reflecting the potential time gap between evidential and suspect samples and also the potential for intra-speaker variation in any case of recorded speech. For Rose et al. (2003) and Morrison (2011), short-term non-contemporaneity in recordings is vital to capture the intra-speaker variation that we all exhibit from day-to-day and month-to month. The magnitude of this recommended delay is not stated, however, and there is little research comparing length of short-term delay effects. Researchers have attributed these potential differences to many factors from health/lifestyle changes (Künzel, 2007) to time of day of recording (Rose, 2002) to maybe most importantly, the complexity of articulatory movements and coarticulatory processes which makes every gesture different (Nolan, 1983). Importantly, these changes may not reflect overall trends (of change) in the individual across time, but are merely a function of the flexibility of the linguistic and vocal system.

A study by Rose (1999) investigated the difference between this kind of day-to-day variation and what he calls 'longer-term' variation over one year in respect to intonation and the first four formants. He concluded that there was very little difference in the two kinds of non-contemporaneity.

2.1.3.2 'Long-term' non-contemporaneity

However, for the purposes of this study and in respect to cases such as the Yorkshire Ripper Hoaxer (R v John Samuel Humble, 2005), 'long-term' is taken to refer to periods generally longer than one year, especially in light of the findings of other RT studies. One of the first long-term studies of interest from a forensic perspective was carried out by Endres et al. (1971). They took formant and fundamental frequency measurements from four male and two female speakers at five year intervals over a period of around 14 years (male speakers were 42 years or older in the first recording and females 29 or older). They found that formant frequencies decreased over time (for six monophthongs and two diphthongs) and that mean F0 and the variation of F0 also decreased with age. They

attributed this to the degenerative effects of aging (see further §2.1.4.5). This study was, however, carried out from the point of view of using voiceprints to characterise speakers, a method that is widely discredited (Nolan, 1983); despite this fact, the measurements are still relevant to formant-based FSC.

Suzuki et al. (1996) carried out a study of similar variables, including cepstral coefficients, for eight male speakers across a 20 year delay. They found that change was extremely variable and individual but that formants remained relatively stable, F0 decreased in 6/8 speakers and cepstral coefficients were raised slightly. This shows moderately less change than Endres et al. (1971) and also highlights a point raised from sociolinguistic studies: that change is individually variable.

Künzel (2007) carried out a study of non-contemporaneity in regard to both auditory and automatic speaker identification over 11 years in answer to following question, which forms the basis of part of this study:

> Is it possible that physiological aging or other factors have caused alterations of parameters of voice, speech and language to such a degree that the material basis for identification may no longer be regarded as accurate when a new identification task is due (Künzel, 2007, p. 110)

He found that likelihood ratios (LRs) from an automatic speaker recognition system (*BATVOX*) were all correct despite a short-term delay, with LR magnitudes between 10^2 and 10^8. Unfortunately, there is no control sample to compare these scores with to observe what effect the delay itself has on the strength of the LR, although they clearly perform well. After 11 years, he claims that aging had almost no influence (although one speaker shows much lower performance).

Künzel (2007) also makes important observations about the effects of aging and non-contemporaneity on speech. He highlights the important notion that the "magnitude of the time delay is an important parameter, the greater the delay, the greater the likelihood for any of the related factors to occur, irrespective of the fact whether this be sudden or gradual" (Künzel, 2007, p. 111). These related factors include certain temporal changes, such as smoking and drinking habits (especially repeated overuse), surgery, disease, dentition and of course the natural degenerative process of aging, alongside many more. The paper also points out that most forensic cases are not straightforward,

and that these effects are also complicated further by the typical characteristics of the forensic condition, for example non-cooperation, disguise or channel effects.

Kelly and Harte (2011) use an automatic GMM-UBM (Gaussian Mixture Model – Universal Background Model) ASR system to test the effects of aging on speaker verification. ASR systems use Mel Frequency Cepstral Coefficients (MFCCs) to try and remodel the vocal tract, and use that a biometric identifier (for more detail see §3.4.2.7). Using a database of 13 speakers across a 30-40 year delay, the researchers found that variability and test scores beyond ten years generally fell outside normal inter-session variability for speakers. Therefore, speakers' voices were becoming unreliable for testing in a verification situation after ten years, as the variability was falling outside the parameters set for that speaker at 'enrolment' in the system. In further testing (Kelly, Drygajlo, & Harte, 2012) they applied this model to a forensic setting (backwards testing, i.e. the aging effect is from oldest to youngest sample, modelling suspect and evidential sample) as well as a verification setting (forwards testing, i.e. testing the normal chronological effects of aging), and found that LLR scores gradually declined, with increasing error rates. A similar period of around 10 years was found to be the limit of performance. Given the development of ASR systems and their emergence in the forensic market, it would be useful to examine the effects of age on this system for the present study and be able to compare the change in performance with that of the acoustic features.

2.1.4 **Physiological research**

Longitudinal sources from forensic and sociolinguistic research are clearly limited. There is, however, a wealth of research carried out from a physiological perspective, as part of the explosion of gerontological research from the start of the 20th Century (Linville, 2001). While a proportion of this research is more focussed on the aging process rather than linguistic production, there are also directly relevant findings on changes to articulation and acoustic output. What this research does do is to provide a scientific basis from which a hypothesis can be formed on causes of changes in the voice and its acoustic output.

This section first presents more general findings relating to the aging process and the voice (a number of these are not longitudinal, but are still useful) and subsequently illustrates relevant longitudinal studies which have taken place in the context of gerontology. One concern is that some of this work is concerned with the health and

social implications of elderly speakers' vocal changes, which is not relevant to the subjects in this study. In order to maintain clarity, the phrase 'elderly' is preferred for speakers in the research aged around 70 years and over, where physiological literature uses 'old' or 'older'. In this context, 'older' will mean beyond young adulthood. Terms such as 'young adulthood, adulthood, and middle age' are used and, wherever possible, the age or at least age decade of the subjects is presented.

2.1.4.1 Aging and the voice

Within this subsection biological aging factors are described and the potential effects that these processes have on the voice are elucidated. Most of this work is collated in two comprehensive overviews of vocal aging by Linville, encapsulating both the processes of senescence (2001) and their effects on measurable features of speech (1996; 2001).

Before inspecting the individual changes, it is worthwhile to look at the general theories of aging. Linville (2001) summarises a dichotomy between two of the key models of aging: 'purposeful events' and 'random events' theories. Within the former, aging is seen as programmed within the individual and dictated by predetermined events with concomitant changes in hormonal or genetic programming. This idea has also been further expanded by Birren and Schroots (1996) who have coined the phrase 'gerodynamics'. Citing that at critical life stages, the individual does not just undergo changes in biology, but also behavioural patterns and social functioning, which are reflected in observable patterns (such as speech). The latter, 'random events' theory sees aging as less coordinated, more as the consequence of the cumulative effects of random 'events'; essentially wear and tear to the body. Sataloff et al. (1997, p. 157) postulate that whatever theory is followed, the notion that this decline occurs "gradually and progressively (linear senescence) is open to challenge". The effects of aging are, therefore, not always predictable and stable and knowing the magnitude of the time delay may not allow us to predict the specific effects of aging. We have also seen from studies mentioned above that this change is very variable between individuals. The following research demonstrates that, as can be expected, both theories account for change across the lifespan: as critical events and also as the effects of 'wear and tear'.

With respect to the physiology of aging regarding the vocal apparatus specifically, Linville (2001) identified two main aging processes (in bold). Within the research collated in her book there are four identifiable sub-processes which also play a prominent role:

1. **Loss of tissue elasticity**
    a. **largely due to calcification or ossification of the tissue**
2. **Muscle weakening**
3. Muscle/tissue atrophy
    a. as a result of age-related changes in neuromuscular function
    b. hypertrophy also occurs in some structures
4. Fibrosis
5. Loss of blood vessels and decrease in their efficiency
    a. numbers of vessels decreases
    b. transfer rate decreases due to thickened walls
6. Decline in sensitivity due to degenerating innervations
    a. neuromuscular junctions (NMJ) (responsible for much neural 'traffic' between muscular tissue and neural network) need to grow and regenerate, these processes are all limited with age
    b. 'motor units' decline in numbers, become more forceful to compensate, therefore are larger and slower – increased force at the expense of fine motor control
    c. general decrease in brain matter

The following subsections summarise the effects of these processes on specific parts of the anatomy involved with the production of speech. The effects of these changes are explored in the subsequent section (§2.1.4.5 onwards).

2.1.4.2 Respiratory system

Overall lung function is known to increase from childhood to adolescence, where it increases rapidly with growth, at which point it plateaus at young adulthood (20 years for women, 25 for men) and then declines throughout life (Linville, 2001). There are lifestyle factors which influence this. For example, smoking accelerates the normal aging process to the lungs by enhancing the oxidative damage to the tissue of the lungs.

Three main aging processes influence respiratory change:

1. Decreased lung elasticity
2. Decreased strength of respiratory muscles
3. Increased stiffness of the thoracic cage

Their effects are shown in Table 10 below:

Table 10 - Summary of aging effects on the respiratory system

Respiratory structure/process	Nature of aging change	Reason
Lung size	Reduced	Degeneration and long-term dehydration
Blood transfer	Less effective	Vessel wall thickening, and lining becoming fibrotic due to collagen deposits or hyalinisation
Alveoli	Fewer functional, less effective	Degeneration, alveolar ducts widened
Bronchioles	Less effective	Bronchioles widened
Position of lungs/bronchi	Lowered in the thoracic cage	Weakened support structures due to thinning of intervertebral discs and reduced vertebral height
Tissue elasticity	Reduced	Stiffening as part of changes to elastic recoil characteristics in the connective tissue of the pulmonary system
Thorax	Reduced flexibility, changed shape	Ossification, can become bowed in old age due to vertebral change (known as kyphosis or 'dowager's hump'), although conversely can also increase in diameter, reported as 'barrel chest'
Costal cartilages	Reduced flexibility	Ossification/calcification
Diaphragm	Less powerful, more fatigue suffered in old age	Atrophy/degeneration

Source: Linville (2001)

2.1.4.3 Laryngeal system

Changes in the larynx are generally more noticeable for men. Moreover, onset of aging changes is generally earlier for male speakers (starting in the 30s and mostly noticeable by 50-60 for males, starting at 50s for females). Although the specific parts of the vocal folds age in slightly different ways, the principal overall effect is that they shorten and stiffen. This process of ossification or calcification is likely to have significant effects on the voice, as the glottis acts as the source for speech (Fant, 1960). The specific differences are laid out in Table 11 below:

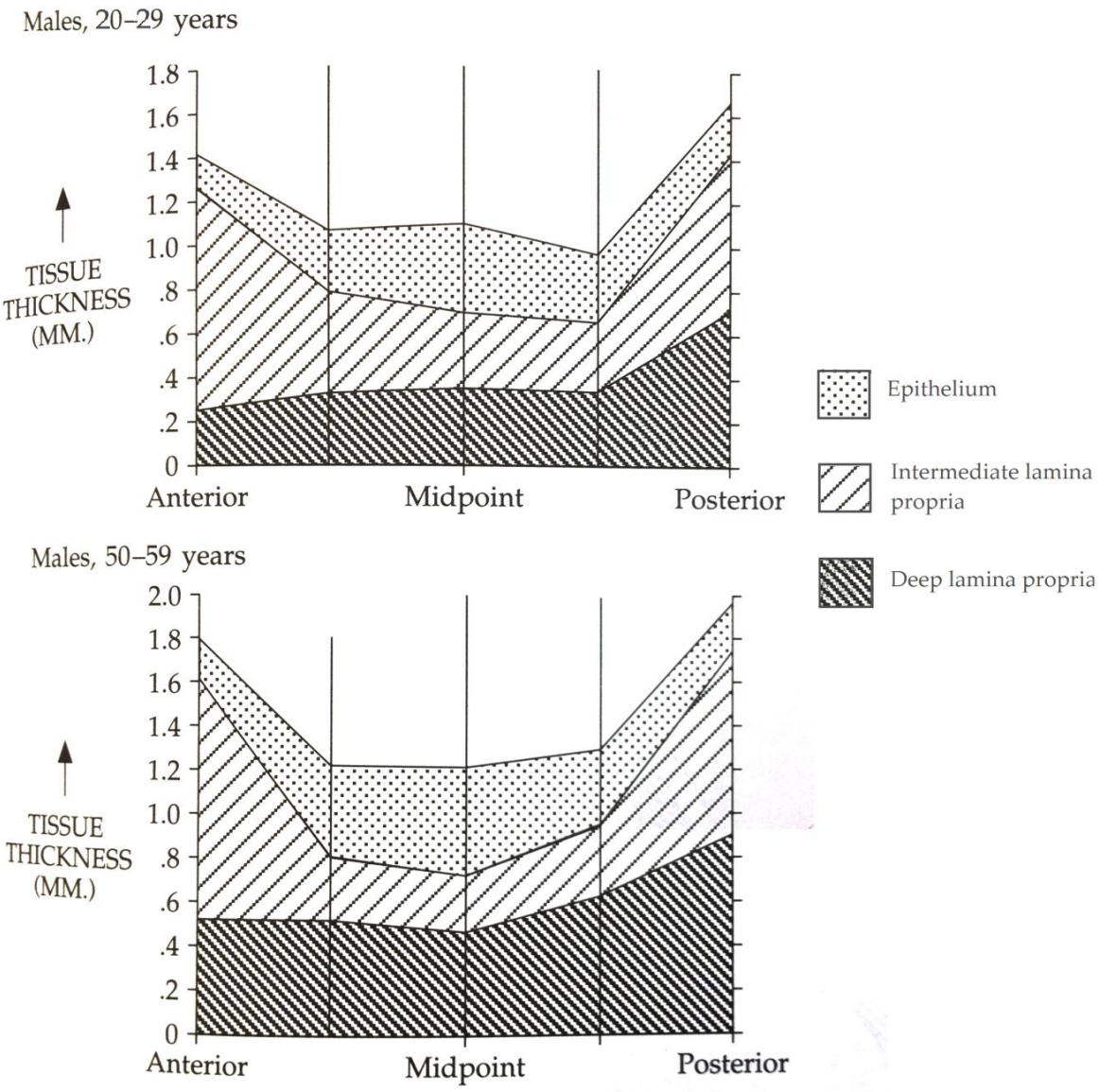
Table 11 - Summary of aging effects on the larynx and gender differences

Laryngeal structure	Nature of aging change	Gender difference
Cartilages	Ossification/calcification	More extensive with earlier onset in males
Cricoarytenoid joint	General deterioration	More evident in males
Intrinsic muscles	Atrophy	Reported in males, limited data in females
Epithelium	Thickening Fatty degeneration or keratinisation (Honjo & Isshiki, 1980)	Progressive for both More marked in males
Mucous glands	Degeneration	Reported in males, no female data
Lamina propria (mucous lining)	Degeneration, stiffer more viscous tissue	More evident and extensive in males
Macula flava (join between vocal folds)	Degeneration of fibroblasts	No data on difference
Conus elasticus (lateral cricothyroid ligament)	Thinning of fibres, fragmentation of fiber bundles	Few changes observed in women
Thyroarytenoid muscle	Atrophy/degeneration	No data on difference
Innervation	Some evidence of disruption, conflicting data reported	No data on difference
Vascular supply	Reduced blood vessel diameter, thickening of capillary walls	No data on difference

Sources: Linville (2001), (Beck, 1997)

Significant changes affecting acoustic output are those to the vocal folds and its component tissue layer structures; importantly the relative thickness of the layers, which is demonstrated for the male subject age range of this study by the figure below.

Figure 5 - Age related changes in the tissue layers of the vocal folds in males



Source: Beck (1997) based on data from Hirano et al. (1982)

2.1.4.4 Supralaryngeal system

The supralaryngeal system consists of the articulators and therefore is particularly important with respect to the features of speech that are typically measured in a forensic setting. The effects of aging are summarised in Table 12 below:

Table 12 - Summary of aging effects on supralaryngeal system

Supralaryngeal structure	Nature of aging change
Facial skeleton	Symmetrical enlargement across life, 3-5% from 30-50
Facial muscles	Loss of tone, blood supply & elasticity; atrophy, breakdown of collagenous fibers
Temporomandibular joint	Reduced mandibular height, placing the articulating surface at a lower level (affects women more than men)
Maxilla (upper jaw) and palate	More width and length in males (pubertal growth). Thicken in older age
Mandible (lower jaw)	Growth generally in accordance with body growth
Mucous layer and salivary glands	Reduce performance in old age – leaves mouth more susceptible to microorganisms, causing mucosal inflammation
Dentition	Hardened enamel surface and discolouration. Cracking of enamel
Tongue	Fissuring on epithelium (surface), muscle atrophy, declined motor skill
Pharynx and soft palate	Thinning of epithelium, muscle weakening (dilation of pharynx), increased variability in diameter (due to atrophy and compensatory hypertrophy)

Sources: Linville (2001), Beck (1997)

It is worth noting that there is relatively little research about the nasal cavity and its changes throughout adult life, which has a significant influence on acoustic output. This is made even more complex by the fact that "the properties of the oral-nasal port will be affected by the lumen of the pharynx, the size and carriage of the tongue and the mass of the lymphoid tissue" (Beck, 1997, p. 274), all of which undergo concurrent and varying age related changes. There are other symbiotic processes, such as movement or [limited] growth in the mandible and maxilla, which generally develops simultaneously to ensure correct dental occlusion.

2.1.4.5 Changes in articulation

Concomitant with these physiological processes are their effects on articulation, which have a subsequent impact on the characteristics of speech produced by older speakers. A number of these findings relate more specifically to elderly speakers, but the processes are ongoing throughout adulthood; the two main general changes are *imprecision* and *slowing of articulatory gestures*. Table 13 below highlights changes to articulation with aging:

Table 13 - Summary of aging effects on articulation, with evidence from different research methods

Articulation change with aging	Manifestation
Acoustic evidence	
Slower tongue movements	Smaller rate of frequency change along formant transitions
Deterioration of muscle control	Longer stop closure intervals
Less extensive lip and tongue movements	Noisy stop closure intervals, spirantisation
Loss of full range of movement in cricoarytenoid joint Reduced maximum opening of glottis during devoicing	Shorter voiceless interval for voiceless consonants
VOT findings are conflicting	Some shorter VOT, VOT distinction compressed
Ultrasound evidence	
(possible) Decline in oral motor precision	Subtle differences in tongue movement and position during isolated phoneme production
EMG (electromyographic) evidence	
Decline in lip muscle activity	Reduced average peak EMG activity
Decreased flexibility of fine motor control	Higher correlation of EMG activity among sites surrounding lips, increased lip muscle coupling.
Kinematic evidence	
Reduced spatiotemporal stability	Higher STI (sum of SDs of kinematic waveforms from multiple utterance repetitions)

Source: Linville (2001)

Of course, a number of these tests are not relevant to forensic cases, where it would not be possible to use ultrasound or EMG with a suspect, let alone at the point of an evidential recording. What we are interested in is, for example, acoustic evidence of spirantised or longer stop closures (although longer stop closures would be expected only in long-term cases which featured an elderly speaker). The effects of these processes are discussed in light of evidence based on speech characteristics in §2.1.5.

2.1.4.6 Changes to speech characteristics

These aging and articulation effects are posited to have a significant effect on the speech of the individuals. This section relates these findings regarding speech output, presenting causes and magnitudes of possible changes in features of speech.

2.1.4.7 Fundamental frequency

For male speakers, F0 lowers from young adulthood to middle age (from 20-40 yrs (Hollien & Shipp, 1972)) and rises from that point into old age. The drop in F0 from age 20 is not explained in detail, although Hollien and Shipp suggest that it is due to 'subclinical trauma' associated with day-to-day voice use. The rise in F0 in older age (post 60yrs) is predictable from anatomical changes in the larynx with aging i.e. atrophy of muscle tissues and ossification of vocal folds and cartilages.

Figure 6 - Scatterplot of male speakers' F0 (N=175). Solid line connects mean for each decade

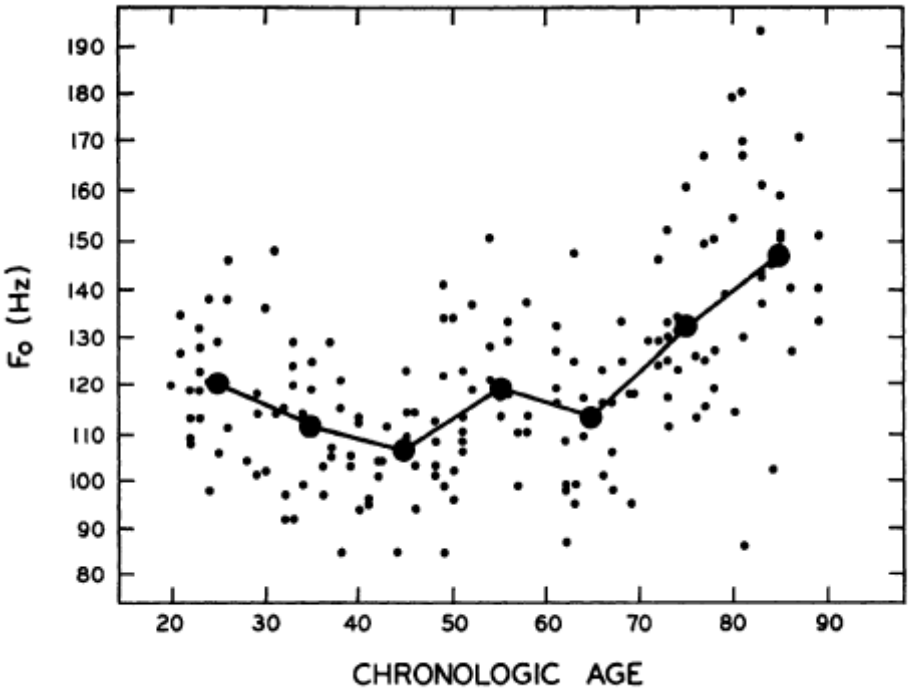
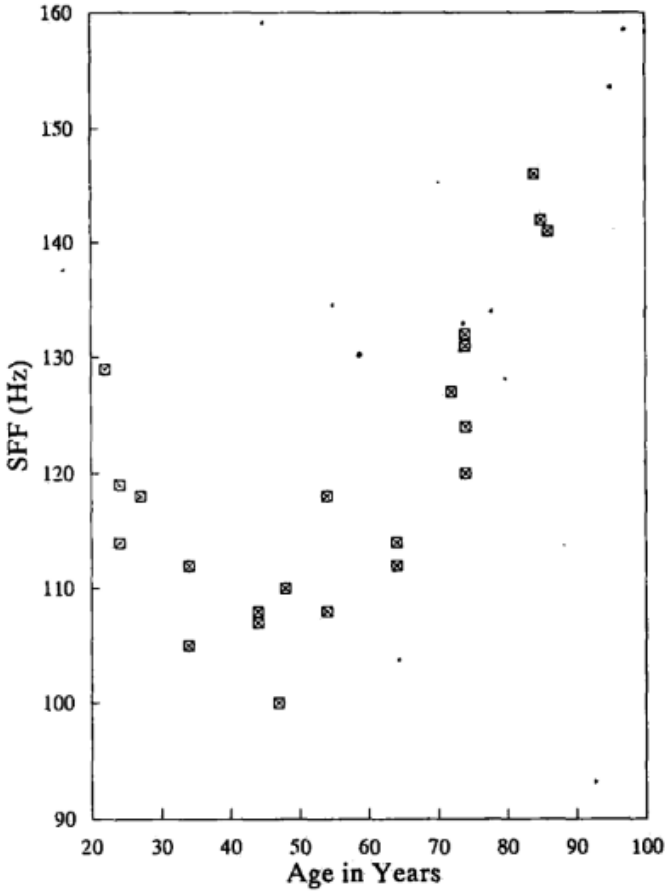


FIGURE 1. Scatterplot of 175 subjects' mean SFF. Solid line connects mean values for each age decade.

Source: Hollien & Shipp (1972)

Data in Figure 6 above show this pattern in 175 male speakers recorded by Hollien and Shipp (1972). Longitudinal data from DeCoster and Debruyne (2000) corroborate across a 30 year period comparable to this study, finding a 14Hz mean drop in F0, although this was variable across speakers. Figure 7 below presents data from five studies which also support this finding:

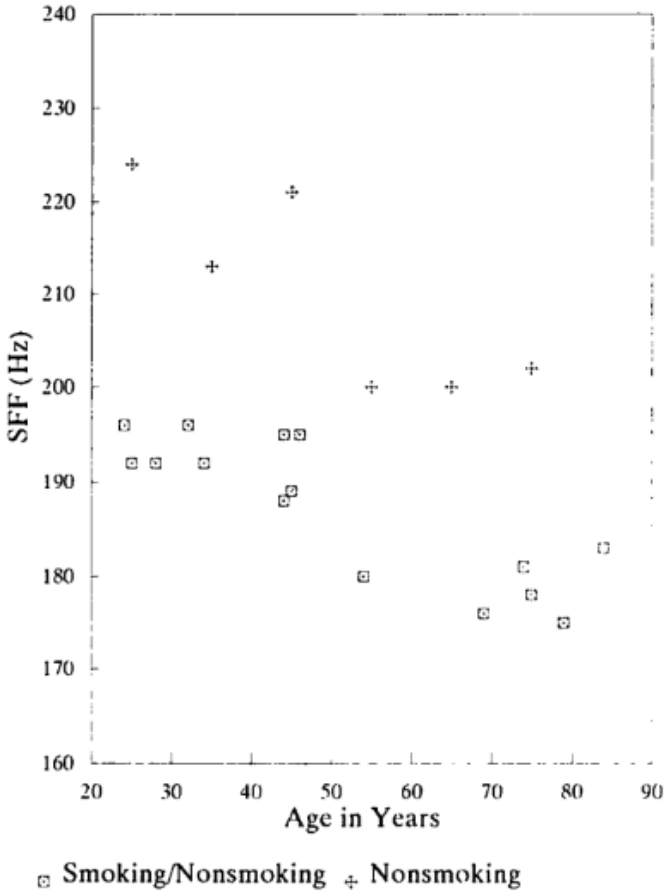
Figure 7 - Summary of data on speaking F0 as a function of age for male speakers from five studies



Source: Linville (2001)
Data: Mysak (1959), Hollien & Shipp (1972), Hollien & Jackson (1973), Krook (1988), Brown et al. (1991)

For female speakers the pattern is quite different; F0 remains fairly stable throughout adulthood until around 50-60 years, when a drop in F0 occurs, cited as being due to the menopause, as hormonal changes can result in vocal cord edema (Kahane, 1983). This F0 remains fairly constant in older age, although can increase at very old age. Figure 8 below presents speaking F0 data for female speakers, as well as indicating the effect that smoking can have on fundamental frequency.

Figure 8 - Summary of data on speaking F0 as a function of age for smoking and non-smoking female speakers from five studies



Source: Linville (2001)
Data: Saxman & Burk (1967), Honjo & Ishiki (1980), Stoicheff (1981), Krook (1988), Brown et al. (1991)

Although these data demonstrate a drop in F0, Linville (1996) suggests that longitudinal data (DePinto & Hollien, 1982; Russel, Penny, & Pemberton, 1995) show a much steeper decrease in F0 after menopause than is suggested by the cross-sectional data in Figure 8. What is clear from Figure 8 is the effect of smoking, which appeared to lower the F0 of subjects across all ages; this effect is also consistent for male speakers (Verdonck-de-Leeuw & Mahieu, 2004). It has been suggested that although cigarette and other types of smoke cause edema of the vocal folds, these effects are somewhat reversible (Verdonck-de-Leeuw & Mahieu, 2004). Furthermore, it has been demonstrated that these effects are reversible even at a period of 40 hours of non-smoking (Murphy & Doyle, 1987). There are effects of smoking, therefore, that are relatively short-term and smoking within a day of being recorded could have affected the data in Figure 8 significantly. It is also unclear what the long-term effects of smoking are if the drop in F0 is reversible in both the short- (Murphy & Doyle, 1987) and the long-term (Verdonck-de-Leeuw & Mahieu,

2004). However, what is clear is that whether a subject smokes or not is an important piece of information in research or in forensic practice.

Another corollary of age is increased variability of F0, due to a combination of the inaccuracy of the older larynx because of degeneration, ossification of musculature and connective tissue and respiratory changes, such as diminished elastic recoil and reduced vital capacity (Kaltieider, Fray, & Hyde, 1938; Mittman, Edelman, Norris, & Shock, 1965; Pierce & Ebert, 1965). Linville and Fisher (1985) and Orlikoff (1990) all found massive increases in F0 standard deviation in older age. SD in old age was almost double that of younger speakers for males (Orlikoff, 1990) and showed a 71% increase for females (Linville & Fisher, 1985). It is also reported (Beck, 1997) that fatty degeneration or keratinisation of the vocal fold epithelium can cause dysperiodic vibration, perceived as 'harshness' (Honjo & Isshiki, 1980) (although this is largely in older age).

2.1.4.8 Jitter and Shimmer

Jitter is a measure of cycle-to cycle fluctuations in the fundamental period of vocal cord vibration, whereas shimmer is cycle-to-cycle variation according to waveform amplitude (Orlikoff, 1990). Orlikoff (1990) found that for older male speakers, jitter and shimmer variability was much higher. However, after normalising for health and fitness patterns in his subjects, only differences in shimmer were still significant. These patterns are found in much older speakers than those in questions in the present study, however. Jitter and shimmer measures may be used to supplement a systematic voice quality analysis, but there are a number of factors which could affect these measures (Orlikoff, 1990) and without more detailed information about health and fitness it appears these parameters would not be useful. These measures were also found to be significantly different in two different age groups (Vipperla, Renals, & Frankel, 2010) used for testing an automatic speech recognition system used for dictation.

2.1.4.9 Vocal cord closure

Another more subtle feature of the aging voice is an increase 'glottal gaps', which have a couple of consequences for speech production. Glottal gaps are an incomplete closure of the glottis during phonation, attributed to atrophy of the intrinsic laryngeal musculature or atrophy of the connective tissue in the region (Linville, 2001). Weakening in different parts of the larynx can cause different types of gap, for example:

- Thyroarytenoid weakening can cause an anterior gap
- Inter-arytenoid weakening can cause a posterior gap
- Adductor weakening can cause incomplete closure along the length of the glottis

Research has shown that for younger men, glottal gap occurs at around 20-38% of closures (Bless, Biever, & Shaik, 1986; Soderston & Lindestad, 1990) and at around 67% for an older group (Honjo & Isshiki, 1980). For women, however, age has not been shown to have a significant effect (Linville, 1992) and in one study, older women made more complete closures than the younger group (Higgins & Saxman, 1991).

This can have three main effects on speech. Firstly it is posited that glottal gaps cause speakers to make fewer syllables per breath, possible decreasing speech rate and increasing the respiratory load. Secondly, incomplete closure, along with irregular glottal cycles, can cause spectral noise by creating turbulent airflow. Finally, Higgins and Saxman (1991) suggest that speakers use compensatory strategies which include increasing laryngeal adductory force, resulting in a strained voice quality. Although Hillman et al. (1989) put forward that this increased pressure, or 'vocal hyper-function', is merely an indicator of stiffening of the vocal folds.

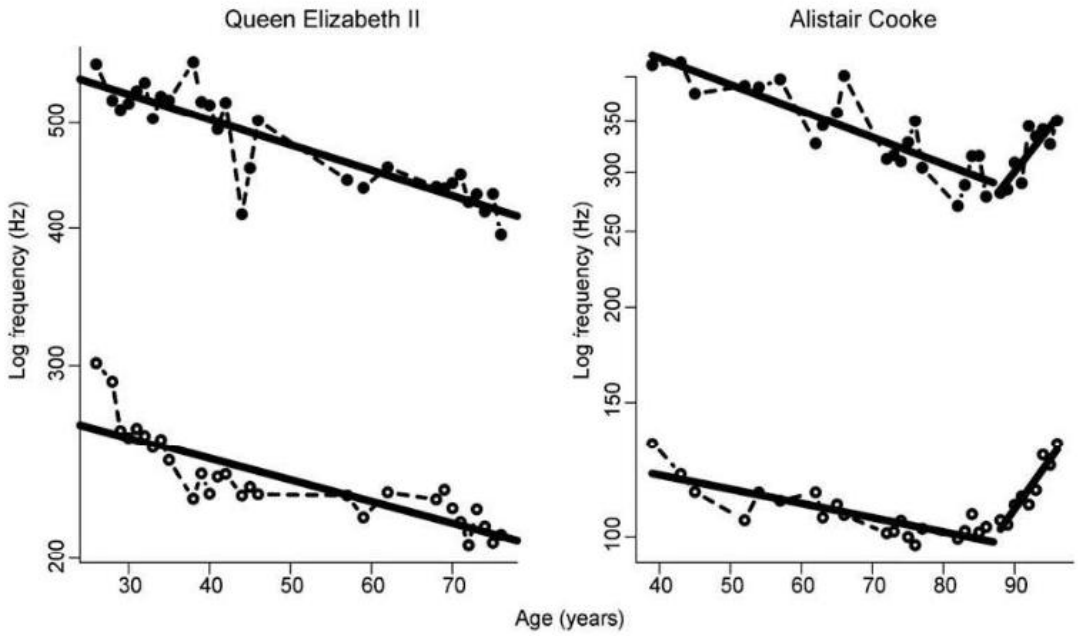
2.1.4.10 Formants

Limited information on formant changes with increasing age has been reported in §2.1.3.2. For example, Endres et al. (1971) demonstrated that formants lowered in ten monophthongs and two diphthongs in males and females, a finding corroborated for women by Linville and Fisher (1985). Linville and Rens (2001) used long-term average spectra (LTAS), a technique used in forensic research which captures average frequency measures across a recording, to investigate a more holistic picture of formant characteristics of aging male and female speakers. They found that F1 decreased over time in both male and female subjects, and that F2 and F3 lowered significantly in women and showed a tendency to decrease in men. This corroborates other findings showing an overall reduction or compression of the vowel space. It has been suggested by Reubold et al. (2010) that this reduced F1 could be caused by restricted jaw opening; this would be plausible as movement of the temporomandibular joint becomes more restricted with aging (Kahane, 1980).

It has also been suggested that the overall pattern of decreased formant frequencies is due to an extension of the vocal tract (Wind, 1970), defined by Ferreri (1959) as 'laryngeal ptosis' and possible extension of the facial skeleton. The source-filter theory of speech production (Fant, 1960) dictates that increasing the length of the filter decreases formant frequencies. Wilder (1978) hypothesises that the larynx might lower in the neck with advanced age as a result of stretching of ligaments and atrophy of the strap muscles of the neck, thus lengthening the vocal tract. Laver and Trudgill (1979) and Linville (2001) also speculate that as the lungs and bronchi lower in the thoracic cage, this causes all the vocal apparatus to descend. However, Flügel and Rohen (1991) did not find any lowering of the larynx during later adulthood using computed tomography (CT) scans. Xue and Hao (2003) utilised an acoustic reflection technique to measure the vocal tract length of young (18-30yrs) and old (62-79yrs) speakers and found no difference in overall vocal length, although older speakers had a slight increase in vocal tract volume and length of the oral cavity specifically (of around 0.5-1cm). Studies of this kind with longitudinal data would provide further useful explanation.

Reubold et al. (2010) set out to test a different perspective, suggested in Harrington et al. (2007). They investigated how age-related F0 changes and formant changes might be related. Using longitudinal data from five speakers, they looked at the relationship between F0 and F1 and found that the rate of change (calculated as the logarithm of the parameter as a function of the year) for F1 roughly tracked that of F0; this was tested and was not merely an acoustic artefact of reduction in F0. Figure 9 demonstrates this correlation:

Figure 9 - F0 and F1 averages at different age stages for two subjects with regression line superimposed



Source: Reubold et al. (2010)

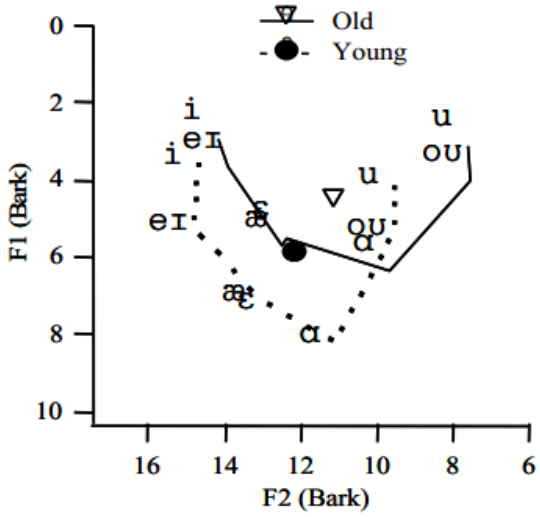
They conclude that changes in F1 may be compensatory to offset a physiologically induced decline in F0, maintaining a relatively constant auditory distance between F0 and F1. They argue that as Tranmüller (1981) has demonstrated that listeners use the difference between F0 and F1 as a cue to vowel height, speakers might be attempting to preserve this distance. They postulate that jaw opening could be manipulated simply to affect F1, as shown by Lindblom and Sundberg (1971), and demonstrated, using a filter, that this would have a small effect on F2 and F3. The male speaker, Alistair Cooke, in Figure 9 above certainly demonstrates an unprecedented rise in F1 which tracks the expected rise in F0, which might support their argument.

Further to this overall reduction, Ratstatter and Jacques (1990) and Ratstatter et al. (1997) illustrated that the vowel productions of older speakers were more centralised than younger speakers, so that as well as decreased formants overall, the speaker's vowel space was de-peripheralised. This finding may have been missed in studies which used more holistic methods and did not investigate differences between different vowels where, for example, although F2 was decreased overall, it may have actually increased in back vowels because of centralised tongue positions. However, much of this research was carried out using carefully elicited carrier phrases, and Benjamin (1997) shows that in elderly speakers, vowel variability is high and speakers' vowel spaces are expanded. This

study examines the effects of aging on formants and the vowel space using real speech across adulthood.

Watson and Munson (2007) offer an alternative hypothesis for changes over time. Although their data is not longitudinal, but comes from different speakers at different ages, it does illustrate the patterns of reduction in formants found throughout the literature:

Figure 10 - Mean formant frequencies for older and younger groups



Source: Watson & Munson (2007)

Differences in formants, assessed by MANOVA, were significant for all vowels in F1, and less so in F2, where only three back vowels were different. What is novel about this study is the explanation given for age-related changes. They put forward that low phonological neighbourhood density and changes in lexical frequency might be at the root of changes, as speakers are shown to produce unusual words or phonotactic patterns with more extreme articulatory movements (presumably to produce clearer realisations). However, differences between speaker groups were not apparent. However this study does support the notion of differences in formants for older and younger speakers, and testing with real-time data could explore the hypothesis further.

In terms of formant transitions, which are relevant to the formant dynamic aspect of the present project, Liss et al. (1990) demonstrated that transitions in elderly speakers were much less extreme, reflecting more conservative movements of the tongue. However, these subjects were much older than those in the current study, and the effects of aging across adulthood on extent of formant transitions are not known.

Linville (1996) identifies the need for close analysis of vowel behaviour across aging, especially looking at the changes between individual vowels to indicate variations in articulatory setting; this project examines these factors.

2.1.4.11 Consonantal changes

There are effects of articulatory degradation on consonantal realisation. Benjamin (1997) demonstrated that older speakers have longer stop closures than for a younger group, and for very old speakers, this poor performance in stop closure leads to high rates of spirantisation (Liss, Weismer, & Rosenbeck, 1990). Increasing imprecision and muscular control also affects voice onset time (VOT) for aging speakers; although findings in cross-sectional studies are conflicting (Linville, 2001), DeCoster and Debruyne (2000) showed a significant increase in VOT for two segments in a longitudinal study of 20 speakers across 30 years (mean 33-63yrs). For [pa], VOT increased by mean 5.8ms and by 9.2ms for [ka].

2.1.4.12 Speech rate

Glottal gap is not the only cited cause for decreased speech rate in older age; lower speech rate in older speakers is well documented, the following have all been cited:

- Neuromuscular slowing (Hartman & Danhauer, 1976; Ryan, 1972)
- Degeneration of the respiratory system (Oyer & Deal, 1985; Ramig, 1983)
- Physiological condition (Ramig, 1983)
- Increased cautiousness/expectations of society (Mysak, 1959; Mysak & Hanley, 1959)
- Fatigue (Hollien & Shipp, 1972)

Source: Linville (1996)

Linville et al. (1989) also correlated slower reading rates with the reductions of F0 stability that was mentioned in §2.1.4.7, attributing both to a general loss of physiological control. Other research has also attributed reduced speaking rate to respiratory factors, such as reduced maximum intensity level (Linville, Skarin, & Fornatto, 1989), respiratory difficulty and reduced elasticity and recoil of lung tissue (Kaltieider, Fray, & Hyde, 1938; Mittman, Edelman, Norris, & Shock, 1965; Pierce & Ebert, 1965). It is, however, worthwhile to remember that, particularly for speech rate, some of these concern only elderly speakers and may not be relevant to the present study.

2.1.4.13 Vocal resonance and articulatory setting changes

It has been posited that there are changes in general vocal resonance and articulatory setting due to changes in the supraglottal vocal tract throughout adulthood (these changes are summarised in §2.1.4.4). The following changes have been reported in the physiological literature:

- 3-5% growth of the facial skeleton (Israel, 1968; Israel, 1973; Lasker, 1953)
- Atrophy/hypertrophy of tongue musculature (Balogh & Lelkes, 1961; Cohen & Gitman, 1959; Silverman, 1972)
- Atrophy of pharyngeal musculature (Zaino & Benventano, 1977)
- Restricted movement of the temporomandibular joint (Kahane, 1980)
- [Potential for] tooth loss (Meyerson, 1976)

Source: Linville (1996)

It should be noted that although Linville (1996) noted tongue and pharyngeal atrophy play a significant role in changing the overall vocal resonances, this process is noted particularly in elderly speakers and may not be relevant to the current project. While the result of some of these individual changes has been speculated upon (i.e. facial growth may affect formant frequencies), it is the speaker's own experience of these changes to the vocal tract that affects speech. That is, that older speakers may respond to these changes and "systematically alter their articulatory positioning to compensate...thus affecting the resonance characteristics of their speech" (Ratstatter & Jacques, 1990, p. 318). The idea that speakers compensate for age-related change was also put forward by Reubold et al. (2010) with respect to F1 changes. It may also be possible that as older speakers' hearing performance degrades with age (Braun & Diehl, 2010) they alter their speech patterns to compensate.

2.1.4.14 'Elderly speech'

Moving from physiology to behavioural changes briefly, it has been proposed that this compensation process may lead to a 'socially ingrained' pattern of older speech (Ramig, 1986). The idea that older speakers emulate a sociolinguistic pattern of speech associated with their age was also put forward by Hockett (1950) in his theory of age-grading. The possibility that speakers alter their speech according to their age in this case could be the results of social conditioning or as a compensatory strategy in the face of physiological change; it is also very likely that the two interact. If this is the case, it could

very well be that formants are lowered due to lowered larynx being part of an elderly speech norm, rather than physiological causes.

2.1.5 **Summary**

Of course it would be far too simplistic to imagine that any one factor would be the cause of measurable change in such a complicated system such as that involved with linguistic production. Furthermore, it seems the most consistent finding throughout this body of work is that change is variable between individuals. However, there are findings which have been presented in the previous section which provide reasons for different patterns of change, and might just predict the course of change for subjects in the current study. For example, if a speaker shows a pattern of change where formants are lowered, this might support the hypothesis that the vocal tract is extended throughout adulthood and we could predict pattern in non-contemporaneous forensic cases. Conversely, it might be that speakers' formants are subject to influence from more pervasive varieties of their dialect (as in the Queen's English (Harrington, Palethorpe, & Watson, 2000b)), in which case we may be able to look to those dialects as indicators of likely change for individual subjects.

There are predictions in the literature about the likely age-related changes to frequency characteristics that are relevant to the ages under analysis in the current study:

- F0 will reduce
    - F0 will reduce in men by around 10% or around 15Hz
    - For women the effect will be less marked
    - Smoking significantly lowers F0 and increases rate of decline
- Formant frequencies will reduce
    - F1 will reduce significantly in men and women
    - F2 and F3 reduce significantly in women, and show a tendency to do so in men
    - There may be a relationship between F0 and F1 reductions (preserved by the speaker using jaw opening)
    - Formant transitions will be less extreme (and presumably reduced in Hz)
- Vowel spaces will become contracted
- Speech rate is reduced

- There may be changes to any of these patterns based on specific differences in accent types, due to geographical or social mobility, or to mainstream changes to an accent type

2.1.6 **Questions**

Questions have already been raised out of the debate concerning real and apparent time methods:

1 What is the extent of vocal instability in adulthood and how will this affect how we interpret results from apparent-time studies?

Expanding on this idea, the findings presented from all of these disciplines raise an important further question from a forensic perspective. This was raised by Künzel (2007) and concerns the reliability of forensic evidence from long-term non-contemporaneous samples and relates to the effects of aging in general:

2 What is the magnitude of change in individuals' vocal output during adulthood?

a. Which features remain more stable than others?
b. To what extent are changes predictable from a model of sociolinguistics or gero-physiology?
c. What effect should this have on how we evaluate forensic speech evidence?

2.2 **Formant Dynamics**

The second purpose of this study is to further investigate the discriminant power of dynamic measures of formants. Furthermore it assesses the relative stability of dynamic method in comparison with more traditional measures, observing the stability of each approach across long-term non-contemporaneous data.

Work by McDougall (2005; 2006) has developed strands of research into the speaker-specificity of formant frequencies, using measurements across the duration of a vowel segment to capture the behaviour of the articulators across the vowel, rather than at a single central slice. Although central targets are prescribed by a dialect, speakers have more articulatory freedom moving between them:

> We can hypothesise that targets are highly constrained by the shared language system and that the transitions [between them] present greater potential for individual variation. (McDougall & Nolan, 2007, p. 1825)

This methodology has been termed a 'formant dynamics' approach, and this thesis investigates how well this individual behavioural variation is preserved over time. This section introduces the theory which puts formant dynamics (FD) forward as a potentially excellent parameter for FSC alongside initial research. Furthermore, it introduces a discussion of potential methods for improving on previous findings with realistic speech data and poses questions about the usefulness of FD methodology for forensic purposes.

2.2.1 **Theoretical model**

The theory explicated in this section illustrates the hypothesis that:

> The relatively invariant dimensions of the vocal tract, and the complexity and highly practised nature of skilled articulatory routines, means that the time-varying resonance properties of the speech signal give greatest scope for speaker-characterisation patterns. (Nolan & Grigoras, 2005, p. 168)

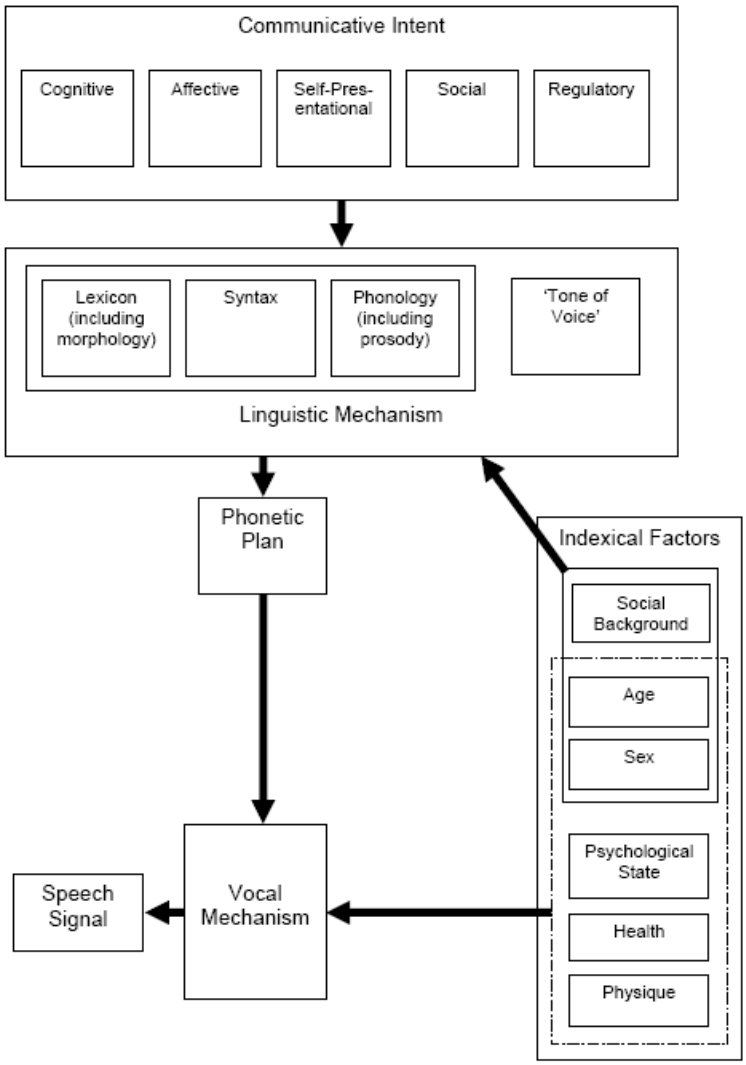
Formant dynamics relies upon assumptions from the task-dynamic model (Saltzman, 1986), from articulatory phonology; this framework equates phonological units with "dynamically specified units of articulatory action", known as gestures (Browman & Goldstein, 1990). Speech then, is a series of goals [targets] and a gesture is defined as "the activities of all articulators in completing a goal" (McDougall, 2006, p. 93). The benefit for FD over other articulatory models is that the task-dynamic model sees speech units as actions rather than static outcomes.

There are five key 'tract variables' in this model (Saltzman, 1986; Haskins Laboratories, 2007): the lips, tongue tip, tongue body, velum and glottis. These all move in combination, within different ranges and at different speeds, to create the supra-laryngeal conditions which are the source of formant frequency variation (Nolan & Grigoras, 2005). It is logical to assume that static central measurements would not register this range of highly complex articulator trajectories. Furthermore, Carré et al. (2007) point out that a dynamic approach incorporates velocity as well as other spatial and temporal properties of the speech signal that are traditionally measured. This is somewhat true, but many dynamic approaches normalise for duration, losing some of this temporal information.

The articulators move in complex configurations to achieve targets which are defined by a dialect. According to McDougall (2005) the focus of formant dynamics is "inter-speaker variation in properties of the acoustic signal *in between* the moments at which phonetic targets are achieved". Therefore, central measurements made at the speaker's realisation of a target are constrained by a dialect and contain some *anatomical* information. Dynamic measures, however, capture a speaker's articulatory *behaviour;* with the increased freedom speakers have moving *between* defined linguistic targets, the premise is that this reveals more idiosyncratic information. This emphasis on the speaker-specificity of behaviour rather than physiology (or nurture over nature) is supported by a study into monozygotic (identical) twins' speech (Loakes, 2004; 2006). Despite having identical DNA, therefore (almost) identical vocal anatomy, and the same linguistic environment, twin speakers had significantly different formant outputs. Loakes concludes that twins "use different articulatory strategies in approaching the same phonological targets" (2004, p. 290).

The idea of speakers using individual 'articulatory strategies' originates in Nolan's (1983) key monograph. These strategies are necessary for speakers to resolve the problem of "mapping communicative intent" (1983, p. 61) onto the vocal apparatus:

Figure 11 - Mapping of communicative intent



Source: (McDougall, 2005, p.7)

Different speakers resolve this complex task in different ways, and form highly exclusive articulatory solutions (varying greatly between speakers) (Nolan, 1997). Furthermore, as these strategies and solutions are developed through trial and error and are therefore highly practised, intra-speaker variation is low for these behaviours (Nolan & Grigoras, 2005).

While FD focuses more on behavioural patterns, vocal physiology still plays an important role, not only in defining "the range within which variation in a particular parameter is constrained to take place" (Nolan, 1983, p. 59), but in affecting speakers' preferred gestural solutions in individual ways. McDougall (2006, p. 92) states that "individuals are likely to make articulatory movements in different ways to accommodate anatomical differences" using an example from Johnson et al. (1993, p. 712) to illustrate:

> With respect to degrees of palate doming, a speaker with a shallow palate might move the jaw and tongue together in the same single motion, while a speaker with a deeply vaulted palate might use a larger, more independent movement of the tongue to achieve a functionally equivalent outcome.

In summary, the complexity of cooperative articulatory movements coupled with the problem of 'mapping communicative intent' entail that articulatory solutions (influenced or bounded by vocal anatomy) between target configurations are highly variable between speakers. Moreover, the highly practised nature of these gestures entail that these patterns are relatively invariant within a speaker. For FSC these are highly desirable traits and, in theory, warrant the extraction of dynamic formant measurements in order to discriminate between speakers.

2.2.2 **Research**

Several studies have tested the discriminatory power of formant dynamics, and are summarised in Table 14 below. While many of these studies share a similar basic methodology, and most use a discriminant analysis (DA) test, there are differences which are described in this section.

Table 14 - Summary of studies into formant dynamics (updated from Hughes et al., (2009))

Study	Language	Speakers	Segment(s)	Test	Result
Greisbach et al. (1995)	German	80	6 long vowels	Matching procedure	**94%**
Ingram et al. (1996)	Australian English	15	21 'chunks'	DA	**93%**
McDougall (2005; 2006)	Australian English	5	/aɪ/ (+ other segments)	DA	**88-95%**
McDougall & Nolan (2007)	English (SSBE)	20	/u:/	DA	**52%**
Eriksson & Sullivan (2008)	Swedish	5	/jœ:/	DA	**88%**
Morrison (2008)	Australian	27	/aɪ/	Likelihood ratios (LR)	**varying LR scores**
Morrison (2009a)	Australian	27	5 diphthongs	Likelihood ratios (LR)	**varying LR scores**
Hughes (2009)	English (SSBE) (spontaneous)	20	/aɪ/	DA	**45%**
Rhodes (2009)	English (Derby variety) (spontaneous)	8	/aɪ/	DA	**50-54%**
Atkinson (2009)	English (SSBE) (spontaneous)	25	monophthongs	DA	**24-40%**
Enzinger (2010)	Viennese German	30	/aɛ/ (controlled contexts)	Likelihood ratios (LR)	**7-16% Equal error rate (EER)***

*Equal Error Rate (EER) represents the thresholds for false acceptances and false rejections of the system, where both are equally frequent; the lower the EER, the more accurate the system.

The table above demonstrates that dynamic measures have shown promise in separating between speakers (note that many discrimination scores are the optimal test result). While the number of speakers is relatively small (in a forensic exercise the expert may be theoretically discriminating a speaker from thousands), early studies show that a small number of vowels are able to separate speakers in up to 95% of cases. However, there are differences between these studies, and their results, which need addressing.

2.2.2.1 Expressing strength of evidence

The major difference between those studies by McDougall (2005; 2006; McDougall & Nolan, 2007) and Morrison (2008; 2009a) is the expression of strength of the FD data, not the data itself. Rather than DA, Morrison (2008; 2009a) uses a Bayesian likelihood ratio approach, arguing, from Rose (2002; Rose & Morrison, 2009), that it is the only 'logical and legally valid' method for expressing conclusions and that DA is unsuitable for forensic casework. In fact McDougall (2006, p. 119) acknowledges this, explaining that DA is "a statistical tool for assessing and comparing the degree of speaker-specificity offered by sets of variables". In the studies presented above then it should be noted that a DA is merely testing the effectiveness of the parameter at discriminating between speakers and is not intended to be a part of any forensic investigation, quite naturally, as the test itself requires the knowledge of which samples belong to which speakers. The use of LRs and other methods for expressing the strength of evidence are discussed in §2.3.

2.2.2.2 Processing dynamic data

Most studies use time-normalised measurements of dynamic formant contours across different segments for comparison. In some cases (see Table 14 above), discriminant analysis (DA) is used to test how well the formant dynamic method separates speakers, giving a percentage of data correctly attributed to the speaker who produced it. Further to this, more recent studies have included polynomial regression (McDougall, 2005; 2006; McDougall & Nolan, 2007) or direct cosine transforms (DCTs) (Morrison, 2008; 2009a; Enzinger, 2010) in order to reduce data to curve coefficients. This has a benefit for DA as it can reduce the number of tokens necessary to include all data points in an analysis by reducing the number of predictors. It can also be used to characterise a speaker's habitual formant transition behaviour in order to compare samples using a numerical strength of evidence estimate, such as an LR. Studies by McDougall (2005; 2006) and McDougall & Nolan (2007) using polynomial equations for regression state that for /aɪ/ and /u:/, (PRICE and GOOSE (Wells, 1982)), quadratic and cubic polynomials perform best, even performing better than raw data, at 89-96% (with a small data set of five speakers). It is important to remember that in these studies which used DA, a smaller number of speakers in any study will exhibit increased classification rates. Morrison (2009a) disagrees, proposing direct cosine transforms (DCTs) as performing better than polynomials, although he admits that there is little difference in the best performing

formants. Morrison (2009a) also sensibly states that the curves used for polynomial regression should be selected on a case-by-case basis, and that the order of the curve should reflect the shape of the formant contour. For instance, an S-shaped, or sigmoidal, formant contour should be reduced to a curve using a cubic polynomial (third degree).

2.2.2.3 Formant dynamics and real speech

Although the discrimination scores for earlier studies shown in Table 14 are very promising, they all analysed carefully recorded laboratory data from read speech; in the majority these tokens were elicited in desirable phonological contexts. This is a concern for application to forensic casework where most recordings mostly feature styles other than read speech and tokens from a range of phonological contexts. For those studies where spontaneous data were elicited, classification rates were much lower: 24-40% for monophthongs (Atkinson, 2009); 45% for /aɪ/ in SSBE (Hughes, McDougall, & Foulkes, 2009); 54% for /aɪ/ in Derby English (Rhodes, 2009)). Perhaps more importantly, these studies using spontaneous speech did not produce rates that were much higher (or even higher at all in some cases) than traditional central formant measures. Hughes (2009) performed DA with spontaneous data from 20 speakers and found that for /aɪ/ in SSBE, classification rates were much lower, although still well above chance (5%) at 45%, than for McDougall's (2005) Australian five speakers in read speech.

Moreover, while the speech in Hughes' study was 'spontaneous', certain target words were still elicited and tokens were taken from between voiceless obstruents to augment ease of extraction and reduce coarticulatory effects. The tokens in Atkinson (2009) and Rhodes (2009) were from more realistic data (i.e. target words were not elicited and therefore vowel tokens had to be extracted from a number of consonantal contexts). Atkinson (2009) found that for monophthongs, classification rates were fairly low, between 24-40%, depending on the vowel in question. In Rhodes (2009), where tokens of /aɪ/ were taken from all contexts, except surrounding nasals, /r/ and /w/, which had too strong an effect on the contours, classifications with the best-performing parameters were still fairly promising, at around 50%. However, this was only for eight speakers and in a variety of English (Derby) where /aɪ/ is very sociolinguistically variable, which ought to have augmented the separation of speakers. McDougall (2005) reports a similar notion that the high (88-95%) classification in her study may be partly due to the high sociolinguistic variability in Australian English of the PRICE vowel, along with the fact that

her speakers were not matched precisely for dialect. Hughes et al. (2009) also posit the opposite effect as a reason for reduced performance in SSBE.

Another finding from Hughes (2009), which may be damaging for a dynamic approach, was the performance of DA using only two measurements for each formant, taken at the steady-state or target portions of the /aɪ/ vowel. Classification rates for 10 best performing predictors (45.1%) (i.e. a dynamic approach) were marginally outperformed by measurements taken at 20% and 70% of each formant (45.7%) (measures of the two targets). Considering that this steady-state analysis used fewer predictors (6), and that generally increasing predictors yielded better classification rates (McDougall, 2005), this finding questions the idea that transitions present more speaker-specific information in spontaneous speech. Rhodes (2009) uncovered similar results for formant measures taken at the 'target' of each part of the /aɪ/ diphthong; the best dynamic classification rate based on the most discriminative predictors (n=17) was 53.7%, however, based on only six predictors at the 'target' portion of the vowel, data were correctly attributed to a speaker in 52% of tests. It seems that for these spontaneous data, dynamics does not present an effective advantage over traditional central formant measures. For McDougall's (2006) elicited speech, however, central 'instantaneous' measures performed at 57-68%, far worse than the dynamic measures (88-95%). However, many of these studies compared static points of transitions to try and capture speaker behaviour, and polynomial regressions would lead to inclusion of the more information.

2.2.2.4 Improving performance

Given the promising results from laboratory studies, and the encouraging discrimination rates from those studies which looked at spontaneous data, attention is now being turned to methods of augmenting the effectiveness of FD methods. There are a number of propositions which may improve or elaborate further on the success of formant dynamics. It would seem sensible, having seen discrimination rates for single vowel classes, to investigate a more comprehensive approach to a speaker's realisation of diphthongs in multivariate analyses. These could even include monophthongs to see how successful a more holistic impression of dynamics plus traditional measures is. Of course, the problem with multiple diphthongs is their availability in real speech and especially forensic recordings. Even in recordings which make up part of the present study, many

speakers exhibited only a few tokens of PRICE and FACE words, with other diphthongs even more infrequent.

Another approach would be to investigate more and more diverse dialect populations, such as Derby English, to elucidate if more socially variable diphthong productions yielded better speaker-characterisation. This would give a better picture of the suitability of formant dynamics for FSC in more than a handful of dialects.

A problem for FD methods is the effects of consonantal context. However, this might be neutralised with novel consonant-environment normalisation techniques which are being developed with forensic exercises in mind (Clermont, 2009; 2011). This would presumably remove obfuscatory effects of consonantal influence and give a more accurate picture of a speaker's vowel formants. Conversely it may remove idiosyncratic information about how speakers move between target, *including* different consonants. To the author's knowledge, no study has characterised transitions between different consonants to see their speaker-characterising effects, though presumably this would be difficult with limited forensic data. Another approach might be to investigate consonantal formant transitions themselves, across nasals and laterals, to see if there are any dynamic processes which would contribute to discriminating between speakers (although cf. work on discriminant power of consonants by Kavanagh (2012)).

2.2.3 **Questions**

It was illustrated in a previous section that there is limited literature on formant transitions and aging, but that one study (Liss, Weismer, & Rosenbeck, 1990) indicated that gestures would become 'less extreme' in much older age. Given the physiological factors which, it is theorised, affect a lowering in formant values, it would also seem sensible to suggest that overall formant frequency would decrease in diphthongs as well as monophthongs. This study is the first (to the author's knowledge) to investigate stability of dynamic parameters in the face of aging through early adulthood.

This section has presented a number of questions regarding aging and formant dynamic measures, which this study investigates:

3 What are the effects of aging on formant transitions?

   a. How do these effects compare with monophthong formants?

b. Are speakers more stable in 'target' sections of a diphthong, or in their movements between targets?

c. Is making dynamic measures of vowel formants worthwhile and/or reliable in cases of long-term non-contemporaneity?

2.3 **Expressing the strength of forensic speech evidence**

If evidence is presented in court, it is vital that the trier of fact can make an assessment of the strength of that particular piece of evidence and how it should affect their judgement. The final review section of this chapter addresses the issue of how the strength of forensic speech evidence is assessed. In particular, it discusses and explores the use of likelihood ratios (LRs) for forensic speech evidence.

2.3.1 **Context**

The issue of the presentation of forensic evidence has become a prominent debate in the field of forensic speech science over the past 10-15 years. This is in part due to advances in other fields of forensics such as DNA and toxicology analysis, where strength of evidence estimates are now canonically delivered using likelihood ratios (Aitken & Taroni, 2004). Before presenting recent UK legislation and regulations which address this debate, the present framework for presenting speech evidence in the UK is illustrated.

2.3.1.1 Current practice in the UK

If not formally established, the following practice has been undersigned by almost all forensic phonetics practitioners in the UK and serves as the conclusion framework for most work carried out in that jurisdiction (therefore, this approach is henceforth referred to as the 'UK' framework). It is also by a large proportion of forensic speech analysts working outside of the UK, especially those working within an auditory-acoustic method (Gold & French, 2011).

The UK framework is set out in French and Harrison (2007). It was originally developed to counteract impressionistic conclusions such as 'I believe the suspect is the perpetrator', which are legally and logically invalid (see §2.3.2.4). The voices in the evidential and suspect recordings are first determined to be 'consistent' or not, and if so, the expert phonetician must determine the 'distinctiveness' of the features analysed. Naturally a decision of 'no consistency' rejects the suspect as the speaker in the evidential sample. This 'no consistency' decision has been criticised by Rose and Morrison (2009) for being

just as logically and legally invalid as the conclusion above: 'I believe the suspect is not the perpetrator'. French et al. (2010) responded by highlighting cases where there were practical reasons for being able to eliminate a suspect, e.g. obvious differences in gender and accent. These kinds of eliminations are also stated in Rose and Morrison (2009) and in general throughout their approach, it is unclear how their objection to the cases in French et al. (2010) relates to non-LR based exclusion of evidence in the LR approach. In practice it would be useful for analysts to be able to express a gradient degree of consistency, alongside categorical rejections. It is likely that some samples could be strikingly similar where others are similar enough to be compared with respect to distinctive features. In order to make this approach analogous to numerical approaches, the level of consistency (or similarity) should be incorporated into final assessments of the strength of evidence.

The distinctiveness element of the framework is similar to the 'typicality' component of the LR, although in this framework it is theoretical, generated using the expert's knowledge and research into a dialect or accent, rather than being calculated by reference to data from a relevant background population (see §2.3.2.5).

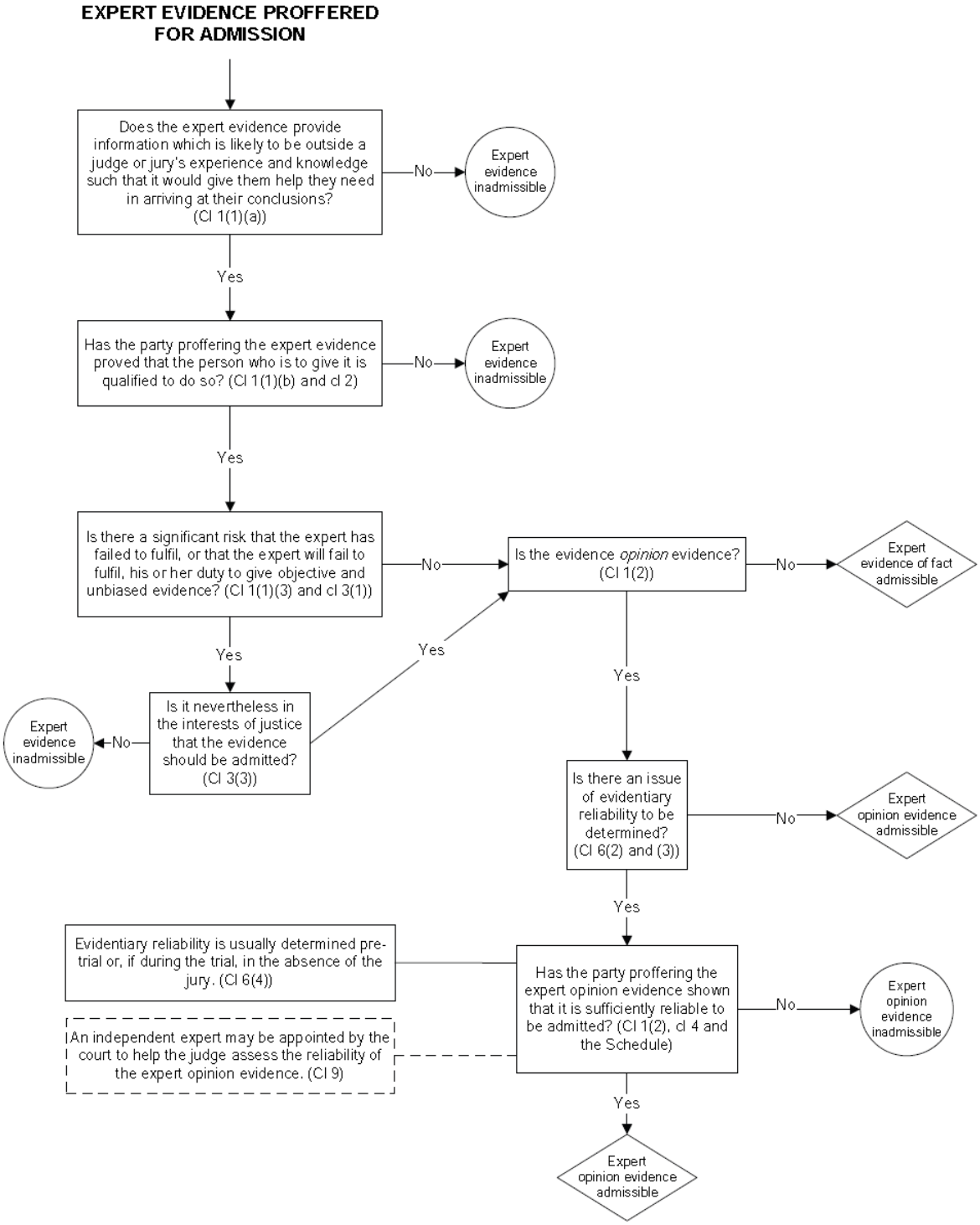
2.3.1.2 Guidelines on evidence

There are a number of legal precedents and guidelines which inform the way evidence must be presented and processed. Relevant rulings from the UK are summarised with background from broader jurisdictions including the US, which have generally influenced the UK context.

Historically, evidence has been subjected to a number of admissibility conditions. One of the original conditions was set out in the case of Frye v United States (1923), which stated that evidence must be formulated from a 'well recognised scientific principle' and that methods be established and have 'general acceptance in the particular field'. These conditions were expanded upon with the *Daubert* rulings (Daubert v Merrell Dow Pharmaceuticals Inc, 1993) which demanded that the 'rate of error is known and is acceptably low'. Known error rates are problematic for experience expert based forensic disciplines, where both the method and the expert should ideally be tested for error rates, which may prove difficult given the circumstances that in real cases the base truth is rarely known or discoverable. There have been calls for rigorous blind-testing of experts (Brown & Willis, 2009). Courts in the UK also routinely apply admissibility criteria

based on guidance from an Australian case (R v Bonython, 1984) as to whether an expert is required (based on whether the evidence is beyond the ability of a layperson to interpret) and the reliability of the witness' knowledge and experience (French, 2011). The UK government have sought to tighten admissibility and have published a public bill (Criminal Expert (Experts) Bill, 2011) and a set of guidelines (Law Commission, 2011) designed, in part, to bring the UK into line with general EU legislation. This document demands that opinion is soundly based and that the strength of the opinion is warranted, with a number of caveats presented for ruling out evidence. One of the likely causes for argument may be the caveat that opinions should not be based on flawed data, which causes problems for judgements of typicality either by experience or LR where not enough is known about reference population composition. The procedure for evidential admissibility in the UK is set out below:

Figure 12 - Admissibility of expert evidence in criminal proceedings



Source: (Law Commission, 2011)

There have also been further standards (*Competence of testing and calibration laboratories* ISO 17025) introduced by the Forensic Science Regulator in an attempt to bring all laboratories in line; these standards require a number of policies relating to staff, audit and information to be maintained. In terms of evidence, it also requires validation of methods and accreditation and testing of staff on a continuous basis. There are

political factors which have influenced the introduction of these guidelines, including a desire to improve public confidence in forensic evidence after high profile miscarriages of justice, such as that in the Sir Roy Meadow case (General Medical Council v Meadow, 2006), where inappropriate use of statistics lead to a defendant being wrongly convicted of murdering her two baby sons.

In summary, there is increasing pressure on forensic science providers to be able to demonstrate that testing methods are both valid and reliable and that analysts are accurately performing these tests. New regulations have also put forensic speech analysis under question with a suggested dichotomy between scientific and experiential evidence. It is unclear where speech analysis falls given that it is based on scientific principles, but relies on expert interpretation in some cases, and numerical calculation in others.

2.3.1.3 Market factors

There are also several factors in the UK which may affect funding for forensic examinations (in general), which in turn have an impact on practical demands for analyses. These are largely dictated by structural reductions to public funding made in the government's *Spending Review* in 2010-11 (HM Treasury, 2010). A large focus of the public sector spending cuts fall on external police expenditure (Cellmark Forensic Services (FSS73), 2010), of which a sizeable amount is forensic examination. For example the Metropolitan Police Service (London) is already seeing reductions in both budget allocations to forensic examinations and number of examinations carried out, with further reductions planned (see Table 15 below).

Table 15 - Statistics for expenditure on forensic examinations by the Metropolitan Police Service

Year	Budget*	Actual spend*	Number of examinations carried out**
2008/09	£101,634,726.24	£93,686,369.65	128,472
2009/10	£92,832,379.72	£94,173,435.65	126,067
2010/11	£86,963,892.97	£81,039,916.09	99,018
2011/12	£86,113,910.54		

*figures for forensic charges, DNA testing and staff

**figures are for Total Notifiable Offences as required for reporting to the Home Office. The figures do not include those incidents that are classified as "crime related incidents" or any forensic examinations which do not have an MPS Crime Reporting Information System (CRIS) report

Source: (Metropolitan Police Service (FOI request), 2011)

Furthermore, the MPS is provided with Home Office funding for DNA forensic testing only, which limits the uses of other kinds of forensic examination, including speech.

Alongside these cuts came the closure of the Forensic Science Service in 2011, which in turn leads to increased privatisation of the forensic market in general. Despite a relatively high level of research output in forensic fields (ranked 5th in world tables for citations in 2011 (Frankel, 2011)), the FSS was closed as it was not profitable. This comes in the context of widespread cuts to public services in favour of privatisation from a Conservative-Lib Dem government, and there is no doubt that the private sector benefits from such action. Despite this, there is also a general expectation that cuts will also cause the external forensic markets to fall. James Brokenshire, parliamentary under-secretary of state for the Home Office suggested, based on advice from the Association of Chief Police officers (ACPO) and Her Majesty's Inspectorate of Constabulary (HMIC), a reduction in the general market from £170m in 2009 towards £110m in 2015 (Parliamentary Questions (#295-354), 2011). He also cited that he expected a private marketplace to drive down cost and response times, evidence or projections for this were not provided.

In general then, despite increased legislation for forensic businesses and evidence admissibility, there is a significant downturn in police funding for forensic examinations. This is particularly true of non-DNA testing (which is sometimes carried out by Police force laboratories (Parliamentary Questions (#229-354), 2011)). There is a strain on the forensic market to provide quick, accurate forensic evidence at a much lower cost, which also satisfies admissibility requirements (i.e. has known error rates and other reliability requirements).

2.3.2 **The likelihood ratio approach**

> If it can be accepted that nothing is absolutely certain then it becomes logical to determine the degree of confidence that may be assigned to a particular belief [or set of beliefs] (Kirk & Kingston, 1964, p. 435)

The likelihood ratio is one method of assigning a level of confidence to a particular belief, such as 'suspect *x* is the source of the evidential sample *y*'. The likelihood ratio is one part of Bayes' theorem, which was established by Bayes (1763) to express the measure of the strength of evidence in support of two competing hypotheses (although it should be noted that calculating an LR is not a 'Bayesian' exercise, as this implies assignment of prior odds (Champod & Meuwly, 2000)). For forensic speaker comparison then, Morrison

(2010e, p. 2) proposes that forensic scientists should provide a 'strength-of-evidence statement in answer to the following question':

> How much more likely are the observed differences between the known and questioned samples to occur under the hypothesis that the questioned sample has the same origin as the known sample than under the hypothesis that it has a different origin?

There has been a recent focus on statistical applications of Bayes' Theorem for assessing strength of evidence across forensic analyses in general (cf. Robertson & Vignaux (1995)) and for speech evidence (there is a wide literature from these authors, for comprehensive introductions see (Rose, 2002; Rose & Morrison, 2009)). Rose and Morrison (2009) argue that the LR approach is the 'only legally and logically valid' method for presenting speech evidence. Below is a version of Bayes' theorem which is commonly presented, the component parts are explained in the following sections.

$$\underbrace{\frac{p(H_{ss} \mid E_{sp})}{p(H_{ds} \mid E_{sp})}}_{\textit{Posterior Odds}} = \underbrace{\frac{p(H_{ss})}{p(H_{ds})}}_{\textit{Prior Odds}} \times \underbrace{\frac{p(E_{sp} \mid H_{ss})}{p(E_{sp} \mid H_{ds})}}_{\textit{Likelihood Ratio}}$$

Source: Rose and Morrison (2009, p. 144)

2.3.2.1 The Likelihood Ratio

The magnitude of the LR represents the measure of the strength of evidence within Bayesian approaches to probability (Rose, 2002). If we examine the formula above, the LR is the ratio of two probabilities, that of the speech evidence (E_{sp}) given a same speaker hypothesis (H_{ss}) and given a different speaker hypothesis (H_{ds}). The numerator of the LR represents the probability of the evidence given a 'same speaker' hypothesis (i.e. the suspect produced the evidential sample), which is weighed against the 'different speaker' hypothesis (or random match probability (Champod & Meuwly, 2000)). If the numerator is greater, this provides support for a 'same speaker' hypothesis, and vice versa if the denominator is greater this provides support for a 'different speaker' hypothesis (in idealised cases). In practice these hypotheses are much more refined and depend on the intricacies of each case (see §2.3.2.5).

The probative value of evidence is defined by the magnitude of the LR, or the distance from 1. For example, if LR=10, the evidence would be 10 times more likely to have been produced by the suspect (and *vice versa* for 0.1) (Robertson & Vignaux, 1995). LRs are

often converted using logarithmic scales to make them easier to interpret, where positive values indicate support for the same-speaker hypothesis and negative values support the different-speaker hypothesis. If independent features are analysed and LRs are calculated for each, then these can be combined to create an overall LR for the case or presented in turn to keep updating an existing probability. In cases of uncorrelated data, a simple multiplication can be used. For speech data this is not appropriate, as different frequency outputs (particularly in vowel systems) are related and therefore methodologies for fusing LR scores have been put forward (Rose, 2010; Morrison, 2011). For two proponents of the LR approach, estimating the LR is the only responsibility of the forensic speech analyst (Rose, 2002; Rose & Morrison, 2009). This position is not universally accepted. Robertson and Vignaux (1995, p. 54), for instance, argue that: "Part of the task of the expert witness should be to explain how and why the court is helped by the evidence they are giving".

For forensic speech science, the LR is essentially a test of the similarity and typicality of a set of features extracted from recordings, much like the 'consistency' and 'distinctiveness' found in the UK approach but expressed with explicit empirical reference to statistics derived from datasets. The similarity aspect is encapsulated in the numerator, so that the 'same speaker' hypothesis is tested by determining the 'overlap of general distribution of features', and the typicality aspect in the denominator, is determined by 'distance from a reference population mean' (Rose, 2002). In a sense the two parameters correspond somewhat to two of Nolan's (1983) criteria, where the numerator encapsulates intra-speaker variation and the denominator assesses inter-speaker variation in a relevant population. The distance from a reference population is very important as it can heavily determine the outcome and magnitude of the LR and therefore the reference populations used to calculate this typicality must be selected very carefully (see §2.3.5.12).

The LR can be expressed in many different ways, and these are discussed in §2.3.5.4, but Morrison (2010a) provides a sample conclusion of a (support for different-speaker hypothesis) LR=0.000001:

> Based on my evaluation of the evidence, I have calculated that one would be one million times more likely to obtain the acoustic differences between the voice samples if the question-voice sample had been produced by someone other than the accused that if it had been produced by the accused. What this means is that whatever you believed before this evidence was presented,

you should now be one million times more likely than before to believe that the voice on the questioned-voice recording is not that of the accused.

2.3.2.2 Prior odds

Returning to the formula above, the prior odds represent the existing probability of each hypothesis 'prior' to the evidence. This depends on the specific defence hypothesis, which is determined by accent or speaker sex or other factors which may be extracted from the evidential sample. For example, if the perpetrator was confirmed as being from a city with 10,000 residents, the prior odds might be 1/9999. If the defence hypothesis is that another man was the source of the recording, this may be reduced to 1/5000 for instance.

In practice there are issues with the calculation of prior odds, this is especially problematic given that "very large or very small prior odds can give some very startling effects" (Robertson & Vignaux, 1995, p. 19) and that jurors are responsible for assigning and understanding priors, including the impact of their own biases (see §2.3.5.4). Robertson and Vignaux (1995) express concerns about the effects of extremely small or large priors and give the example of an HIV test. In 1985, the test had a 100% hit rate (LR=100) and a 1% false positive rate (calibrated as such to avoid missing positive tests). One in 10,000 in the population had HIV (so priors were 1 to 10,000). Thus, LR of 100 multiplied by the prior odds actually gave a [posterior] probability (cf. the following section) of 100 to 1 against having the disease, and a second confirmation test would be administered.

2.3.2.3 Posterior odds

The posterior odds are the result of multiplying the prior odds and the LR, and give the odds in favour of being the same speaker given the evidence, i.e. the conclusion of the forensic exercise. Unlike the LR, the posterior odds is a $p(H|E)$ statement, that is, it gives the probability of the hypothesis, given the evidence (the LR is a $p(E|H)$ statement: the probability of the evidence, given the hypotheses). The posterior odds, as with all $p(H|E)$ statements, are the responsibility of the trier of fact (Robertson & Vignaux, 1995) and it is against the law for witnesses to take this role.

2.3.2.4 Logical fallacies in evidence presentation

There are several reasons why the expert cannot produce a $p(H|E)$ statement, i.e. the posterior odds. Rose and Morrison (2009) present the two main reasons, one logical, and one legal. Firstly, logically, the analyst does not generally have access to the necessary prior odds in a case to calculate posterior odds, especially when, in the case of a jury, the priors may vary between individuals as their individual beliefs about the facts and persons in the case would naturally differ. Secondly and perhaps more importantly, to give the posterior odds would 'usurp the role of trier of fact' (Robertson & Vignaux, 1995; Rose & Morrison, 2009), in breach of what is known as the 'ultimate issue' rule. In English law, the decision of guilt (the ultimate issue) is the responsibility of the jury (or judge in UK Magistrate courts). If the expert were to present a $p(H|E)$ statement such as 'the suspect is the source of the evidential sample', therefore effectively incriminating the suspect, they would be in breach of the following ruling:

> It is not competent in any action for witnesses to express their opinions upon any of the issues, whether of law or fact, which the Court or a jury has to determine (Neville, J. (Joseph Crosfield & Sons v Techno-Chemical laboratories Ltd., 1913))

Another commonplace error is known as the 'prosecutor's fallacy' (Thompson & Schumann, 1987), or transposing the conditional (Evett, 1995). This gives exaggerated weight to the prosecution hypothesis by confusing what are actually prior odds with calculation of strength of evidence. For example if a set of DNA features was only shared by 1% of the population, the chance of the suspect being the criminal is not 99/1, but 1/1% of the relevant population.

The 'defence attorney's fallacy' occurs when population statistics are cited without considering their associative value with the evidence, giving less weight to the evidence (Thompson & Schumann, 1987). For example:

> Suppose, for example, that the defendant and perpetrator share a blood type possessed by only 1% of the population. Victims of the fallacy reason that in a city of 1 million there would be approximately 10,000 people with this blood type. They conclude there is little if any relevance in the fact that the defendant and perpetrator both belong to such a large group. What this reasoning fails to take into account, of course, is that the great majority of people with the relevant blood type are not suspects in the case at hand. (Thompson & Schumann, 1987, p. 171)

While these evidential fallacies are logically erroneous, it is their effect on the trier of fact's understanding of the case that is most damaging.

2.3.2.5 Hypotheses and the reference population

In order to calculate the LR, the expert needs to know the two competing hypothesis of which it is composed. The prosecution hypothesis (or same-speaker hypothesis) (H_{ss}) is usually relatively straightforward: 'the suspect is the source of the evidential sample'. It is the different speaker, or defence, hypothesis (H_{ds}) that is much more problematic and complex. The defence may not always have a clear hypothesis, and this may change during the case. Robertson and Vignaux (1995, p. 33) argue that the process of hypothesis-forming is of paramount importance:

> It is difficult if not impossible to determine the probability of the evidence with a vague and ill-defined hypothesis such as "the person is not guilty" [the suspect did not produce the sample], the value of the evidence will best be realised if the two hypotheses are well formed, positive and specific

This is because the defence hypothesis necessarily affects the prior odds (Robertson & Vignaux, 1995) and determines the reference population against which the typicality of the material is tested (Aitken & Taroni, 2004; Robertson & Vignaux, 1995; Rose, 2002; Rose, Personal communication, 2010). Robertson and Vignaux (1995) propose that logically, the two hypotheses must be exclusive but in practice, they need not be exhaustive. They may be two different versions of events, but it would be unwieldy to formalise every possible version of events. However, an LR with non-exhaustive hypotheses is not a logically 'true' estimate of the strength of evidence. For 'true' odds, the hypotheses must be exclusive and exhaustive, for instance the two sides, and potentially the edge, of a tossed coin.

Forensic scientists have often assumed that the H_{ds} is that the perpetrator could be any other member of the population (Robertson & Vignaux, 1995): 'the default could be that the source of the evidential sample is not the suspect, but someone else' (Rose, Personal communication, 2010). This is certainly vague, and the question has been asked whether this represents the population of the Earth, the country, the city or the speech community; just how is this group delimited? Aitken (1995) discusses the idea of a 'potential perpetrator population' or 'relevant population' to make this process more manageable, trying to find the smallest possible population known to contain the criminal. Buckleton and Walsh (1991, p. 464) pose the question: "should it be persons

completely unconnected with the crime or should it be persons of the type likely to come to the attention of the police?". Either of these approaches raises a multitude of questions and problematic definitions, such as how you define a connection with the crime given unknown guilt and the legal presumption of innocence, and 'types' which are likely to come into contact with police. Rose (Personal communication, 2010) argues that the expert has to use their knowledge to supply useful practical limits, based on the sex or the dialect of the perpetrator, for example. Buckleton and Walsh (1991) point out that in many cases the forming of a reference sample is executed as the 'best compromise'.

Morrison, Ochoa and Thiruvaran (2012) have recently put forward a novel approach to selecting the reference population (or database):

> We present a logical argument that because an investigator or prosecutor only submits suspect and offender recordings for forensic analysis if they sound sufficiently similar to each other, the appropriate defence hypothesis for forensic scientists to adopt will usually be that the suspect is not the speaker on the offender recording but is a member of the population of speakers who sounds sufficiently similar that an investigator or prosecutor would submit recordings of those speakers for forensic analysis.

If we return to the idea of the potential perpetrator population, it seems ridiculous to suggest that it is logical to demarcate a group of potential suspects by the fact that a police officer would identify the suspect as similar to the offender recording. Given that a forensic expert is approached to carry out the analysis, as it is deemed too complex for a naive listener, why are police officers able to perform any better in selecting voices for a testing or reference database? Furthermore there are a number of forensic tests where there are untested relationships with perceptual and auditory cues, even with regards to trained phoneticians. If using an automatic GMM-UBM (Gaussian Mixture Model – Universal Background Model) system, for example, the results of this system may be completely at odds with a police officer's selection by 'similarity'. Moreover, there are bound to be huge between-listener discrepancies in selecting 'similar' voices which could cause problems not only in the analysis stage, but also when such an untested approach comes under scrutiny in court. Given these problems, it would seem more sensible to rely on knowledge of the case at hand, or of the speaker on the evidential recording, in attempting to delimit a potential perpetrator population, rather than a subjective similarity judgement by a single individual.

There are also practical considerations to the reference population, especially considering the extent of accent variation in the UK and the lack of existing large scale data. Even if the potential perpetrator population was limited to the residents of a town of 10,000 residents (if no other clue to the profile of the perpetrator could be found), for example, time and financial restriction would prevent an expert from collecting speech data from all the potential perpetrators. The size of a reference population must be practically determined, especially given its bespoke nature determined by demands of the defence hypothesis, which are unique in each case. Added to this are the restrictions of cost, weighed against regulations for quality of data and tested methods. There are few guidelines on the required minimum size of a reference population. Rose and Morrison (2009) state that the size 'depends on the precision required', that larger samples give more reliable LR scores (for further discussion of issues with reference populations and defence hypotheses, see §2.3.5.12). This statement, however, is operationally vague, and there is a dearth of research into what constitutes a satisfactory reference population (although cf. Hughes (2012), discussed below in §2.3.5.13). Moreover there is a tension with budgetary concerns for any forensic exercise as to what level of accuracy in a reference population is affordable.

It can be too easy to forget that those who are actually forming the H_{ds} are generally legal practitioners and may not have probabilistic expertise or awareness of these factors, or that in some cases the hypothesis might be defined in accordance with multiple forms of evidence. For example, imagine if a DNA match had been found alongside speech evidence, and the defendant claimed it was his identical twin brother who produced the DNA and speech samples. In this case the H_{ds} would be that 'the suspect did not produce the speech sample, it was his twin brother', which would set the priors at 1/2 and delimit the 'relevant' reference population to two speakers (insufficient for LR estimation). The expert would also have to take into account the similarity in vocal anatomy caused by shared genes. In practice, these questions are much more complex than the literature suggests. Few questions have been raised about how these decisions are made and what impact they have on the outcome of a forensic exercise.

2.3.3 **LR-based conclusions in Research**

There has been a wealth of research carried out in which an LR has been used to express the strength of certain parameters in discriminating between speakers. The difficulty

with these studies which feature LR conclusions is that although the results display the strength of evidence as an LR, the magnitude of the LR only tells us about the discriminatory power of the features on which the LR has been calculated. Studies such as Rose et al. (2003) and Morrison (2008; 2009a; 2010b) show the process for calculating LRs with speech data, but there is little comparison of different ways of using an LR approach. There are only limited amounts of research whose results relate to the method of LR calculation itself (although cf. reference population research in §2.3.5.13).

One such study concerns how best to treat multivariate trace evidence: in most forensic cases, and certainly in cases involving forensic phonetic evidence, there are a number of features under analysis, for which LR scores have to be combined or presented together in a way jurors understand. Aitken and Lucy (2004) investigated five different methods for expressing the strength of evidence. Two of those methods utilised significance testing, which performed poorly and were illogical for the task in hand. The best two methods were LR based, which utilised multivariate kernel density (MVKD) projections to model the variation both between and within sources. This method has since been adopting for other studies using multivariate data (Kinoshita, 2005; Morrison, 2009a). It is worth noting that there are concerns with these formulae and how speech LRs are combined (cf. §2.3.5.3).

It was previously mentioned that another area that is vastly (and surprisingly) under-researched is the selection and composition of reference populations for LR tests, given the importance of the typicality factor in calculating strength of evidence estimates. It also seems to be the case that the strength of evidence estimates made by LRs have never been tested with authentic speech data, most studies have used laboratory recordings, which we have seen, for example, from formant dynamic studies (§2.2.2) perform much better in discriminating speakers than 'authentic' recordings, which would naturally be subject to more speaker, situational and channel variations.

What these studies have done, however, is to highlight practical benefits and issues that the LR approach encounters.

2.3.4 **Benefits**

Apart from being put forward as the 'only' logical and legal method for determining strength of evidence (Champod & Meuwly, 2000; Rose, 2002; Morrison, 2009b; 2010e),

there are a number of benefits which an LR approach imparts, some of which are listed in this section.

2.3.4.1 The 'cliff-edge' effect

The method with which evidential conclusions are expressed presents a problem for triers of fact. It has been argued that LRs do not suffer from the same problem. Non-LR frameworks, such as that used in the UK generally, express part of their conclusion using a scale with a set number of intervals, for instance, using a 1-5 or 1-9 point scale. This can be found in the UK framework, where distinctiveness is attributed a rating of 1-5, which can be expressed in court using a verbal scale. These scales, it has been argued (Champod & Evett, 2000; Rose, 2002; Rose & Morrison, 2009), suffer from a 'cliff-edge' effect, as differences of say 10Hz frequency may be expressed as moderately strong, whereas 9Hz might not (a simplistic example). Champod and Meuwly (2000) contend that to make any kind of threshold at all is an 'ultimate issue error', as "establishing thresholds is the utility function of the court, not...of the statistician" (Fienberg, 1989). Proponents of the LR approach argue that as LRs are gradient they do not suffer from the same effect. This is true in theory; however, for other types of forensic conclusions, and even for numerical speech evidence, many practitioners use a Log_{10} scale as a verbal base for expressing the LR (Champod & Evett, 2000). Such that LR=12 is in the same 'bin' as LR=89, meaning that the final evidential output suffers from the same effect.

2.3.4.2 Multivariate data

Rose and Morrison (2009) question the UK framework's strategy for dealing with multivariate evidence, questioning the idea that different speech analysis strands may be distinctive to different extents, or not even consistent, and how the approach deals with combining these disparate elements. Certainly this is not set out in detail in the UK framework, with the implication that experts balance these individual 'strands' of analysis, based on expertise and knowledge. Rose and Morrison (2009) propose that LRs are better suited to the problem of multivariate data, as LRs can be combined to update the posterior odds based on a number of different strands of evidence. The simple process is to multiply the LRs, but this is only a true reflection of the odds if the LRs are independent or unrelated, which is not the case in terms of features of speech.

All vowel formants are part of a wider vocal system, and to an extent reflect the harmonics of the source, for example. The extent of correlation of features of speech is generally unknown (although cf. initial findings of French et al. (2012) and Gold & Hughes (2012)). This is an area where further research is vitally needed. There have been attempts to find best practice when dealing with multivariate data, for example the research by Aitken and Lucy (2004) mentioned in the previous section, however, without comprehensive study of the connectedness of vocal patterns, these are only ever an estimate. There are also similar concerns with the formulae used to estimate LRs for speech data, as they have been developed with other fields and data-types in mind (see §2.3.5.3).

2.3.4.3 Estimating validity and reliability

It was mentioned in the introduction to this section that there is pressure for scientific evidence to be able to report on its reliability and validity in accordance with the *Daubert* ruling. This is very difficult for evidence based on the judgements of experts, although Robertson and Vignaux (1995) suggest that this might be achieved in fields such as speech analysis by providing reliability measures through blind testing of experts.

For speech-based LRs, however, Morrison (2009c; 2010c) illustrates a method for establishing the reliability and validity of a forensic exercise. To measure the validity (or accuracy) of a forensic system, Morrison (2010b; 2010c) describes a metric which captures a gradient measure of the proportion of LRs which correctly identify a same or different speaker pair, as well as their magnitude (weighed against a false positive to prevent miscarriages of justice): this metric is called the log-likelihood-ratio cost, C_{llr} and was developed by Brümmer & du Preez (2006). This measure has advantages over EER judgements (explained in 2.2.2) as it takes into account the magnitude of the LR, whereas EER is based on a binary error distinction. Reliability (or precision) is measured using a credible interval (CI), which tests within and between speaker discrimination in a fixed set of speakers. In simple terms, these metrics characterise how likely a test is to correctly discriminate speakers and also the internal variability within a test.

Although these metrics test the system used in a forensic setting, the C_{llr} depends on knowledge of the speakers' identity, much like with DA, and would not be applicable directly to data from a forensic case, where this is unknown. Essentially these metrics test the system, but the performance of that system is always affected by individual

conditions of each case. In this case, it could be argued that using the LR approach is of little relevance, as for instance, DA results for analyses based on formants could be quoted as validity measures for these kinds of evidence in much the same way. The same is true with the recommendation for reliability (Morrison, 2010c), where Morrison suggests that as the evidential sample is fixed, that this should be used to test the system. However, the evidential material is only one part of the forensic analysis and the reliability of the system cannot be truly determined with only one recording, particularly with significant mismatches between materials. Considering the effects of channel and speaker variation that vary from case-to-case, this could present erroneous system-testing scores to the court.

2.3.5 **Issues**

Several issues have been raised with respect to the application of an LR framework to speech data, some of which are presented below.

2.3.5.1 Accent variation

There are two important points concerning accent variation. Firstly, from a practical perspective there are an extensive number of different language varieties in the UK. Accent variation is extremely marked, so much so that speakers from one town or village may be distinguishable from those in another from accent alone (this happened in the Yorkshire Ripper hoaxer case (R v John Samuel Humble, 2005)). Secondly, accents are not uncomplicated, unchanging and homogenous. They are affected by class, gender, age and numerous other factors (Foulkes & Docherty, 2006). Moreover, people who move or are educated in different locations across the country and the world are likely to pick up features from the host accent or language or lose features of their 'home' lect. Speakers also accommodate to those people who they are talking to (Herman, 1961; Labov, 1966; Giles, 1971), affecting phonetic realisations of speech features.

If the judiciary or forensic phonetic practitioners in the UK were to assemble large scale databases of language which were representative of all accents, the cost would be prohibitive, even if accents were stable and located in bounded areas. Changes in accents would mean that databases would have to be updated much more regularly than for DNA, for example. Furthermore, accents not only change in themselves, but can be multifaceted within speakers: what is the reference population for a perpetrator

determined to have an upper class Liverpool accent, but who has some SSBE features. Maybe they have developed a TRAP/BATH distinction, yet still carry Liverpool features? What if that speaker was recorded in a situation where the interlocutors were RP speakers in a position of power, increasing the likelihood that the perpetrator would accommodate to an accent variety that they are familiar with? There are many situations which affect intra-variability of speech parameters in the same individual. This is unlike DNA and other fields where LR approaches have been successfully applied, where for example, allele distribution remains stable within the individual for the course of their life (Aitken & Taroni, 2004).

2.3.5.2 Qualitative judgements

A point raised by French et al. (2010) pertains to those features which do not lend themselves to quantitative analysis. Voice quality (VQ), for example, can be a very idiosyncratic feature, and is a widely used and powerful discriminant feature, but some aspects of voice quality can be very difficult to describe quantitatively. Although some numerical parameters can reflect voice quality differences, experts largely categorise voice quality using auditory means, sometimes in formalised scales. Within a strictly numerical LR approach, it would be very difficult to assign probability functions to parameters that are not entirely quantitatively measured. In some cases (such as qualitative voice quality judgements) it is necessary to use the LR approach as simply a theoretical approach (Rose, 2002). It is unclear how these features are combined with numerical LR scores, for example.

2.3.5.3 LR estimation formulae

First, it is important to remember that LRs are not *directly* related to certain truths, but are an estimation of the probability of evidence, given two competing hypotheses. It is impossible to know the true LR for pairs of samples; the LR is just an estimation of that truth. This is why the LR method has to be tested in its accuracy and precision in estimating that truth (Morrison, Thiruvaran, & Epps, 2010). There are a number of ways of estimating those likelihoods, which provides another problem for analysts and the court (and presumably presents another issue of standardisation). Lindley's univariate LR formula was initially used to model LRs (Aitken, 1995), but this is insufficient in dealing with correlated data (such as speech).

To deal with cases where a number of features have to be combined, Aitken and Lucy (2004) subsequently proposed a multivariate likelihood ratio, which can combine traditional parameters that may be strongly correlated, such as F2 and F3 (Kinoshita, Ishihara, & Rose, 2009). That is to say, it can combine these in a more accurate way. This formula is still only designed to take account of two levels of variance which is not sufficient to characterise a system as complex as that of speech production. Kinoshita et al. (2009) call this a 'leftover' from its origins in dealing with glass fragments, as glass fragments, like DNA, are invariant in the course of their 'lifetime'. This is not totally analogous to speech data, which can be highly variable within a speaker due to a plethora of factors.

2.3.5.4 Trier of fact (juror) understanding

One of the biggest concerns for the LR approach, and any technical or scientific evidence, is how jurors understand the evidence presented to them. Although a minority of practitioners believe the role of the expert to be purely presenting the LR (Morrison, 2010e), it is also argued to be the role of the expert to make sure that their evidence is comprehensible to and comprehended by jurors (Aitken & Taroni, 2004; Robertson & Vignaux, 1995; Lynch & McNally, 2003). In the UK context, jurors are almost completely laypersons with no formal statistical training. Robertson and Vignaux (1995, p. 54) state: "Part of the task of the expert witness should be to explain how and why the court is helped by the evidence they are giving". Essentially, forensic evidence cannot be considered 'fit for purpose' if it cannot be processed by the trier of fact. (cf. R v Adams ruling, § 2.3.5.6). There are also issues with jurors' perceptions of evidence presented using an LR approach which is not based on numerical calculation, where this may give a false veneer of science.

2.3.5.5 Verbal scales

Champod and Evett (2000) propose a verbal scale to phrase numerical LRs to triers of fact (this scale was used by the UK's Forensic Science Service). This converts LR scores into a Log_{10} scale (converting to log likelihood ratios, or LLRs) which is then expressed as support for either the defence or prosecution hypotheses using the following phrases:

Table 16 - Scale of LRs and strength of verbal support for the evidence

Likelihood Ratio	Log_{10} LR	Verbal expression
>10000	5	Very strong evidence for the prosecution hypothesis
1000-10000	4	Strong evidence for the prosecution hypothesis
100-1000	3	Moderately strong evidence for the prosecution hypothesis
10-100	2	Moderate evidence for the prosecution hypothesis
1-10	1	Limited evidence for the prosecution hypothesis
1-0.1	-1	Limited evidence for the defence hypothesis
0.1-0.01	-2	Moderate evidence for the defence hypothesis
0.01-0.001	-3	Moderately strong evidence for the defence hypothesis
0.001-0.0001	-4	Strong evidence for the defence hypothesis
<0.0001	-5	Very strong evidence for the defence hypothesis

Source: Champod and Evett (2000)

While this scale can make numerical LRs seem more understandable to jurors and judges, there are problems with it. Firstly, it appears to suffer from the same cliff-edge effect as explained in §2.3.4.1. Secondly, the terms 'limited' and 'moderate' are likely to be interpreted in different ways by different individuals. It is also the case that many mathematical concepts and processes are not well understood by potential jurors or even legal experts, so it is unclear to what extent these groups would be aware of logarithmic scales. One suggestion might be to aid jurors where likelihoods such as these are presented with real world exemplars of events with similar probabilities.

2.3.5.6 R v Adams Case

Lynch and McNally (2003) discuss the use of LR based evidential conclusions with reference to a particular rape case in which DNA evidence was crucial (R v Adams, 1996). They examine the postulate that "justice systems struggle to incorporate DNA [complex] evidence into a system of justice that stresses lay participation and public accountability" (2003, p. 85). They argue that jurors have 'little acquaintance with, and very little opportunity to learn about' the key technical aspects of experts' presentations of evidence. The expert in this case attempted to get the jurors to calculate the posterior odds in court, by applying subjective judgements of their own priors, or at least estimating pragmatic prior estimates (such as the chance of the perpetrator being a resident in the area), in combination with the DNA LR. Although this case was fairly novel, it did show how the jury were ill prepared to deal with LR based testimony. The Court (R v Adams, 1996) released the following advice:

> we have very grave doubt as to whether that evidence [using Bayes for non-DNA evidence] was properly admissible, because it trespasses on an area

> peculiarly and exclusively within the province of the jury, namely the way in which they evaluate the relationship between one piece of evidence and another. (Paragraph 15)

The Court was wary of using Bayesian estimation of priors explicitly, for the following reasons:

> The mathematical formula, applied to each single piece of evidence, is simply inappropriate to the jury's task. Jurors evaluate evidence and reach a conclusion not only by means of a formula, mathematical or otherwise, but by joint application of their individual common sense and knowledge of the world to the evidence before them. (Paragraph 15)

The judge concluded the report by stating that:

> We regard the reliance on evidence of this kind in such cases as a recipe for confusion, misunderstanding, and misjudgement, possibly even amongst counsel, but very probably among judges as well, and we conclude, almost certainly among jurors. (Paragraph 12)

Although this case was very peculiar, the fact that the expert witness guided the jury through the application of the LR to their own understanding of the case is more training than would be offered in a normal case. The task the expert and jury performed in the courtroom is an application of Bayesian statistics, that is, telling a jury "how they should rationally update subjective, probabilistic beliefs in light of evidence" (Drygajlo, 2010). This is the task that jurors are faced with when presented with an LR.

In this case they were unable to process the LR with comprehensive guidance. It would thus be difficult to see how a jury processes the LR unguided. There is the very real possibility that the Court does not know that part of its responsibility is to apply its own understanding of the rest of the case (prior odds and other LRs) to the LR at all, and just follow the findings of the LR. This could lead to the erroneous interpretation that if for one piece of evidence the LR=1000, the suspect is 1000 times more likely to be guilty. Even if an LR is logically and legally correct, if it is delivered to an audience who are not able to process it, then it does not fulfil its purpose adequately.

2.3.5.7 Understanding prior odds

The main problem in the *R v Adams case* was the calculation of prior odds. The key issue with juror priors is not just their calculation, but the idea that prior odds do not always reflect the conditions of the case. Morrison (2010e) contends that for theoretical work,

you can attach pragmatic priors, e.g. the number of men in a house at the time a call was made. Though, of course, in the practical setting this is the trier of fact's jurisdiction.

Thompson (1989) raises the point that in many cases jurors' underlying assumptions outweigh a rational appreciation of a probabilistic system for evidence. He raises the (slightly odd) example of 100 suspects, where ten are lawyers and 90 are builders, asking whether most jurors would assign a straight 1/10 probability to the chance that the perpetrator is a lawyer or a builder or whether individual bias about the kind of person who follows each profession would affect their judgement. Lindh et al. (2010) stress that prior odds need to be discussed more fully, especially as they have such a big impact on the outcome of a Bayesian exercise (Robertson & Vignaux, 1995).

Lindh et al. (2010) also contend that the LR approach is dependent on the belief that all jurors share the same prior odds, and that this is rationally determined. If not, if the individual jurors frame the LR within their own 'discourse' or bias, then the conclusion will be disparate within a jury. Indeed, little research has considered how juries behave as a group, and particularly in response to complex evidence.

2.3.5.8 Studies into juror understanding

Lewontin (1991) argues that juries are simply not trained or prepared to be dealing with complex probabilistic statistics such as those associated with the *R v Adams* case. He points to undergraduate statistics students, stating that they take months or years to come to terms with the mathematical reasoning, and they are in favourable conditions to do so. This echoes the question that even if our evidence is scientifically and logically accurate, but triers-of-fact cannot interpret it correctly, is it 'fit for purpose'? Research in this section seeks to answer that question.

McQuiston-Surrett & Saks (2009) tested a number of different logical and anti-logical conclusion types, including experts 'impinging on the Ultimate issue' (i.e. stating a probability of the suspect's guilt) and providing qualifying statements about the limitations of forensic evidence. They presented these different conclusions to legal practitioners. In a simple test of calculating a LR with a random match probability, only 41% of subjects correctly calculated the LR, and this was only 25% where the test was made more complex (i.e. the formula involving a decimal place). The problems of legal and logical fallacies were overshadowed by the subject's inability to carry out simple

mathematical tasks. They found that there was little difference in fact finder's decisions of guilt if evidence statements committed fallacies, and that there was little difference between judges and jurors, although jurors were more cautious.

They conclude that:

> Experts in most forensic fields have no empirical RMPs (random match probabilities) to share, but only intuitive guestimations which they customarily express in qualitative terms...While some witnesses temper [fact finders'] impressions...others reinforce the misunderstanding [that evidence constitutes a unique match] to achieve an exaggerated belief (McQuiston-Surret & Saks, 2009, p. 437)

Thompson and Schumann (1987) also set out to test how likely it was that jurors (in this case university students) were to commit one of the fallacies (§2.3.2.4) in evidence which was presented, either by a percentage conditional probability, i.e. 'a 2% chance the defendant's hair would be indistinguishable from the perpetrator if he were innocent' versus the same statement accompanied by a random match probability. Having been presented with realistic evidence statements, some featuring fallacies, a quarter of subjects were judged to have fallen victim to one fallacy, with the first conclusion [no RMP] largely inviting the Prosecutor's fallacy and the second [with RMP] leading to the Defence attorney's fallacy. They also note that most subjects revised their initial impressions of guilt based on other factors of the case once they were given these conclusions, but not to the extent that a Bayesian type analysis would demand.

In a second experiment, a similar group of subjects were presented with fallacious statements about evidence, and only 22% could recognise errors. 68.5% of participants [wrongly] labelled a statement featuring a Defence-type fallacy correct, while a larger number could [correctly] recognise the Prosecutor's fallacy as incorrect (72%). They conclude that people are not good at drawing correct inferences from associative evidence and incidence statistics and that they are unable to see crude fallacious errors in interpretations of evidence. They also raise the interesting point that although many studies investigate an individual's performance in understanding evidence, few studies have examined the decision making process in a group, such as a real jury (Lindh et al. (2010) also raise this point with specific reference to prior odds). This is clearly important if we are to understand the impact of evidence statements on juries in authentic settings.

Taroni and Aitken (1998a; 1998b) examined the way in which the statistical presentation of evidence can influence a verdict, by presenting different types of statistical conclusions to forensic medicine and science students (who had undertaken at least a year's training in probabilistic methods) and legal practitioners. These conclusions included different LR scores and also different methods of expressing them, by percentage exclusion or in frequency form (i.e. 10000 to 1). Both students and practitioners varied in their calculation of the posterior odds based on LR conclusions. Those subjects presented with LRs and posterior odds were less likely to pass guilty verdicts and "a reasonable proportion of subjects failed to detect errors in the arguments made by experts at trials" (Taroni & Aitken, 1998a, p. 173). The report which the authors assigned as the 'best' assessment, using an LR, was the least understood and misinterpreted as wrong by 28 lawyers – 60% of legal participants. These findings lead them to conclude that evidence is not correctly interpreted, even by future legal experts, and that jurors need better statistical education.

Cudmore (2011) compared jurors' responses to numerical LR ('LR=5') and verbal LR ('moderate support for prosecution hypothesis') conclusions for a FSC case, in terms of support for hypothesis and the level of guilt perceived. She found that numerical expressions lead to fewer subjects (correctly) attributing evidence as supporting the prosecution, although those subjects with a numerical LR were also less likely to commit the Prosecutor's fallacy. In general numerical LRs were found to be confusing and 'jurors' preferred the comprehensibility of a verbal interpretation.

In general, although the LR might be the 'logically and legally correct' method for expressing strength of evidence estimates (Rose & Morrison, 2009), there are strong questions raised about how jurors and even law/legal statistics experts understand or are influenced by different conclusion frameworks. Certainly there needs to be more research into trier-of-fact understanding (including how they behave in groups) if forensic evidence can be considered 'fit for purpose'.

2.3.5.9 Bayesian inference

One of French et al.'s (2010) principal concerns with the use of LRs is in those cases where empirical background data are not available to form a suitable reference population - which is, in practice, in most UK cases. It is possible to use the LR approach to express conclusions based on the expert's knowledge and judgement, and assign mathematical

scores to subjective judgements. Aitken and Taroni (2004) describe this Bayesian inference as quantifying 'the degree of belief of a certain individual in the occurrence of a certain event'. For example if an analyst estimates an event or piece of evidence to be ten times more likely given H_{ss} they can assign LR=10 to that feature (Robertson & Vignaux, 1995). Rose (2002) champions this approach in those cases where reference population data are not available.

There has been academic and regulatory opposition (described below) to the use of Bayesian inference for forensic evaluation of evidence, as it risks giving subjective judgements the veneer of empirical science. The Law Commission's (2011) report on forensic evidence expressly condemns:

> Expert evidence which is presented as scientific...there is a danger that juries will abdicate their duty to ascertain and weigh the facts and simply accept the experts' own opinion, particularly if the evidence is complex and difficult for a non-specialist to understand and explain (paragraph 1.9)

Halliwell et al. (2003) show that subjective probability assessments are not generally precise and it has been claimed that it is 'misleading to seek to represent them precisely'. They quote a paper relating to risk assessments from the committee of the US National Research Council (1981):

> [Members have] an important responsibility not to use numbers which convey the impression of precision, when the understanding of relationships is less secure. Thus whilst quantitative risk assessment facilitates comparison, such comparison may be illusory or misleading if the use of precise numbers is unjustified.

There appears to be little research or guidelines relating to subjective Bayesian expressions of strength of evidence, this seems to be an area where future research is warranted.

2.3.5.10 Intra-speaker variation

One of the reasons that these subjective judgements have been suggested is that speech data differs from other forms of evidence, such as DNA, in that it is highly variable within an individual. One of the aspects of the present project concerns one type of intra-speaker variation (over time), but the voice is subject to a complex arrangement of variable factors, including stress, health, age, social setting, language variety spoken and many more. The arrangement of DNA alleles remains consistent through an individual

and relatively constant across populations (Aitken & Taroni, 2004), but for speech there is a high level of variation, from short-term differences in complex articulatory gestures to longer term physiological changes such as those highlighted in §2.1. Rose (2002) highlights the fact that no two instances of a word are ever realised the same way.

Robertson and Vignaux (1995) present a comparable example from ballistics, where the unique barrel inflicts certain recognisable patterns on the bullet that it produces, this pattern is unique at the time of firing, but the qualities of the barrel may change over time (especially between a evidential and test firing sample). In cases such as these they state that "it seems impossible to create probability models for traces left by rifle barrels...which may change characteristics over time...this is due to the number and complexity of the factors to be considered" (Robertson & Vignaux, 1995, p. 58). They stress that in these analyses, experts rely on experience to inform judgements about the strength of evidence. LR-based approaches are deemed implausible for rifle barrels, for which variation is limited to external factors (such as cleaning or firing). This is not as multiplex as the huge conglomeration of factors which affect speech output; therefore is it not plausible to develop a statistical model for speech, according to Robertson & Vignaux?

2.3.5.11 Determining potential perpetrator populations

Determining reference populations may be straightforward for DNA analysis, for example, where large population databases are available and analyses are rarely limited (i.e. sometimes according to racial categories (Buckleton & Walsh, 1991)). Determining potential perpetrators for speech LRs is more difficult. Rose (2010, personal communication) describes how the analyst can make sensible limitations based on the features present in evidential samples (i.e. excluding sex or age types). French et al. (2010) point out that exclusion based on judgements of sex or age is a p(H|E) type decision, and that a number of these kinds of decisions are made by the analyst in other parts of the FSC process (in editing and preparing the samples, including which parts of the sample are analysed). Furthermore, legal cases are not inflexible and new evidence can be presented in court. As speech databases would have to be more fragmented than the national DNA database due to accent differences, new evidence in a case (such as a new secondary suspect, or a witness statement talking about accent) might lead to a

reference sample becoming useless and a new sample being needed. This would have implications for the time and costs needed to compose a reference sample.

2.3.5.12 Reference population databases

Aitken and Taroni (2004) surmise that transfer or trace evidence is amenable to statistical analysis 'because data are available to assist in the assessment of variability'. However, in the case of speech evidence, this is precisely the problem: data are not available in most cases. Nolan (2001) argues that there is simply not the data available to carry out overtly numerical LR analyses, and that analysts should move towards providing $p(E|H)$ statements in a theoretical way only (i.e. stating the competing hypotheses and which the expert believes is strongest, backed by evidence). Morrison (2009b; 2009d) argues that if DNA databases could be collected, then it should not be a problem, with significant investment, to solve the same problem for speech: "The challenges should not, I think, be harder for forensic scientists working in other fields [than DNA]" (Morrison, 2009d, p. 160).

It does seem, however, to be much harder for speech scientists, due to the plasticity of speech within a speaker in accordance with an array of different factors and the huge variety of dialects in the UK. Along with practical concerns, investment is a key problem, especially considering recent cuts to police budgets (14% reduction to police resource funding from 2009-2015 (HM Treasury, 2010)) and reduced external police spending in forensic areas. It has already been mentioned in §2.3.5.1 that due to accent variation, the process would be hugely expensive and difficult and would need frequent reinvestment. The problem not only entails the cost of collection, but the nature of the reference population for each case. To truly reflect the H_{ds}, the reference population should be tailored very closely. Even if an estimate based on the features present in the evidential sample were to be made, most analysts working in the UK would extract a bespoke set of requirements from an evidential sample which would be expensive to collect a reference population for. This idea was voiced by the National Manager of Forensic and Data Centres of the Australian Federal Police, who warned experts about the likely costs of forensic exercises, and warned against turning cases into research projects ((Robertson J. , 2007) cited in French et al. (2010)). Given the lower evidential status of speech evidence against DNA, the police are unlikely to commission a large scale costly investigation on a case-by-case basis, given that the average cost of forensic examinations at the MPS was

around £850 in 2010/11 (across all types of evidence) (Metropolitan Police Service (FOI request), 2011).

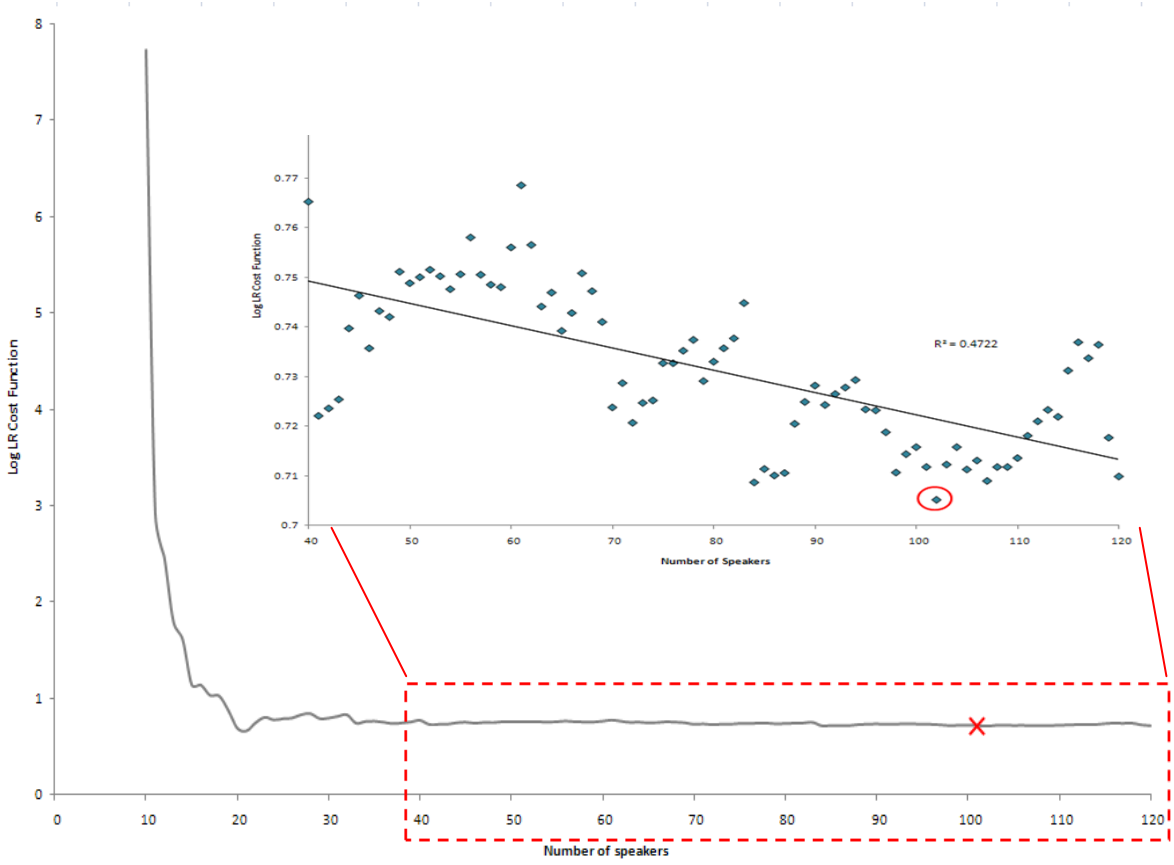
There is insufficient research on the use of reference populations and how specifically tailored these need to be, what size performs to an acceptable standard and how closely mapped they need to be for age, location and background, for instance. For example, a number of findings in Rose (2002) are based on the Bernard data set (Bernard, 1970), which was recorded in Australia in the late 1960's. Although Alderman (2004), in an undergraduate dissertation on the Bernard dataset, states that the set performed well as a reference population, this statement has no comparative basis as it was not tested against a more current population which matched his sample of speakers. Loakes (2006) expressed doubts about using this same dataset as a reference population as it contained speech patterns that were different from her twin subjects. Her results show that all of her speakers were strongly discriminated from this group, largely due to sound changes over the past 40 years. If this were the case in real forensic practice it could be severely dangerous as it would prejudice the LR heavily in favour of the prosecution. Imagine, for instance, a UK reference population of one variety of Received Pronunciation from the 1970s, with a group of homogenous speakers. If two similar-sounding (as might be in a case) SSBE speakers, recorded in a suspect and evidential sample in 2010, were compared they would both score highly on similarity, but would be classed as atypical on features which have developed and changed significantly since standard Received Pronunciation was in general use. This is precisely the kind of question that the non-contemporaneous data in this study addresses.

2.3.5.13 Reference population research

One of the only studies into the issue of reference population composition is reported by Hughes (2012). He investigates three key components of reference populations make-up using /u:/: accent mismatch between reference sample and test samples (evidential and suspect), number of speakers and number of tokens per speaker. Using data from the Origins of New Zealand (ONZE) corpus (Gordon, Maclagan, & Hay, 2007) as a reference population and as test data, he was able to test (in a round-robin format, removing one from the sample at a time) the effects of different numbers of speakers (from N=120-1) and tokens (N=13-1) on the error rate of LLR calculation (using C_{llr}). For speakers, there was a lower cut off of around 20 speakers, below which performance deteriorated

dramatically, while performance was still significantly improving with increased speakers beyond N=100 (this effect can be seen in Figure 13 below, with a steep cut off at eight speakers).

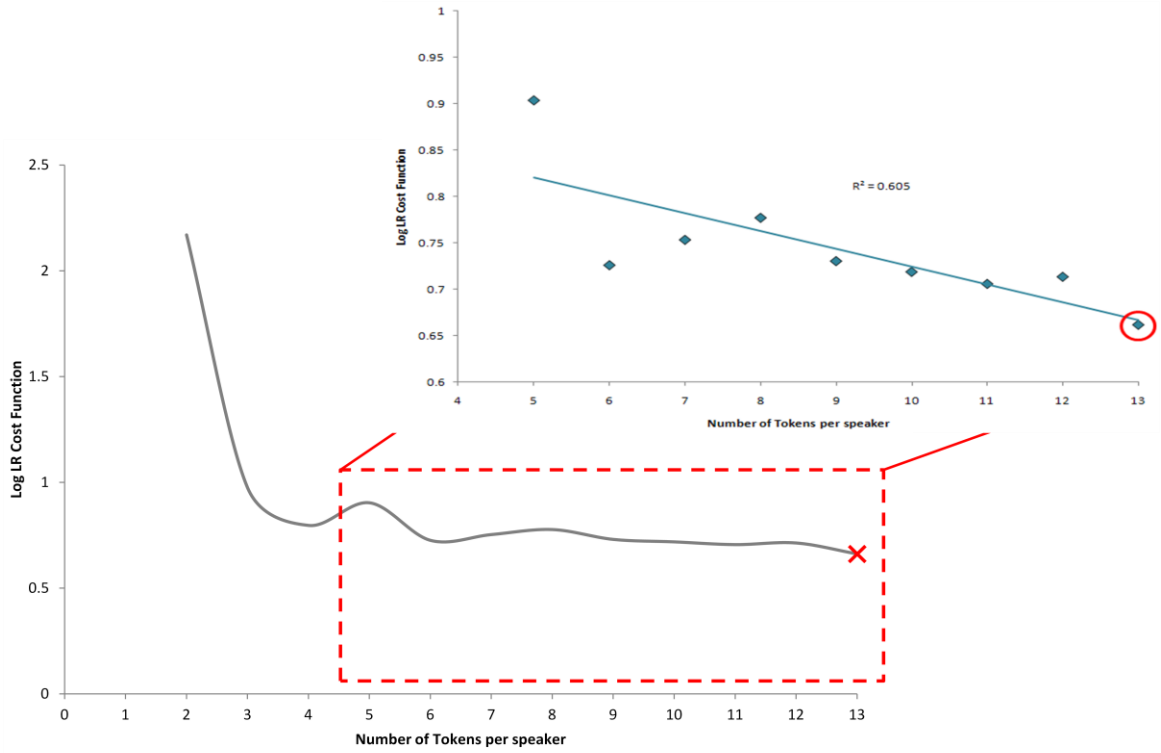
Figure 13 - Log-LR cost function against number of speakers



Source: Hughes (2012)

There is a similar effect for vowel tokens:

Figure 14 - Log-LR cost function against number of vowel tokens



Source: Hughes (2012)

Further to variance in the number of speakers and tokens of /u:/, Hughes tested samples from different accents (Manchester, York and Newcastle English) against the New Zealand reference samples. He found, as predicted, that LLRs for same-speaker pairs were much higher in mismatched accents, in most cases by a magnitude of one (which would entail the next level of support in a verbal LR scale such as that proposed by Champod and Evett (2000)). Perhaps more worryingly in a forensic sense, different-speaker pairs from Manchester and Newcastle yielded high numbers (60% or more cases) of false positives (i.e. LRs higher than 1, or LLRs higher than 0). We would perhaps expect both of these findings, as a reference population which was highly different would naturally favour (in many cases falsely) the prosecution hypothesis, as the test samples would be deemed untypical with respect to the reference population (as typical values for different features would be different i.e. for vowel formants). These are very important findings which stress how significantly LRs can be affected by the composition of the reference population. We should expand on Loakes' (2006, p. 197) statement that "the reference sample must be representative of the population [otherwise] the resulting LR will be misrepresentative" *in favour of the prosecution hypothesis.*

The make-up of reference populations, defined by the defence hypothesis, demonstrably needs more careful examination and testing. This project examines the effects of different age properties of test samples and reference populations on LRs, as part of an investigation in the effects of changing defence hypotheses.

2.3.6 **Questions**

This section poses questions about reference populations which this study investigates with respect to differences in age.

4 What effect does age mismatch between criminal and suspect recordings have on LR estimation?

2.4 Research Questions

Previous sections have presented a number of research questions which are collected and repeated here for clarity.

Question 1 relates to sociolinguistic methods of data collection and sampling and is addressed in §4.6.1:

1 What is the extent of vocal instability in adulthood and how will this affect the interpretation of results from apparent-time studies?

Questions 2-4 relate to forensic contexts in particular; question 2 is designed to explore the effects and consequences of age-related changes to frequency characteristics commonly used in forensic exercises. Question 2 is addressed principally in chapter 4 (§4.6.2), but the themes are explored throughout the later chapters.

2 What is the magnitude of change in individuals' vocal output during adulthood?
   a Which features remain more stable than others?
   b To what extent are changes predictable from a model of sociolinguistics or gero-physiology?
   c What effect should this have on how we evaluate forensic speech evidence?

Question 3 is directed at one forensic approach to analysing vowel diphthongs and examining its efficacy in discriminating between speakers in a non-contemporaneous context. This question is addressed in §5.3.1; 6.4.1.

3 What is the discriminatory performance of FD measures of different polyphthongs in different varieties of English in spontaneous speech?
   a How do they compare with traditional measures of monophthong formants?
   b How do formant transitions change during adulthood?
   c Is making dynamic measures of vowel formants worthwhile?

Question 4 originates from concerns about the dearth of research relating to reference population composition and the subsequent effects this has on LR scores in forensic casework. Question 4 is addressed by chapters 6 and 7.

4 What effect does age mismatch between evidential and suspect recordings have on LR estimation?

3 Method

This chapter describes the methodology of data collection and extraction, summarises the subjects and materials and presents analysis techniques. Following this, it outlines statistical interpretation for those materials. It also discusses issues with longitudinal data that are raised in the sociolinguistic literature and considers the effect that variation in instrumental techniques can have on potential results.

3.1 Subjects

This project examines data collected from eight subjects. The subjects were all born in 1957, and recorded at 7 year intervals from the age of 7. These speakers were taken from a series of Granada television documentaries produced and (mostly) directed by Michael Apted. The initial programme is entitled *7 Up* (Apted, 1964) and there are six subsequent programmes at 7 year intervals; as such, this documentary offers a unique opportunity for real time study of longitudinal changes to the voice. The latest programme considered by this thesis, *49 Up,* was released in 2007 and shows subjects at age 49. This is not the first time this resource has been used in a linguistic investigation; Sankoff (2004) looked into two vowel changes in two speakers in the series.

Although there are more speakers in the documentary, the speakers investigated are those who provided enough speech to be included, and also those that were recorded at sufficient intervals throughout the series. The sampling method that Apted followed for recruiting subjects is not totally random, and although it was not based on linguistic factors there may be implications of the selection on language varieties included. However, the production company did select subjects from different areas of the country and those at extremes of the social spectrum (noted in DVD commentary (Apted, 42 Up, 1998)) and in some cases the accent and linguistic patterns of the speakers reflects this.

The subjects included in this project are summarised in Table 17, in terms of their location, accent, mobility and any other factors which might have an effect (accents were determined mainly by location, but also from close auditory analysis with reference to material such as Wells (1982), Hughes et al. (1996) and Ferragne and Pellegrino (2010)). This information is based on the first three programmes, ages 7, 14 and 21. Some speaker's locations had already changed by age 21 so the table includes their 'home'

location from earlier programmes, which, according to sociolinguistic theory, is when accents become stabilised.

Table 17 - Participant matrix with summary of biographical information

Name	Location (1964)	Accent 'home' (1964)	Accent at 21 (1978)	Mobility across series	Other locations	Other potential factors
Andrew	Kensington, London	RP	RP	Limited	Cambridge University	
Bruce	SE England	RP	RP	Limited	Oxford University, East London, Hertfordshire	Teacher (professional voice user)
Lynn	London	Cockney	Cockney	Limited		Smoker, physical health
Neil	Liverpool	Liverpool	SSBE/RP	High	London, Scotland, Shetlands, Hackney, Cumbria	Ongoing mental health issues
Nick	Yorkshire Dales	Yorkshire	Yorkshire/SSBE hybrid	High	Oxford University, California, Wisconsin, USA	Lecturer (professional voice user)
Symon	London	London, some afro-Caribbean features	London, A-C	Limited		
Suzy	S England, Scotland	U-RP	U-RP	Moderate	Bath	Smoker at 21
Tony	East London	Cockney	Cockney	Moderate	Essex, Spain	Part-time extra (professional voice user, training)

3.2 Preparation

The first stage of data preparation was to extract sections of the video files which feature speech from the subjects in an interview setting (see §3.3.3). *Edius v5* was used to create concatenated video and audio tracks of speech from individual subjects. These tracks were then converted to individual video files. The recordings were originally in .avi video format (.ac3 format audio), so .wav files of the audio track were extracted onto a single mono track using *Edius 5* in order for the audio to be analysed independently.

3.3 Summary of materials

Although there are currently seven programmes published, this study is only concerned with changes in the voice across the adult lifespan, and so in the first recording analysed in this study, all subjects are aged 21. Therefore the five age stages, with seven year intervals are:

21, 28, 35, 42 and 49 years

3.3.1 Samples

The five episodes for this age range have a running time of 769 minutes. This is not distributed equally across all speakers or age stages, potentially due to the interest shown in different subjects from a popular audience, and also the sociological interest from a production viewpoint. Some speakers therefore have more coverage than others. However, a large majority of the speech in the series is from recorded interviews with the subjects, which presents high-technical-quality spontaneous speech data. Samples are at least one minute per stage and up to 15 minutes for some speakers. Furthermore, even in cases where there is only a short recording for some age stages, this is not uncommon in the forensic condition, and while this might limit the strength of pattern we can derive for that particular speaker, further analyses from those subjects are comparable with a forensic exercise. Below is a summary of the data for each speaker at each age stage:

Table 18 - Summary of length (seconds) of speech extracted from *7 Up* series at each age

Subject	21 years	28 years	35 years	42 years	49 years
Andrew	130	94	114	156	174
Bruce	667	(323)	325	248	342
Lynn	159	77	201	316	322
Neil	918	766	491	385	469
Nick	239	402	349	373	368
Symon	401	336	-	228	287
Suzy	277	306	285	297	209
Tony	415	480	472	277	390

Symon did not participate in '*35 Up*' for personal reasons and data for Bruce aged 28 was insufficient in terms of formant measurements due to conflicting background noise.

3.3.2 Technical quality of recordings

It was not possible to ascertain the exact specifications of recording equipment and editing procedures for the materials. One possible technical issue might be the mismatch in recording equipment from the 1960s through to the 2000s. However, the audio is high

quality and the DVD versions were produced at standard DVD quality (48 kHz sampling rate, 16 bit, mono-split 2 channel, .ac3 format audio). The recordings are excellent quality, especially in contrast with FSC materials, and this entails excellent spectrographic resolution. This allows for acoustic analysis with greater accuracy than in forensic cases.

3.3.3 **Criticisms of RT studies and this dataset**

It is worthwhile pointing out how this dataset responds to the three main criticisms of RT studies put forward in the sociolinguistic literature. Tillery and Bailey (2003) highlight three main concerns when working with RT data (to put these in context, they address RT studies of sound change in a community):

i. Participants may move or die
ii. The sample becomes less representative
iii. Speech must be recorded in the same context

In response to these concerns and regarding i, in fact some speakers removed themselves from the set which may have caused a problem to the producers, but in this case it was simple to identify those eight speakers who feature throughout. In terms of subjects moving, this is entirely acceptable for the present study as one of the aims is to assess the possible magnitude of the effect of geographical location changes on the voice: the study is focussed on individual cases, not in groups or communities. Quite simply, for the sake of forensic validity, if potential criminals move, a study of language must include that possibility. Point ii, as with geographical mobility, is more a concern of community studies, as in this case it is impossible for a speaker to become unrepresentative of themselves as an individual.

Point iii is important, but rarely satisfied by forensic materials. Nevertheless, the interviewer in the *7 Up* series remains constant throughout, as does the style of interview (a consistent interviewer may also prevent effects of interlocutor accommodation which are shown to affect some parameters of speech: ((Gold, 2012) for clicks). Although it is clearly not a typical speaking environment, i.e. speaking in front of a television crew for a programme intended for national broadcasting, at least it is constant throughout the documentary. Moreover, the spontaneous speech style is constant, along with the 'home' environments in which the interviews normally take place. For the present study,

the environment for the data satisfies these concerns, and is very unlikely to be a significant factor in the variation demonstrated by the speakers.

3.3.4 **Establishing intra-speaker variability: short-term non-contemporaneity**

A benefit of the *7 Up* dataset is that each sample is not simply one recording, they are mostly two or three separate recordings made on different occasions. In some cases they are clearly made at different times of day. These kinds of recordings are usually made over a period of a few weeks or months, and this provides a useful amount of natural intra-speaker variability. According to Rose (2002) and subsequently Morrison (2010a; 2010c; 2010e), it is vitally important when carrying out research into the speaker discriminating potential of features, to make use of short-term non-contemporaneous data. This means that estimations of discriminatory power capture intra-speaker variation over time. Moreover, these recordings are relevant to forensic recordings which are usually at least short-term non-contemporaneous, and may comprise of a number of recordings from different times. The current dataset satisfies this criterion.

3.3.5 ***7 Up* Video**

A further benefit of this dataset for future research is that most speech by the participants is from recorded headshots, providing video of the speaker which may prove extremely useful for some parameters. In voice quality analysis, for example, video is extremely useful for corroborating auditory judgements of features such as lip rounding or spreading degrees or even jaw opening.

3.4 **Parameter extraction**

3.4.1 **Equipment**

The following equipment and software was used to perform acoustic analysis: *Praat* v.5.0.22 (with forensic add-ons developed by Philip Harrison) was run on a *Sony Vaio VGN-FW51JF* laptop computer using a *Realtek ATI HDMI* sound driver. All acoustic and auditory analysis was carried out using *Sennheiser HD 280 Pro* closed-cup style headphones.

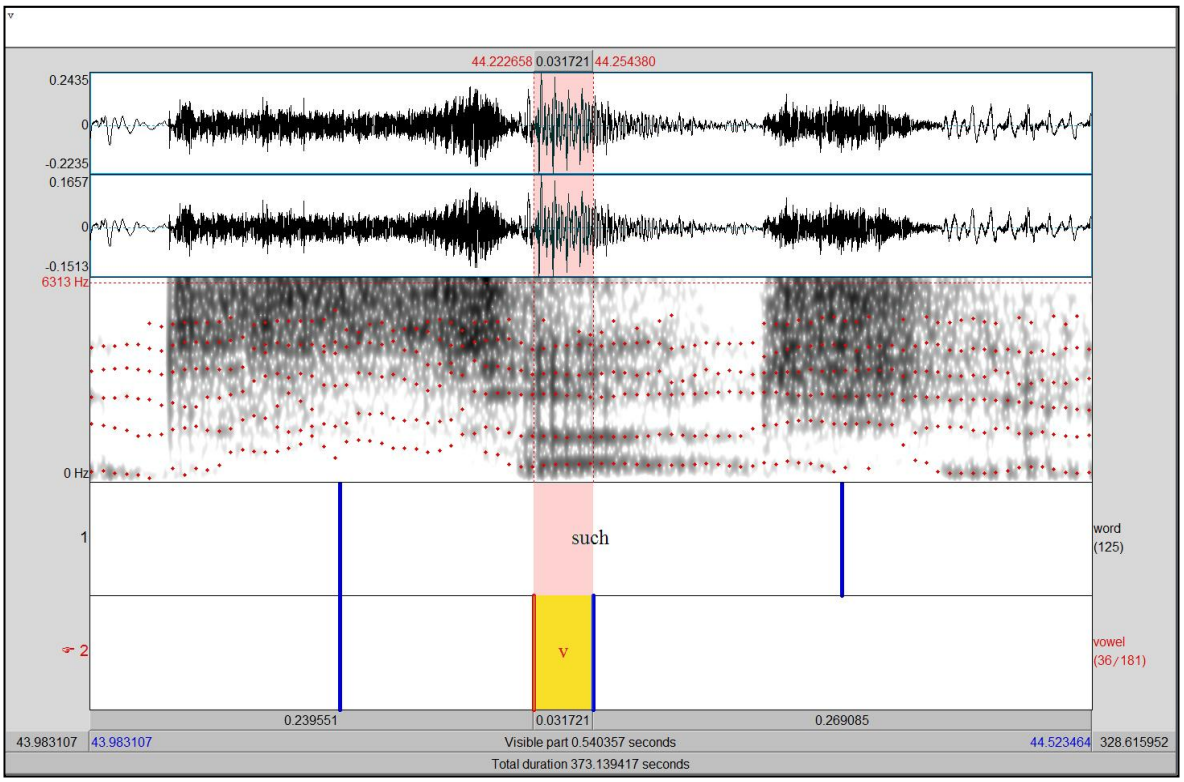
3.4.2 **Features analysed**

A number of features were analysed, particularly those that were predicted to show an effect of age, and those which are prevalent in forensic research projects and in casework (Gold & French, 2011).

3.4.2.1 Monophthongs

Monophthongs in stressed syllables were extracted using a script (developed by Philip Harrison) which records the formant frequency for F1-4 as well as time, *number of formants* setting and also any information the researcher wants to input. These formant measures are recorded from *Praat*'s LPC formant measuring tool (sometimes erroneously referred to as a tracker), the settings for which were modulated for each individual speaker; where possible, these setting were kept consistent across recordings from individual speakers so as not to cause differences in formant measures due to the LPC tracker itself (although of course these setting differed for male and female speakers). Within the audio files created in *Edius*, monophthongs were marked using a text grid and the script was used to extract an average value across the central point, or the target section, of the vowel.

Figure 15 - Example of a text grid in *Praat* used to extract monophthong vowel formants



This average was taken across a central section (somewhere around 30-50ms) to avoid variability in single points. The central point of the vowel was determined by a number of factors working in harmony; principally by the stability and intensity of the formant bands on the spectrogram, but also by using *Praat's* intensity measure and amplitude in the waveform to indicate the nucleus. The former features were the main indicator, however, as the latter could be influenced by changes in degree of mouth opening across different combinations of sounds. Methodological concerns have been raised about the efficacy and accuracy of LPC formant trackers, particularly in *Praat* (Harrison, 2011). To try and counter this effect, the accuracy of LPC tracks was checked manually when applying the script and where the tracker was inaccurate formant measures were corrected by hand. The script allows these to be input through the same interface so information regarding settings and time, for example, is preserved. Even so, Harrison reported errors in *Praat*'s formant tracker, using an automated analysis, of 10% Hz differences in either direction, and this should be kept in mind (for all formant-based studies).

The aim of this study is to emulate as closely as is plausible those conditions facing forensic speech analysts. In order to investigate real, spontaneous speech, as many vowels as possible were taken from each sample, from almost all contexts. However, some consonantal contexts were excluded due to their coarticulatory interference with the vowel in question. Although all consonants have a coarticulatory effect (Clermont, 2009), /r/, /m/ and /w/ contexts were deemed to have too significant an effect on the neighbouring formants. This is also recommended by Turk et al.'s (2006) guide for segmentation, although they also include initial /l/ in this 'avoid where possible' bracket, which was not excluded from the present study as its effect was determined to be no more significant than other consonants.

In most cases the following monophthongs were extracted, although frequency of tokens varied between speakers, mostly dependant on the length of sample. Those vowel categories in brackets were not sufficiently present in all speakers' data and are therefore excluded from general analyses:

/iː, ɪ, e, a, ɑː, ɒ, ʊ, ʌ, uː,(əː, eː, ɔː)/

(or the FLEECE, KIT, DRESS, TRAP, START/BATH, LOT, FOOT, STRUT, GOOSE, (NURSE, SQUARE and NORTH) sets (Wells, 1982))

There were potential issues in classifying vowels for speakers who showed mobility, i.e. in the case of Neil and Nick, where Sankoff (2004) had shown (by auditory means) mixed patterns of developing a FOOT/STRUT split due to mobility. However, for Neil, for both FOOT/STRUT and START/BATH categorisations, plotting the formants showed that from age 21 onwards, all vowels were categorised if classed using an SSBE based system (i.e. classifying all possible STRUT vowels STRUT). So, /a/ for Nick represents only vowels in the TRAP set and /ɑ:/ only those in the START set.

3.4.2.2 Vowel space area estimations

Monophthong data were also used to plot estimated habitual vowel space area for the speakers. Vowel space area estimations (VSAe) were calculated by using the same nine monophthongs as points which described a polygon, the area of which was calculated (using a formula written by Mark Adams, modified by Bernhard Fabricius (Fabricius, Watt, & Johnson, 2009)). This formula has not been used in forensic settings, (vowel space area has been previously been estimated using long term average spectra). However, it should present an insight into what Nolan (1983) terms the habitual speaker setting, i.e. those plastic limits of the vocal space within which a speaker normally behaves. The benefit of this formula over LTAS is that the same vowels are considered at each stage, giving an element of consistency in judging typical vowel realisations as part of a wider 'habitual' system.

There are problems with using this measure, as it only takes into account the mean value of each monophthong, and does not reflect speakers' natural variation in targets. A method which was able to measure not only the size of the vowel space, but also the variation in how extreme articulations were, would give a more accurate picture of speaker behaviour. This vowel space with variability-based 'blurring' at the edges would be preferable in future research into vowel spaces. Another problem is interpreting results from a geometric area measure, where small changes in one dimension can lead to very large changes overall. For the purposes of this study, however, the simple polygon formula at least gives a simple idea of the averaged extent of vowel realisations.

3.4.2.3 Vocal tract length estimations

Vocal tract length estimations (or VTLe) were made using Paige and Zue's (1970) equation, based on an extension of the quarter wave formula. In basic terms, the quarter

wave formula can estimate the length of a uniform tube (vocal tract) based on frequency outputs; this is only really useful for looking at central or schwa vowels, however (where the tract is likely to be more uniform). To make estimations for other vowels, Paige and Zue extended the formula (below) to account for deviation from uniformity, thus allowing for VTLe based on multiple formants. In the present study three formants were used for VTLe, which is reflected below:

$$VTLe = \left(\frac{c}{4}\right) \frac{\sum_{i=1}^{3} {F_i}/{(2i-1)}}{\sum_{i=1}^{3} \left[{F_i}/{(2i-1)}\right]^2}$$

(i=number of formant, i.e. 1,2,3;
c=speed of sound in air; F in Hz)

Source: (Paige & Zue, 1970)

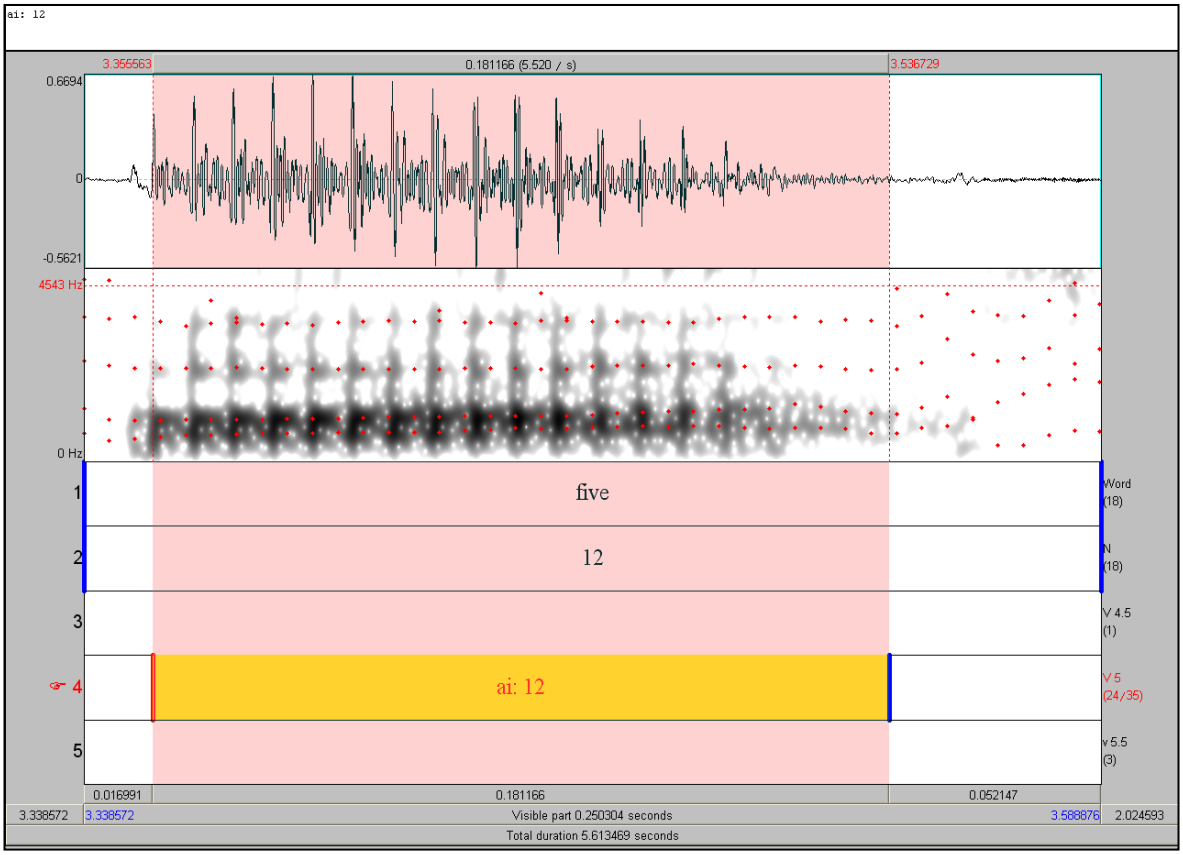
The use of this formula is illustrated by Clermont and Mokhtari (1998) and Clermont (2002; 2007). The performance of the formula (applied in the same manner as it is in the present study) is noted in Paige and Zue (1970) by comparing VTLe measures with physical measures of length using 'real' length and formant data from Fant (1960). The average error across six vowels was about 6%, although this was skewed by an 11% error for [i], whereas three of six other vowels showed under 4.5% error. The formula underestimated length in all these cases (thought this was a small test with one dataset).
For the current dataset, VTLe were based on all available monophthongs for each speaker, as comparisons were within-speaker. Using multiple vowels was preferred to estimates based on schwa as this gave a more comprehensive set of estimates of speakers' VTLe during longer periods of speech. All three measured formants were used based on the recommendation that "the greater number of formants used, the more realistic the length estimate is likely to be" (Clermont & Mokhtari, 1998, p. 528). Where overall VTLe measures are presented, this is simply the mean average of VTLe across all vowels.

3.4.2.4 Diphthongs

For those non-monophthong vowel frequencies that are included in the study, the analysis procedure was slightly different. The tokens were, like monophthongs, marked in a text grid in *Praat*; in this case the onset and offset were marked accurately in order for a script (developed by Phil Harrison for McDougall (2006)) to take measurements at

10% time intervals along the duration of the vowel, effectively normalising for duration differences between tokens. The onset and offset of the vowels were principally decided using periodicity in the waveform (Turk, Nakai, & Sugahara, 2006), and subsequent, more fine-grained, adjustments were made afterwards using spectrographic indicators. Vowel onsets were usually fairly straightforward; vowel offsets are generally more difficult to identify. However, the end of spectral energy in the second formant is usually a good indicator of the end of the vowel sound (McDougall, 2005) and this provided good consistency in demarcating the vowel sounds. Different formant tracker setting were used depending on the production of the vowel and the speaker, so different lines of the text grid were annotated for the appropriate 'number of formants' settings, as can be seen in the example, Figure 16, below (i.e. 4.5, 5 or 5.5):

Figure 16 - Example *Praat* text grid of a segmented diphthong /aɪ/



In most cases dynamic measures for the following diphthongs were extracted:

/aɪ, eɪ/

(or the PRICE and FACE sets (Wells, 1982)

Frequency of these features varied highly in the samples, i.e. for PRICE, where discourse markers such as 'like' were used by some speakers. 'I' also appeared a great deal due to the nature of the interviews. Other diphthongs had insufficient tokens to allow for a satisfactory analysis.

3.4.2.5 N tables

This section presents the number of tokens (Ns) for monophthongs and diphthongs in the present analysis, as well as the distribution of Ns across each monophthong vowel category.

Table 19 - Monophthongs Ns for each speaker

Subject	Age	u:	ʊ	ɒ	ɑ:	ʌ	a	e	ɪ	i:
Andrew	**21**	3	3	4	2	6	14	12	13	7
	28	3	2	2	6	2	4	9	5	5
	35	2	2	1	2	3	5	4	9	6
	42	5	0	2	2	2	5	6	12	6
	49	5	1	5	5	3	11	15	8	6
Bruce	**21**	10	4	18	6	21	24	20	19	20
	35	4	2	2	2	6	12	6	13	17
	42	4	2	2	3	3	4	10	9	9
	49	7	1	7	5	6	8	10	9	9
Lynn	**21**	4	4	8	3	7	6	6	12	5
	28	2	2	4	2	10	7	4	12	2
	35	2	1	4	2	6	8	10	9	1
	42	5	5	13	9	3	12	14	13	8
	49	3	3	15	6	11	9	13	9	5
Neil	**21**	2	3	20	4	7	20	26	24	14
	28	6	2	13	6	5	19	24	20	14
	35	2	2	10	6	4	10	16	15	5
	42	4	2	14	9	5	17	9	15	13
	49	8	2	9	11	15	17	14	14	15
Nick	**21**	2	2	3	6	4	4	9	13	7
	28	7	4	9	4	11	11	11	30	10
	35	4	4	4	5	6	4	7	13	5
	42	5	2	3	6	4	6	8	18	6
	49	2	8	3	12	8	3	11	10	10
Suzy	**21**	4	0	9	6	9	7	5	15	14
	28	7	2	9	8	3	8	14	21	7
	35	4	2	9	9	8	19	14	28	14
	42	4	3	11	5	10	9	10	30	12
	49	2	1	1	3	9	7	8	13	11
Symon	**21**	3	3	4	6	9	9	9	15	12
	28	4	4	8	7	9	12	8	21	11
	42	2	2	3	5	6	12	3	4	3
	49	8	6	9	3	4	10	12	11	12
Tony	**21**	3	8	8	8	11	14	16	14	10
	28	4	1	3	8	14	10	14	21	7
	35	4	2	7	5	5	11	13	18	6
	42	2	2	7	3	8	6	11	10	7
	49	2	0	7	7	14	18	16	14	3

Figure 17 - Distribution of monophthongs Ns across vowel categories

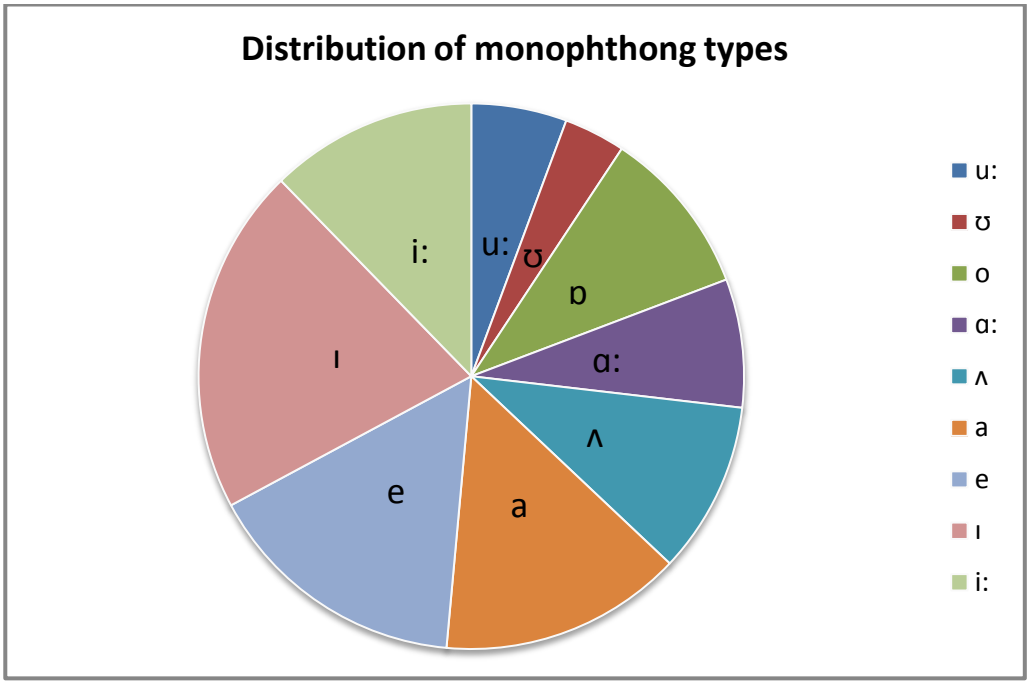


Table 20 - Diphthongs Ns for speakers included in diphthong analysis

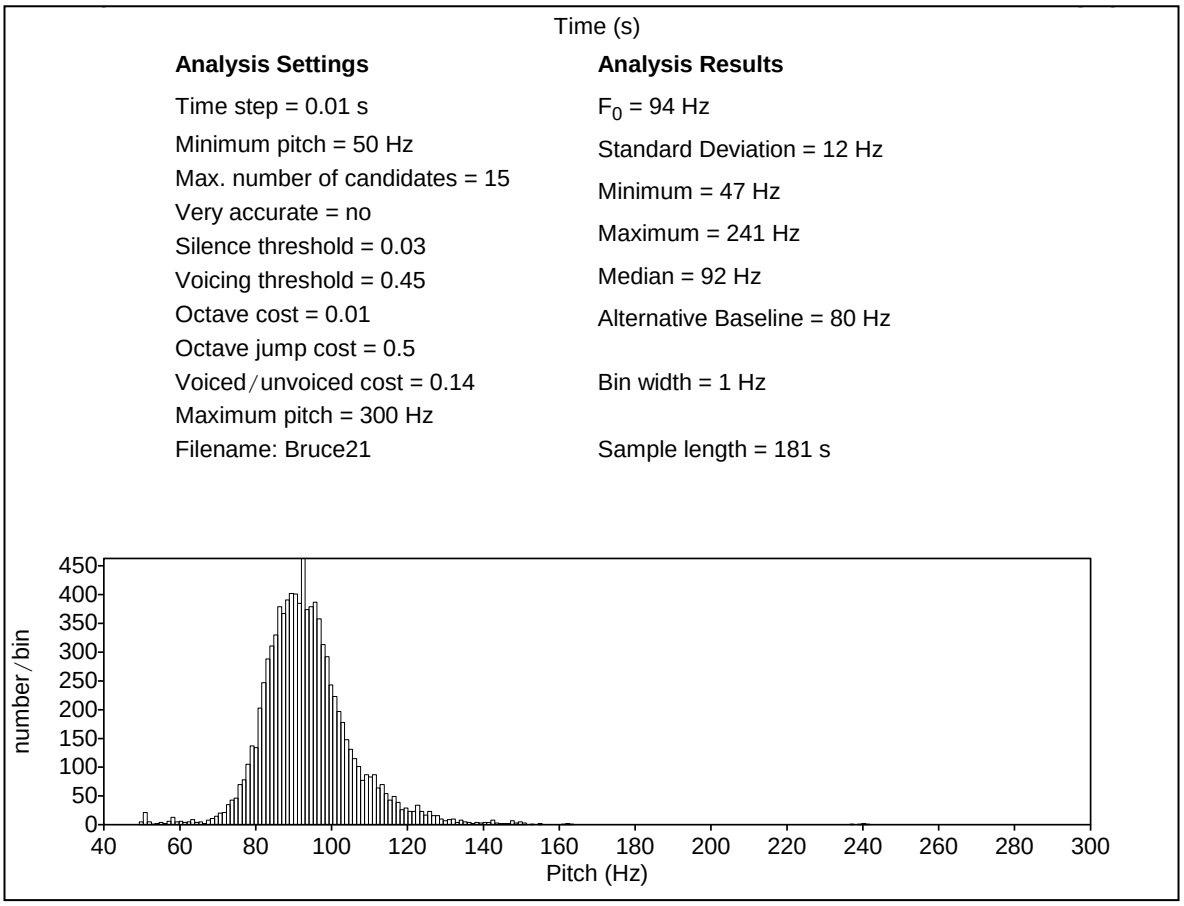
Subject	vowel	21	28	35	42	49
Bruce	aɪ	18	n/a	7	4	8
Neil	aɪ	19	17	16	18	32
	eɪ	22	9	8	5	13
Suzy	aɪ	11	10	16	25	14
	eɪ	3	9	9	9	6
Symon	aɪ	8	7	n/a	5	9
	eɪ	9	4	n/a	4	5
Tony	aɪ	7	15	4	0	12
	eɪ	18	13	5	7	6

3.4.2.6 Fundamental frequency

Fundamental frequency, or F0, was measured using another *Praat* script developed by Philip Harrison. In order for the process to work, all non-speech sounds were extracted from the recordings, and speech which was overlaid with any extraneous noise was also omitted. It is also very important that all speech was of a similar mode and style, so any speech which was not in the standard interview format for the programme, or was, for instance, interrupted by laughter or delivered in a highly emotional state, was removed, as it has been shown that factors such as these can have a significant effect on F0 (Braun, 1995; Jessen, Köster, & Gfroerer, 2005; Jessen, 2009). Having said that, it is very difficult to ignore emotion, especially given the content of the documentary, and although extreme examples were removed, there is no doubt significant intra-speaker variability

due to emotion and/or stress. Periods of heavy breathing, extended hesitation and background noises were all removed. The script then takes frame by frame measurements of F0 across the entire recording and delivers information such as mean average F0, SD, minimum and maximum and provides a histographic profile of the F0 variability; for an example see the display in Figure 18 below.

Figure 18 - Example of pitch analysis read-out in *Praat* (script developed by Phil Harrison)



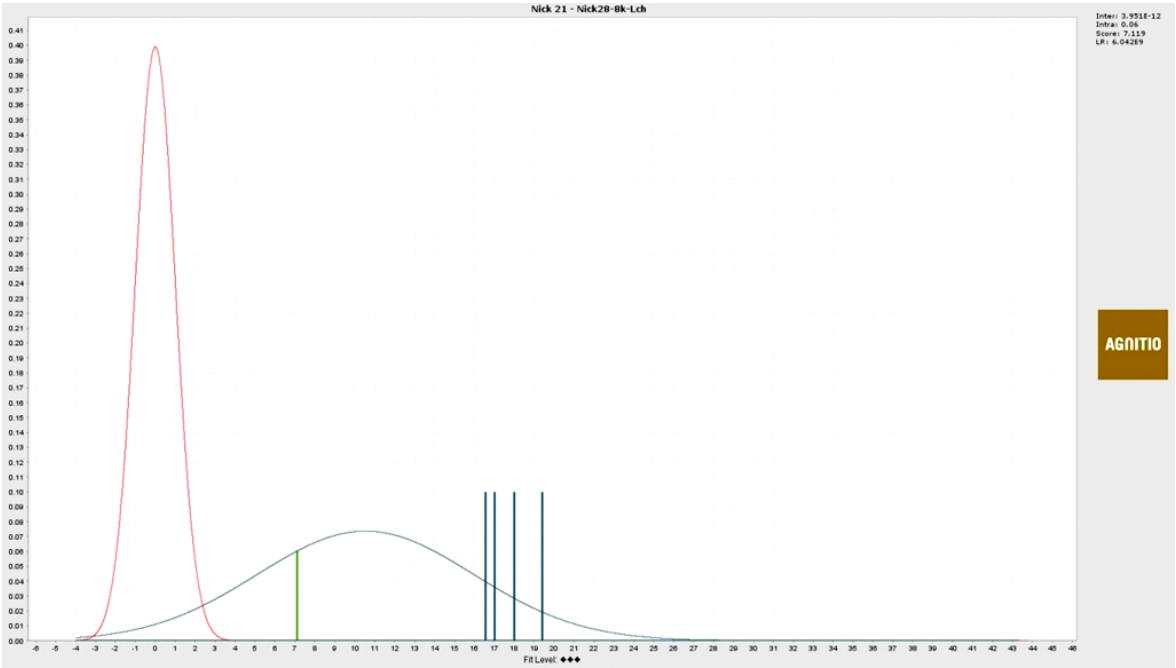
Although it captures an overall impression of F0, it is probably naive to assume that the simple mean and SD measures of F0 in this study represents a speakers' range. For example, Mennen et al. (2012) have demonstrated that an individual's 'average pitch' was best captured using similar level and span features, but measured between individual linguistic units, rather than across long term distributions.

3.4.2.7 MFCCs and ASR system

BATVOX (Basic v3.1) GMM-UBM system was used to extract and analyse MFCC data from the samples. *BATVOX* is automatic in so far as it extracts sections of speech from a recording, assesses their suitability (in terms of length and signal-to noise ratio (SNR)) and also selects a specified number of optimal reference speakers from a given reference

population. A 'model' is then trained from a sample of speech (this is analogous to the known suspect sample) to be compared with test data (evidential recordings). Both the suspect intra-variability and the inter-variability of the reference population are analysed against the evidential sample and used to calculate a likelihood ratio estimate for the strength of evidence. This can be seen in Figure 19 below, where the Gaussian distribution in red represents the reference population inter-variability, the flatter Gaussian in blue represents the model intra-variability, the blue vertical lines represent four measurements from the model and the green line shows the (test) evidential data. As the data in the test below show the evidential sample within the range of the model variability (similar) and removed from the reference population (atypical), the resulting LR is very high: 6.042E9.

Figure 19 - *BATVOX* output showing calculated MFCC distributions for reference and model data against evidential test recording (data for Nick, ages 21 v 28)



The audio files were resampled to 8 kHz and converted to single channel, which are requirements of the system. Only male speakers were analysed using the ASR system as sufficient female reference populations could not be found. The samples for each speaker had to be edited down to exclude any other speakers in overlap (i.e. in the foreground); although background speech was preserved when the subject was speaking in the foreground (*BATVOX* would not identify this 'noise' as the speaker). Background noise sections were not excluded as this would have prevented some samples from being used, and it was preferable to keep signal degradation as to be analogous with the

forensic condition (although the technical quality of the samples was much better than most forensic materials). Non-modal voicing sections were removed in extreme cases, particularly when laughter or other strong emotions affected the voice. *BATVOX* reports net speech durations as well as signal-to-noise ratios for the sections of speech it has selected; these are reported for the edited audio files used below:

Table 21 - Net speech duration (s) of edited audio for ASR, calculated by *BATVOX*

Net speech	21	28	35	42	49
Andrew	64.1	51.8	74.6	81.5	110
Bruce	321.5	-	167.1	156.1	196.5
Neil	388.1	359.9	287.2	250.1	313.3
Nick	144.1	283.7	234.6	269.9	202.5
Symon	194.7	180	-	128.2	162
Tony	259.5	248.9	208.5	148.1	272.9

Table 22 - SNR of edited audio for ASR, calculated by *BATVOX*

SNR	21	28	35	42	49
Andrew	28.8	29.9	31.1	26.1	38.8
Bruce	47.7	33.2	24	33.6	29.7
Neil	43.6	27.9	34.5	35.5	38.2
Nick	26.5	28.2	31.6	42.1	30.3
Symon	36.5	42.5	-	28.2	33.5
Tony	30.9	29.8	31.2	33.5	46.3

Data from 100 DyViS speakers were utilised as a reference population, with 35 optimal speakers being selected by *BATVOX* for each test (this is shown to be a suitable number in Bautista-Tapias (2005)). For more information about the reference population used to calculate the LR scores for the ASR test, and discussion on the potential effects of accent differences, see §3.5.3.2).

3.5 Statistical analyses

3.5.1 Monophthong formants

Descriptive statistics that are presented in the following sections were generated in *Microsoft Excel*, or automatically generated by *Praat* scripts. *Microsoft Excel* was also used to run exploratory unpaired t-tests on vowel formant data from the 21 and 49

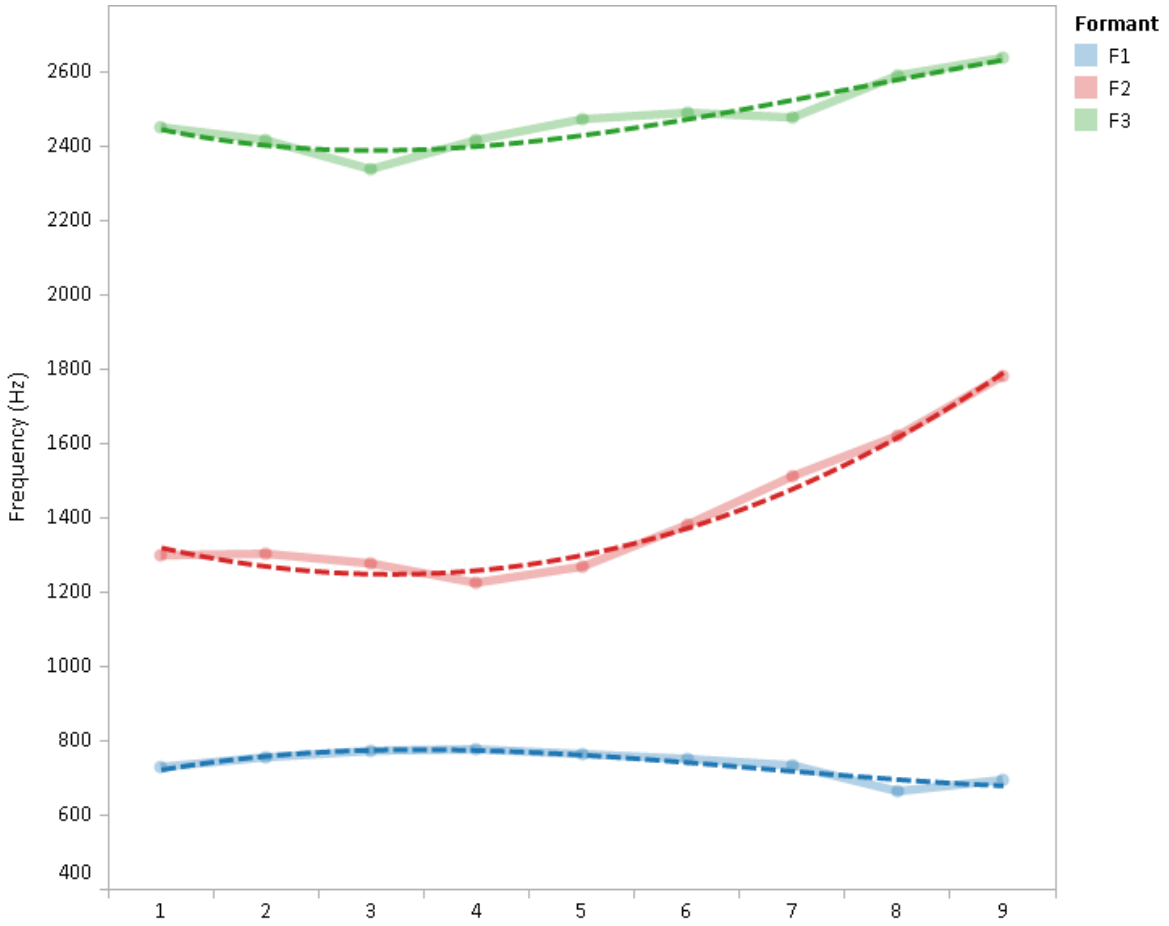
samples within each speaker to determine whether there had been a significant change. *SPSS* was used to test age as a significant factor across all samples in changing vowel formant values by univariate ANOVA. Pair-wise testing of all data from all five ages was then undertaken using post-hoc testing in *SPSS* (v.18) as part of the ANOVA testing.

3.5.2 **Diphthong dynamics**

Descriptive statistics that are presented in chapters 4 and 5 were generated in *Microsoft Excel.* *SPSS* was used to test age as a significant factor across all samples in changing vowel formant values by Multivariate ANOVA (MANOVA). Pair-wise testing of all data from all five ages was then undertaken using post-hoc testing in *SPSS* (v.18) as part of the ANOVA testing; Gabriel and Hochberg post-hoc tests were applied as these are best suited to unequal sample sizes (Field, 2005). There were problems with these tests and these are described in §5.

Diphthong data used in LR estimation in §6.3.2 were transformed to polynomial coefficients using *Matlab* (v.R2010b), using a script by Ashley Brereton. Formant transitions were subjected to cubic regression, as this represented the mostly sigmoidal transitions with the highest efficiency (see below). Regression allows coefficients of the curves to be used as vectors in LR estimation, rather than comparing single interval points, and with minimal predictors.

Figure 20 - Example of cubic regression curves fitted to formant frequency contours



3.5.3 **Likelihood ratio estimation**

Likelihood ratios for acoustic data were estimated in *Matlab* (v2010b/2011a) using Aitken and Lucy's (2004) Multivariate Kernel Density (MVKD) formula. This was implemented in a script by Dr. Geoffrey Stewart Morrison and run iteratively with a script designed by Philip Harrison and modified by Colleen Kavanagh and the author.

The LR estimations in this section are designed to follow the practical LR approach as it would be implemented in these case studies. Suspect and evidential (referred to as offender in LR literature) samples come from data presented in chapters 4 and 5. The analytical and statistical procedures are outlined in depth in §3.5.2-3.5.3. Vowels and formants were all calculated in separate vectors as the correlations and inter-dependencies between them are not well understood (although cf. Gold and Hughes (2012)), and this will affect the accuracy of testing.

3.5.3.1 Suspect and evidential data

The analysis follows what would happen in a case of long-term delay, such as the Yorkshire Ripper Hoaxer (R v John Samuel Humble, 2005), where the evidential sample comes from the youngest age and subsequent tests are made at each seven year interval to assess the effects of an increasingly large delay. Therefore for most speakers there are 5 comparisons, where the comparison between 21 evidential and 21 suspect forms the control LR with which the long-term non-contemporaneous LRs can be compared. As has been previously mentioned, the recordings all feature short-term non-contemporaneity, as recommended by Rose (2002) and Rose and Morrison (2009).

To create separate suspect and evidential samples at age 21, the sample was split with one third of the data as the suspect set and two thirds as the evidential set. While it would be preferable to split different recordings to make these two samples short-term non-contemporaneous, practically the amount of data from each time meant that this was not possible. The evidential set is larger as it is important to have enough data to compare with suspect sets at each stage. Bruce and Neil were selected for monophthong analysis as suitable reference populations could be found for SSBE, while they also had sufficient data at each stage for analysis. Analysis of six vowels is presented: /iː ɪ e a ʌ ɒ/. Diphthong data for four of the male speakers' (Bruce, Neil, Tony, Symon) was analysed as they had sufficient data and (apart from perhaps Tony) there is not a dramatic phonological difference between those speakers and the reference population, with respect to the PRICE vowel.

While this is rather male-orientated, it was not possible to locate a suitable female reference population for Suzy and Lynn. This was problematic both in terms of attaining numbers of speakers and matching a RP/SSBE and London/Cockney type accent.

3.5.3.2 Reference population data

The reference populations were selected as the best available match for the *21 Up* data (from 1978 at age 21) and for larger amounts of speakers. Given the restrictions on available data for reference populations, numbers of tokens for different speakers in the present data and the diverse accent types in the present subjects, this represents a compromise. However, given that the analyses are testing the effects of age, and that the reference sample remained stable throughout, analyses illustrate aging effects effectively.

What might be affected is strength of evidence for individual speakers where a mismatch is apparent, which might exaggerate strength of evidence. Moreover, although it would be ideal to record a bespoke reference corpus for each speaker, this would not be practicable in a real setting.

Bruce and Neil feature in the monophthong analysis, as they had the largest amounts of data across all vowels and both have an SSBE type accent (although Neil's is adopted). Data from Deterding (1997) were originally piloted as a reference population, with 12 speakers of RP/SSBE from a corpus of BBC recordings from the 1970s. There were around 12 tokens per speaker, though this varied between vowel categories. This is shown to be about the lower limit for estimating LRs without massive variability, according to Hughes (2012). Due to concerns around the amount of speakers and the method of measurement, data from 25 speakers (23 for FLEECE) from the DyViS corpus (Nolan, McDougall, de Jong, & Hudson, 2009) were preferred for the analyses shown in this thesis. This data was measured by Frantz Clermont and Nathan Atkinson (for FLEECE), using *Praat.* This number of speakers is shown to be more stable in Hughes (2012) and also afforded more direct comparisons between diphthong, monophthong and ASR results. Moreover, it was judged that although the Deterding data matched the speakers for decade, age and reference size would have a more significant effect on the results than the mismatch between recording dates (apparent for the DyViS material).

For diphthongs, PRICE data from the DyViS corpus (Nolan, McDougall, de Jong, & Hudson, 2009) are used to analyse PRICE tokens from Bruce, Neil, Symon and Tony. Although Tony and Symon have different accent types to SSBE, they were included to investigate the effects of the difference in this mismatch. Although in practice, all speakers are mismatched to the database as it was recorded nearly 30 years after the age 21 recordings from the *7 Up* database. The DyViS corpus was used as these were the only data available 'off the shelf' that was as close to the speakers and featured PRICE tokens.

Data from DyViS task 1 were used for both analyses, where subjects take part in a simulated police interview. While this is not a direct reproduction of the setting of the *7 Up* interviews, it is somewhat analogous. Recording conditions for the database are available in the referenced report. Needless to say the recordings were made in a high-quality laboratory setting and recorded using optimal equipment (audio CD quality, sampled to 44.1 kHz).

The DyViS reference database (task 1) was also used for the ASR analysis by *BATVOX*. Single channel recordings, resampled to 8 kHz and reduced to 4 minutes duration for each of the 100 speakers were selected as reference population data for each ASR test. While for diphthongs it is possible to assess what difference accent might have on productions and therefore formant transitions, the holistic MFCC vocal tract modelling used by *BATVOX* is not so easy to predict. The makers (Agnitio Corporation) claim on their online materials that the system functions regardless of language or accent differences between samples and with the reference population (Agnitio Corporation, 2012). This claim is based on the logic that the vocal tract measures are individual and biological, not language or accent dependant. However, there are presumably some long-term habitual configurations and voice quality 'settings' that are shared across accent communities (French & Foulkes, 2012). Furthermore, Harrison and French (2010) found that there was sensitivity in the selection of reference speakers according to regional background. For the current subjects, there are differences with the reference population and while this study is principally concerned with aging, it is probable that accent difference may have an effect. Using the same task from DyViS for both FD and ASR tests does allow for limited comparison between results. Nevertheless, the parameters are very different and vary in relation to the *Up* data in different ways.

The fact that reference population data were not numerous or easy to find, even for SSBE, indicates that French et al. (2010) are accurate when they assert that there is currently not enough UK population speech data to implement an overtly empirical LR approach.

4 Acoustic analysis of monophthongs

This chapter presents results of acoustic analyses across the 21-49 age period in answer to research questions 1 and 2 posed in §2.1.6. These phonetic and acoustic parameters are investigated as they are typically measured in forensic tasks, and this section sheds light on the likely effects of age.

4.1 Overall results

This section compares the predictions from previous research (§2.1) with findings across all speakers to illustrate general patterns which are consistent even within this small data set. Predictions are briefly summarised with respect to each parameter. Results for each formant are presented generally and with respect to different vowel categories. Speaker summaries are presented in section §4.4. Of course, general findings are coloured by individual differences due to accent changes and mobility, but these are addressed separately, speaker by speaker in that later section.

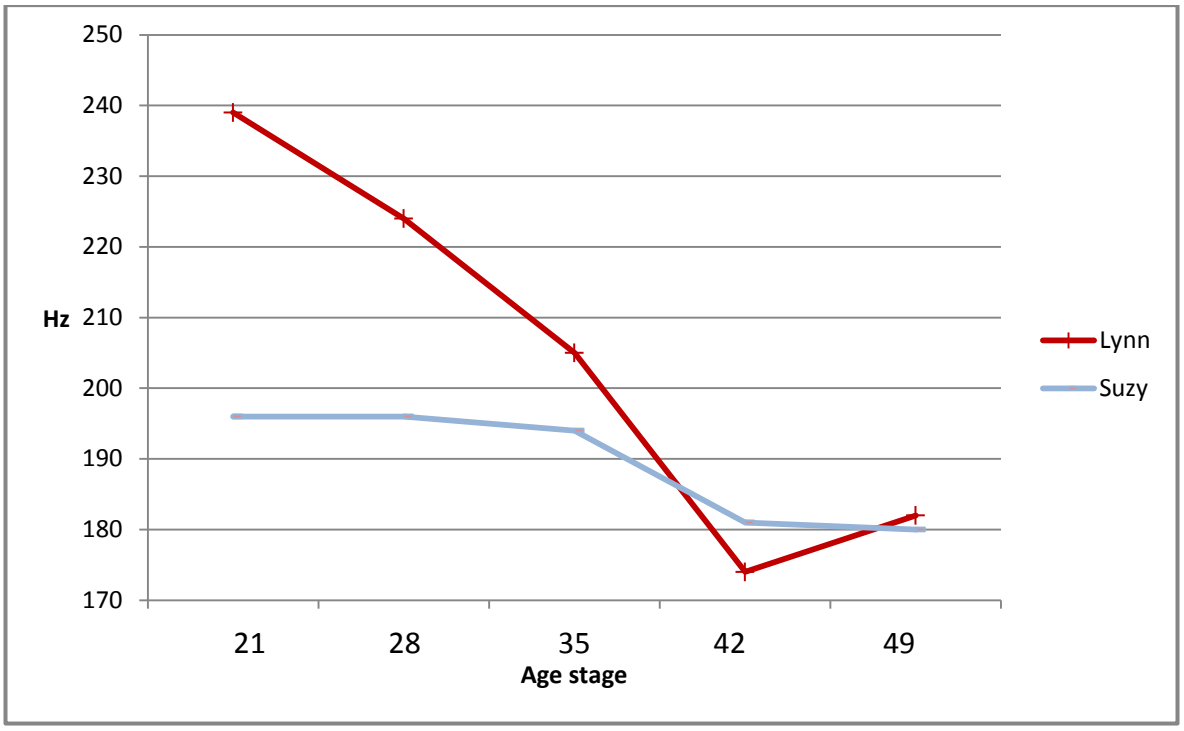
4.1.1 **Fundamental frequency (F0)**

4.1.1.1 F0 Predictions

For male speakers, F0 is predicted to decrease between 21 and 49 years by around 10-14% (Hollien & Shipp, 1972) or 14Hz (DeCoster & Debruyne, 2000). Female speakers are expected to experience a less marked effect, with only a slight decrease (Linville, 2001). For all speakers, there are potential effects of laryngeal health changes due to lifestyle factors, the most obvious being smoking. Smoking is expected to significantly lower F0 and also increase the rate of decrease in F0 across time (Linville, 2001). Although SD of F0 is stated to increase massively in elderly subjects (Kaltieider, Fray, & Hyde, 1938; Mittman, Edelman, Norris, & Shock, 1965; Pierce & Ebert, 1965; Linville & Fisher, 1985; Orlikoff, 1990), there are few findings for the ages of the subjects of the current study.

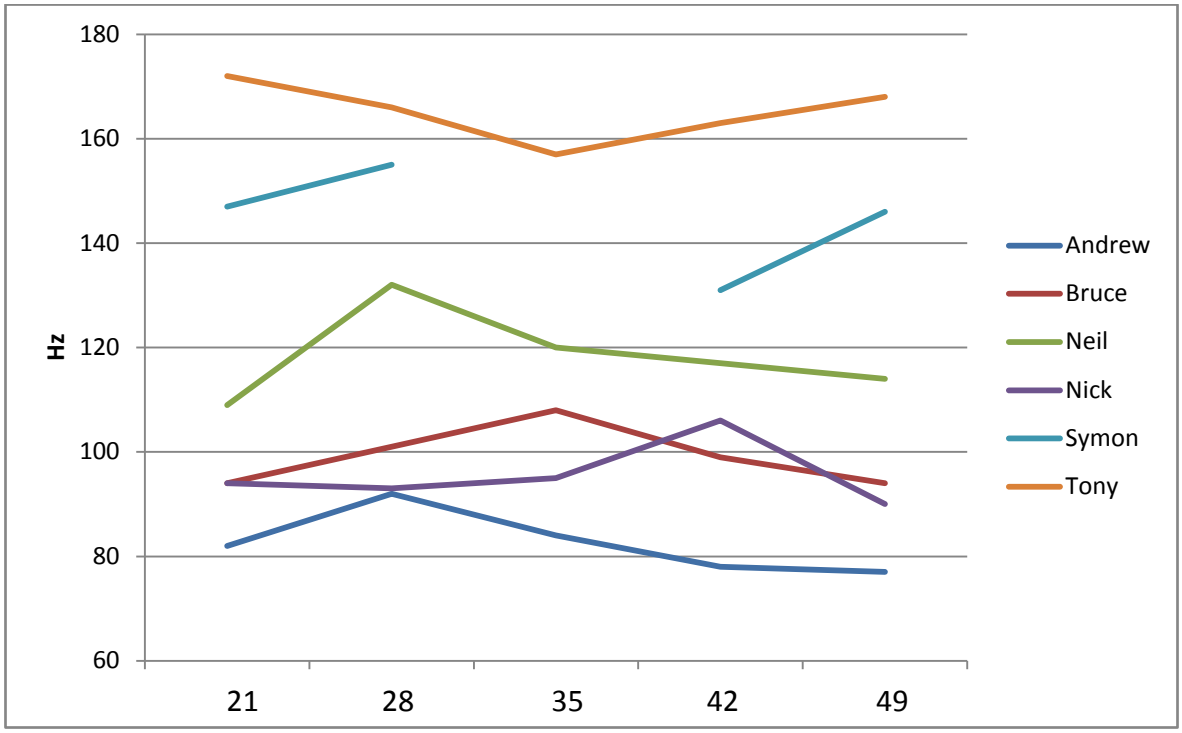
4.1.1.2 F0 Results

Figure 21 - Average F0 for female speakers across all five samples



It is clear from Figure 21 above that there is a clear and consistent reduction in F0 for the two female subjects. For Suzy, there is an 8% reduction between ages 21 and 49. For Lynn, the decrease is much more emphatic, at 23%. This reduction is probably dependent on the fact that Lynn is a habitual smoker (she is shown smoking across all series). It is also plausible that the changes found for female speakers, that are largely not apparent in male speakers (below), are exacerbated by the effects of menopause. These are shown to have a sharp impact on female F0 at a point between 30 and 50 in a similar study (Linville, 1996).

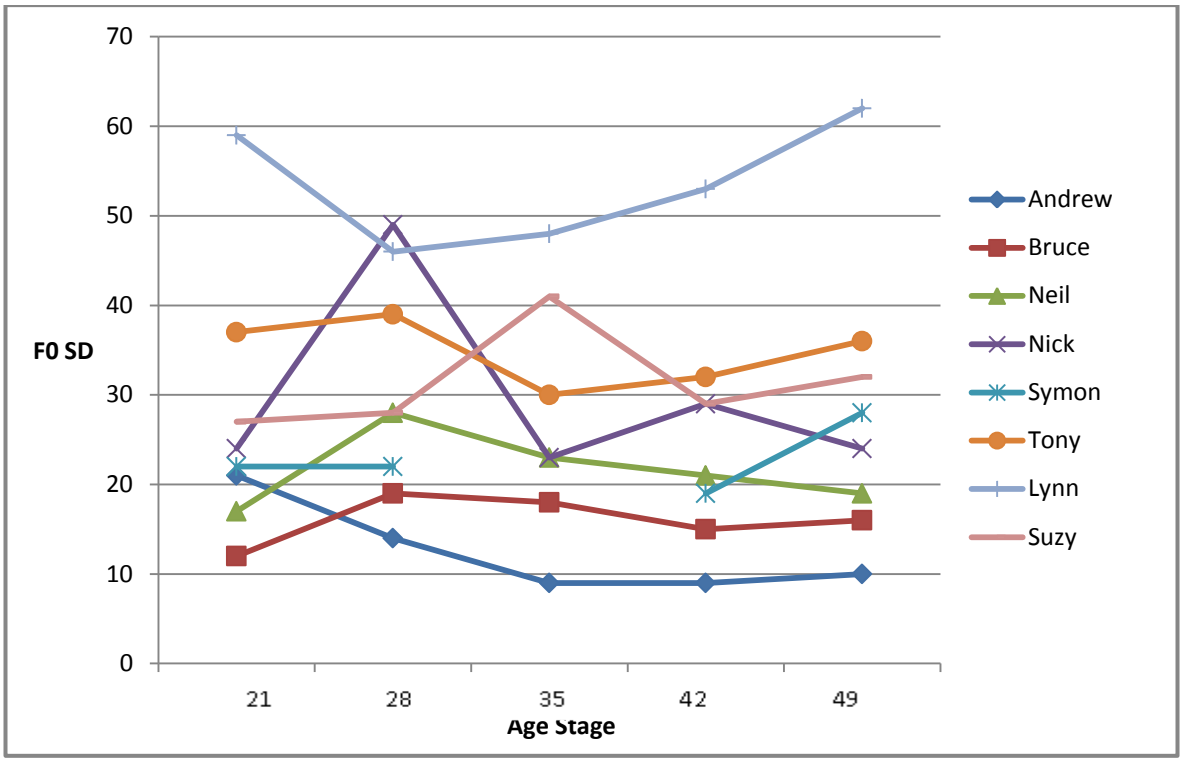
Figure 22 - Average F0 for male speakers across all five samples



For the male speakers, the pattern is less clear. Although there is an overall trend for a slight decrease between 21 and 49 years, this is not to the extent of previous research. Four of six speakers show a decline. However, this is only really noticeable in two speakers at 4% and 6% (Nick and Andrew). Bruce (0%), Tony (2.3%) and Symon (0.6%) show minimal reductions of F0 and Neil exhibits a 4.2% increase in F0 across the period. It can also be seen from Figure 21 that speakers do not display a steady decrease with age, instead displaying a mixture of increases and decreases at each stage. It would be sensible to suggest that most of this fluctuation is due to intra-speaker variability in F0, which could be caused by a number of contextual and individual factors, not limited to but including health, emotion, time of day, background noise (for more detail see §1.1.3 and §1.1.5). The results certainly do not reflect the same (mean between 10-14 Hz) reductions as the laboratory studies of Hollien and Shipp (1972) and DeCoster and Debruyne (2000), the conditions of which would presumably lead to more controlled speech styles and therefore more controlled F0 results.

4.1.1.3 F0 standard deviation

Figure 23 - F0 standard deviation across all subjects at each age stage



The standard deviation (SD) of F0 does not exhibit the same increases as reported in other studies with elderly subjects. There is a slight trend in some of the speakers for SD to increase after 35. However, this is neither consistent nor based on very tightly controlled data. The speakers do not vary hugely in their SD, and there is little crossover between speakers (for example, Andrew is largely stable at 10, while Bruce is quite stable at 16-18). Lynn has the highest SD and this could be related to smoking and having a higher F0 overall.

While these findings are mixed, they do stress the requirement of FSC exercises to take into account all available information to build a model of predictions, such as lifestyle factors including smoking habits. Given the variability in these results, they would seem to support the view widely held and reported that F0 is not an extremely useful parameter in FSC casework due to speaker-internal variability (Nolan, 1983; Jessen, 2009).

4.1.2 **First formant (F1)**

4.1.2.1 F1 Predictions

As discussed in §2.1.4.10, formants are predicted to decrease over this three-decade period, the main cause being expansion or extension of the vocal tract. The extent of this decline is not widely reported; however, F1 reductions are the most prevalent finding across formant-based studies.

4.1.2.2 F1 Results

Figure 24 - Mean percentage F1 decrease between 21 and 49 years for each speaker across all monophthongs

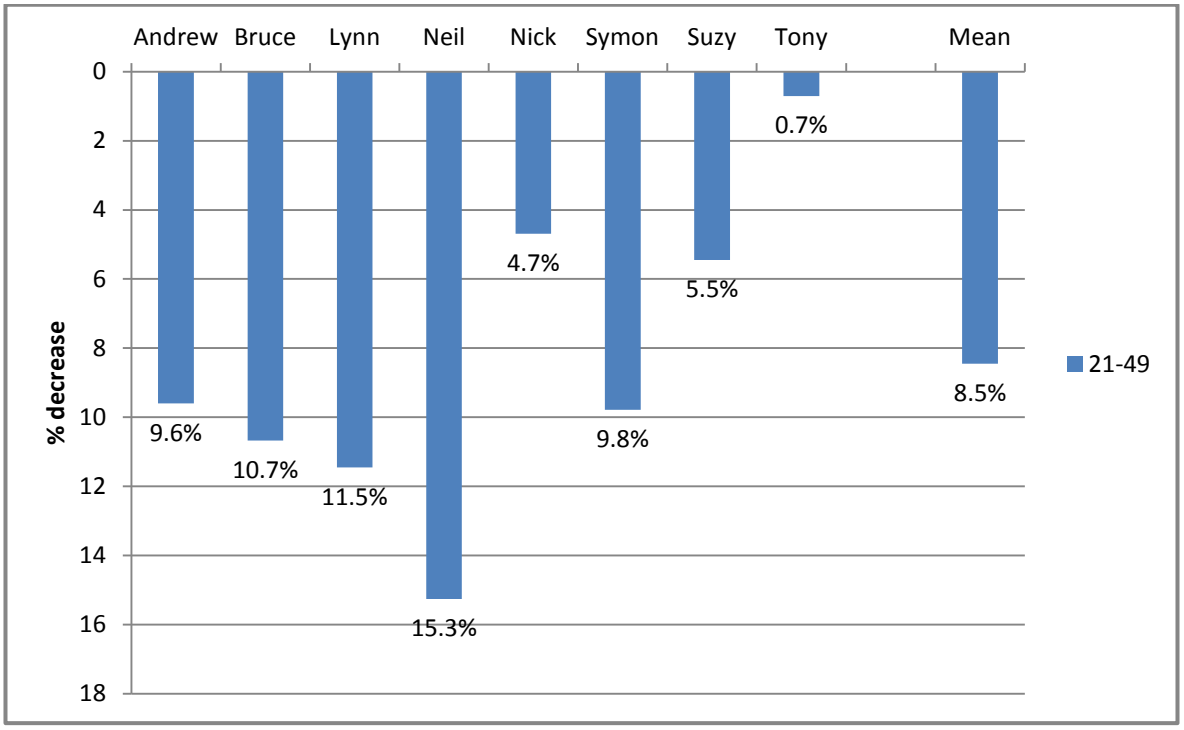


Figure 24 demonstrates a consistent reduction in F1 for all speakers between ages 21 and 49. The overall mean decrease is 8.5%, although some speakers show more of a reduction than others. The most extreme case is Neil, whose F1 is 15.3% lower across all monophthongs. Four speakers' F1 decreases around 10-11%, while two have a reduction of around 5%. Figure 25 below demonstrates that while there is not a decrease in F1 at every stage for every speaker, the reduction effect is fairly consistent. Two speakers (Suzy in orange and Tony in red) show a reduction followed by an increase. In Tony's case this leads to a marginal reduction and in Suzy's case, a reduction of 5.5% between 21 and 49 years.

Figure 25 - Average F1 across all monophthongs at each 7 year interval for each speaker

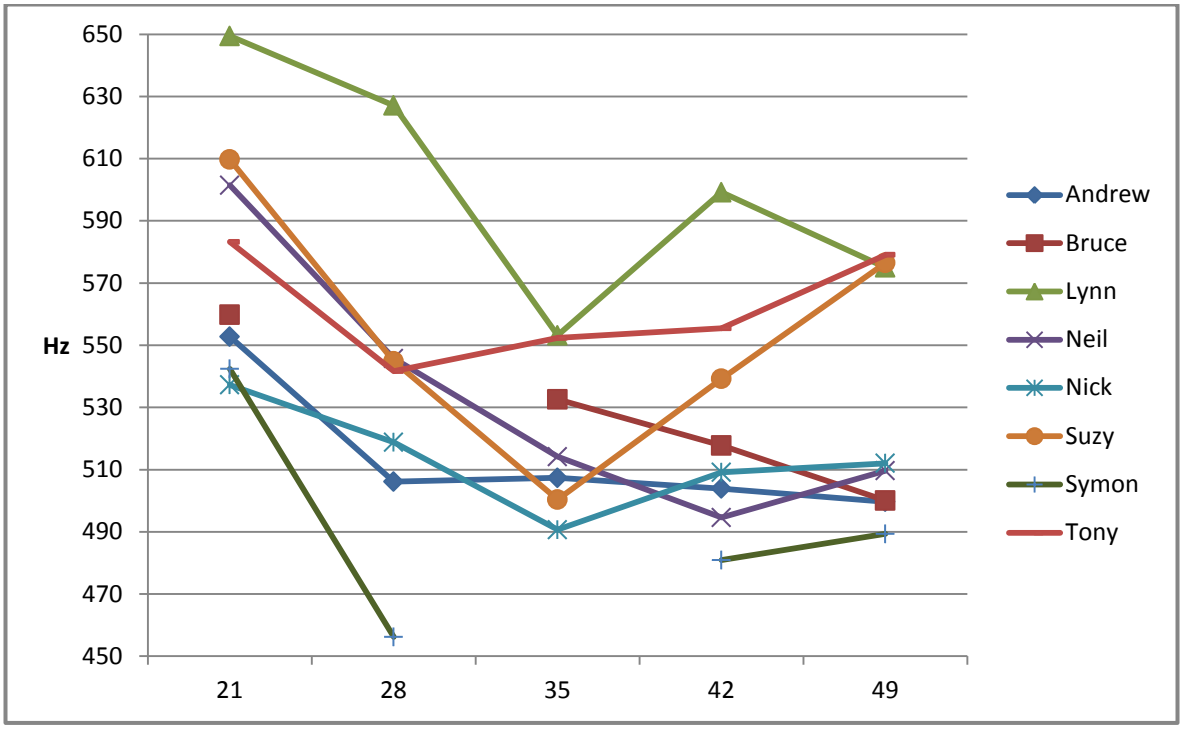


Table 23 shows changes in overall frequency for F1 between each age stage, demonstrating in further the detail the nature of changes. The example table below shows the data in the same format, where the colours represent the magnitude of change. The example table illustrates what would be expected for a completely steady or linear change, with blue in the top right area and red in the bottom left. So, for example, reading down from 21 shows the difference from 21 at each stage, in this example steadily increasing in magnitude.

	21	28	35	42	49
21		21-28	21-35	21-42	21-49
28	28-21		28-35	28-42	28-49
35	35-21	35-28		35-42	35-49
42	42-21	42-28	42-35		42-49
49	49-21	49-28	49-35	49-42	

Table 23 - Showing changes in average overall F1 frequency (Hz) between each age stage

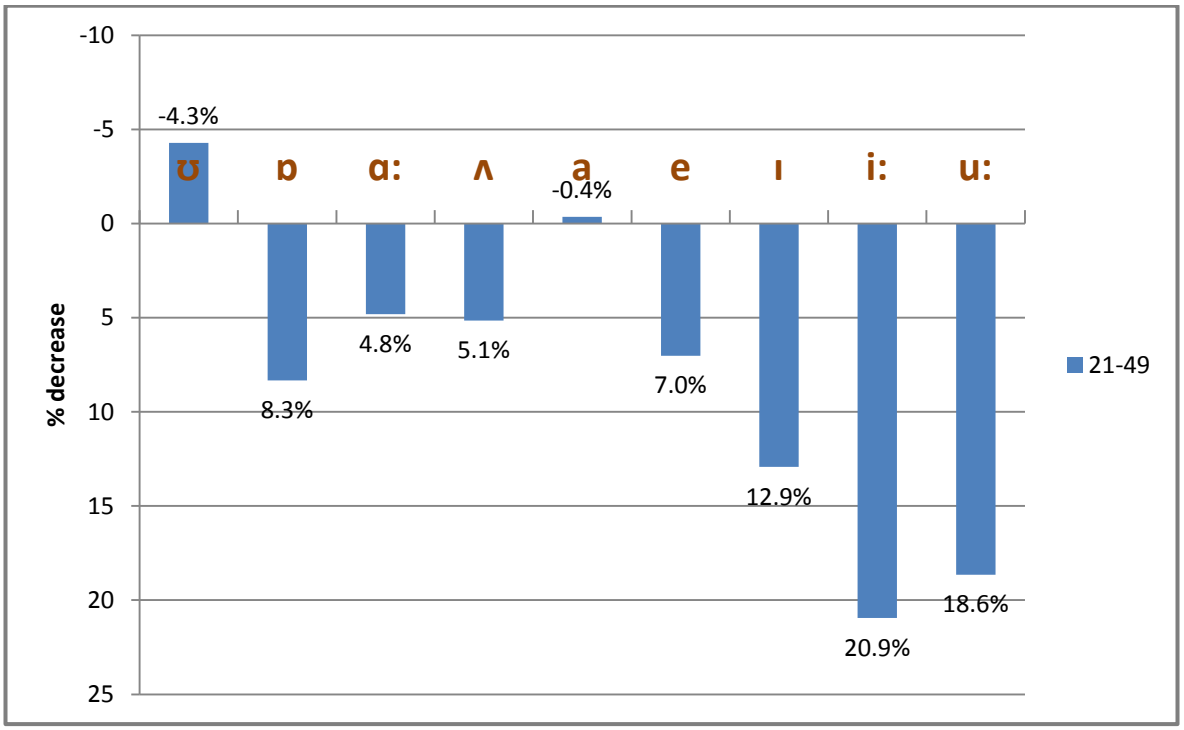
Andrew							Lynn					
	21	28	35	42	49			21	28	35	42	49
21		46.6	45.4	48.8	53.1		21.0		1.9	96.3	50.3	74.4
28	-46.6		-1.2	2.2	6.5		28.0	-1.9		94.4	48.4	72.5
35	-45.4	1.2		3.4	7.7		35.0	-96.3	-94.4		-46.0	-21.9
42	-48.8	-2.2	-3.4		4.3		42.0	-50.3	-48.4	46.0		24.1
49	-53.1	-6.5	-7.7	-4.3			49.0	-74.4	-72.5	21.9	-24.1	
Bruce							Suzy					
	21.0	28.0	35.0	42.0	49.0			21.0	28.0	35.0	42.0	49.0
21			27.2	42.1	59.8		21.0		64.9	109.4	70.5	33.2
28							28.0	-64.9		44.5	5.6	-31.7
35	-27.2			14.8	32.5		35.0	-109.4	-44.5		-38.9	-76.1
42	-42.1		-14.8		17.7		42.0	-70.5	-5.6	38.9		-37.3
49	-59.8		-32.5	-17.7			49.0	-33.2	31.7	76.1	37.3	
Neil							Symon					
	21.0	28.0	35.0	42.0	49.0			21.0	28.0	35.0	42.0	49.0
21		55.7	87.3	106.8	91.8		21.0		86.2		61.5	53.1
28	-55.7		31.6	51.1	36.1		28.0	-86.2			-24.7	-33.1
35	-87.3	-31.6		19.6	4.5		35.0					
42	-106.8	-51.1	-19.6		-15.1		42.0	-61.5	24.7			-8.4
49	-91.8	-36.1	-4.5	15.1			49.0	-53.1	33.1		8.4	
Nick							Tony					
	21.0	28.0	35.0	42.0	49.0			21.0	28.0	35.0	42.0	49.0
21		18.5	46.6	28.1	25.2		21.0		41.5	30.9	27.8	4.1
28	-18.5		28.1	9.7	6.7		28.0	-41.5		-10.7	-13.8	-37.4
35	-46.6	-28.1		-18.5	-21.4		35.0	-30.9	10.7		-3.1	-26.7
42	-28.1	-9.7	18.5		-2.9		42.0	-27.8	13.8	3.1		-23.7
49	-25.2	-6.7	21.4	2.9			49.0	-4.1	37.4	26.7	23.7	

Key (Hz)	
-10.00	10.00
-25.00	25.00
-50.00	50.00

Although the changes are fairly consistent, the table above shows that aging is not simply a monotonic process. For example, it is clear that F1 changes for Andrew are concentrated across the initial seven year period, whereas for Bruce the effects of aging are steadily realised. It also shows that Tony's data do not follow the general pattern as consistently, with blue quadrants in the lower left area and *vice versa*.

4.1.2.3 F1 by vowel

Figure 26 - Mean percentage F1 decrease between 21 and 49 years for each monophthong across all speakers



An interesting pattern is also observable if the average values for each vowel are compared across all speakers. It is clear in Figure 26 that F1 for close front vowels is much more affected by age than other vowels (if we treat /u:/ as a close front vowel, which it is for these speakers, and increasingly for speakers of RP/SSBE in general (Hawkins & Midgley, 2005)). This effect is confirmed if we examine those vowels which were determined as significantly affected by age as a fixed factor in a univariate ANOVA (see

Table 24). In this table, level of significance is indicated (by stars), along with the direction of change (by colour); direction of change was determined using mean values for each vowel at each age stage.

Table 24 - Showing F1 significance results for each vowel by each speaker by Univariate ANOVA (with key)

F1	Andrew	Bruce	Lynn	Neil	Nick	Suzy	Symon	Tony
ʊ					n		*	
ɒ		*	**	***	n	*	***	n
ɑ:		*		*	**	***	*	*
ʌ		*	n	**	n	n	n	n
a	n	n	n	n	n	**	n	n
e	*	***	*	***	n	n	*	*
ɪ	*	*	***	***	***	***	***	*
i:	***	***		***	n	***	***	n
u:		*	**	**	n	n	**	

*a mixed pattern indicates both increases and decreases in F1 across the period

Key	p ≤ 0.001	p ≤ 0.01	p ≤ 0.05
Decrease	***	**	*
Increase	***	**	*
Mixed*	***	**	*
Not sig.	n	n	n
Low Ns			

The table demonstrates the extent to which close front vowels in particular, but also most vowels, exhibit significant reductions in F1. It also demonstrates how some speakers show more significant reductions than others (although it is important to remember that although a number of tests did not yield significant results, there were still trends of non-significant reductions in frequency). For those tests which showed significant increases, there may be sociophonetic reasons for not following the general pattern; for Suzy in particular, there is further discussion in §4.4.6.

In general, the pattern for F1 is fairly consistent decreases across all speakers, and also a trend for close front vowels to exhibit a more marked reduction.

4.1.3 **Second formant (F2)**

4.1.3.1 F2 Predictions

As with F1, second formant frequencies are predicted to decrease over the period due to expansion of the vocal tract (discussed in §2.1). According to one study (Linville & Rens, 2001) F2 is more likely to change in women than men, perhaps dues to differing levels of

vocal tract lengthening (due to sex-related difference in vertebral deterioration) or male compensation strategies. These processes are not widely reported as being sex-stratified, however.

4.1.3.2 F2 results

Figure 27 - Mean percentage F2 decrease between 21 and 49 years for each speaker across all monophthongs

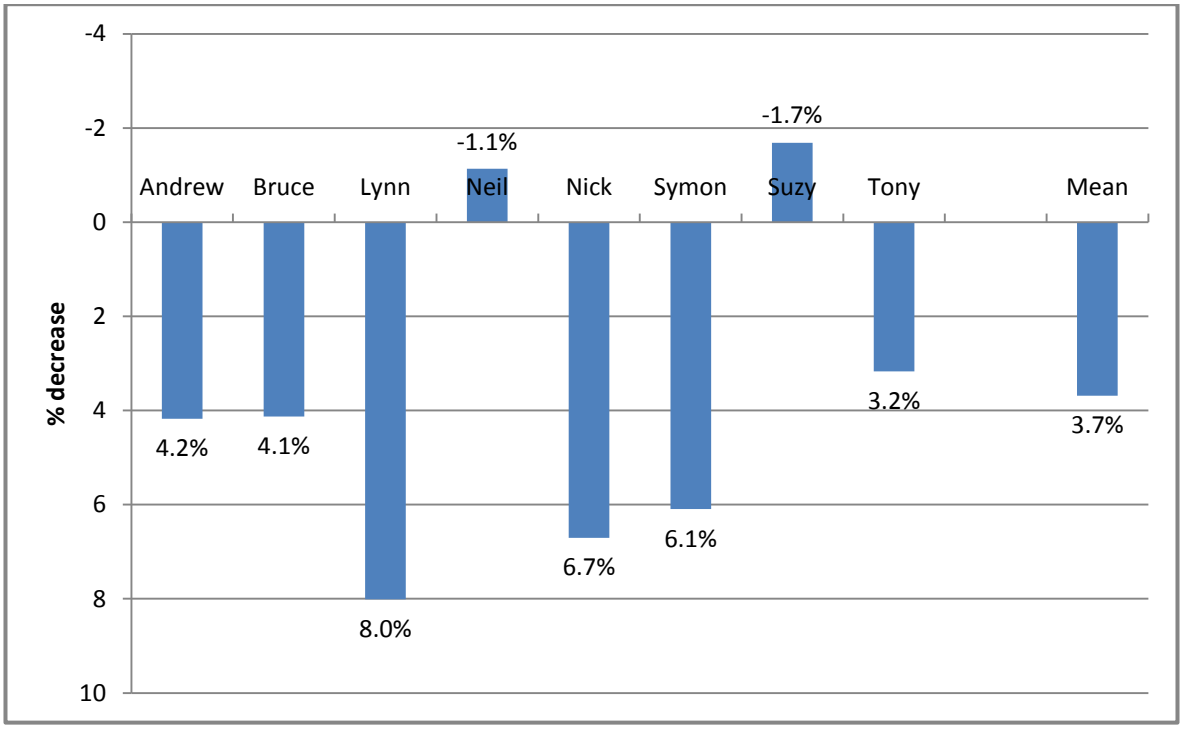


Figure 27 shows that while the pattern for F2 changes is slightly less consistent than for F1, there are still widespread frequency reductions. It also demonstrates that in general, these reductions are less extreme than for F1, with an overall average of 3.7% (or 5.4% across those speakers who show a reduction). For those who show decreased F2, the effect of the 28 year gap is somewhere between 4-8%. Two speakers, however, show a minor increase in F2. In Neil's case, however, average F2 at 21 is fairly low (1561 Hz) and his mean F2 exhibits a decline between 28 and 49 years (1630-1579Hz), thus following the more general trend.

In terms of between sex differences, Lynn does have the most extreme reduction. However, the other female subject, Suzy, does not follow the same pattern (possible sociophonetic explanations are discussed in §4.4.6). Figure 28 illustrates that for all speakers, except Suzy, F2 is following a fairly consistent reduction pattern across the period.

Figure 28 - Average F2 across all monophthongs at each 7 year interval for each speaker

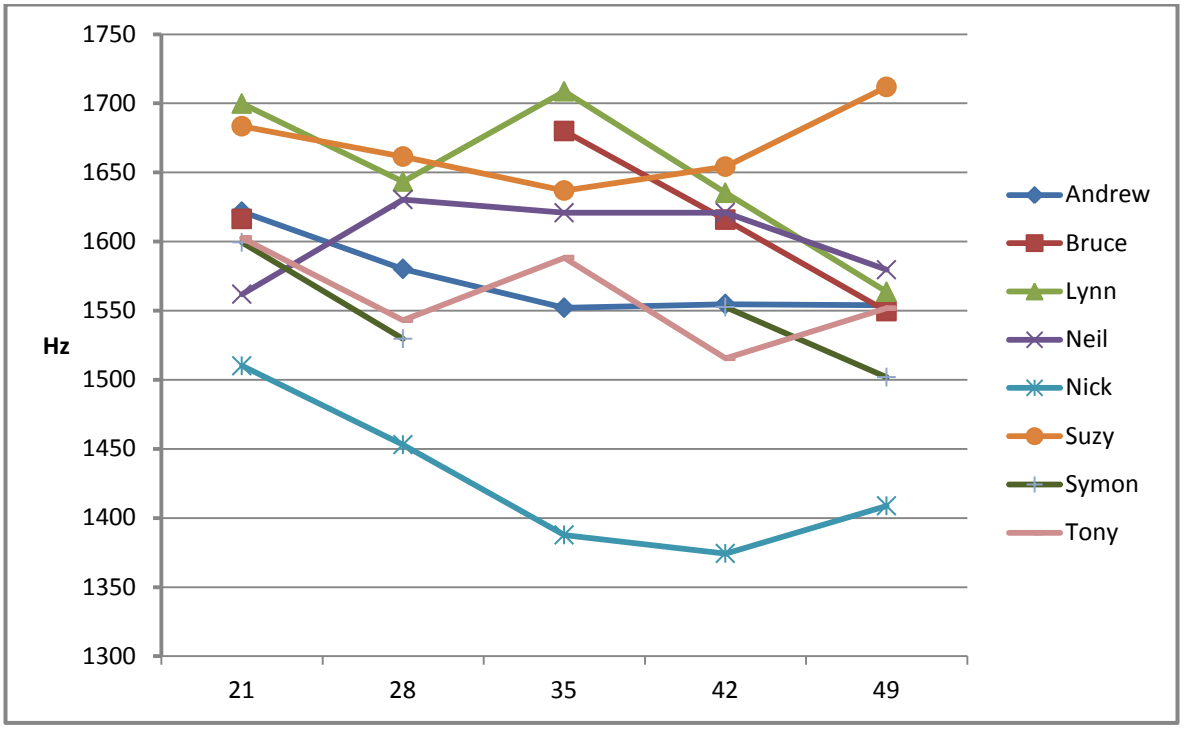


Table 25 demonstrates more individual changes and shows that, as for F1, Andrew presents larger changes in the first period. It also shows fairly steady deductions for most speakers. Generally, there are more results contrary to the overall pattern than for F1 (demonstrated by blue quadrants in the lower left area and *vice versa*).

Table 25 - Showing changes in average overall F2 frequency (Hz) between each age stage

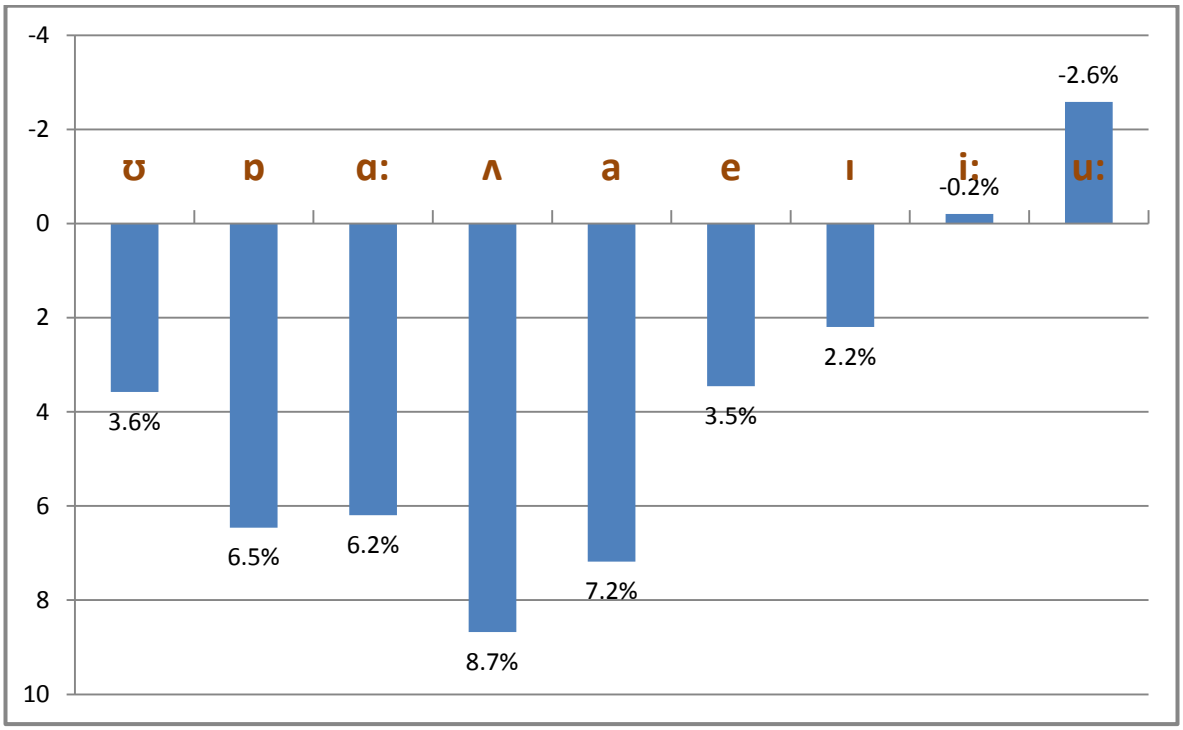
Andrew							Lynn					
	21.0	28.0	35.0	42.0	49.0			21.0	28.0	35.0	42.0	49.0
21		41.4	69.5	66.9	67.7		21.0		56.4	-8.9	64.5	136.2
28	-41.4		28.1	25.4	26.3		28.0	-56.4		-65.3	8.1	79.8
35	-69.5	-28.1		-2.7	-1.8		35.0	8.9	65.3		73.4	145.1
42	-66.9	-25.4	2.7		0.8		42.0	-64.5	-8.1	-73.4		71.6
49	-67.7	-26.3	1.8	-0.8			49.0	-136.2	-79.8	-145.1	-71.6	
Bruce							Suzy					
	21.0	28.0	35.0	42.0	49.0			21.0	28.0	35.0	42.0	49.0
21			-63.6	0.5	66.7		21.0		21.9	46.5	29.2	-28.4
28							28.0	-21.9		24.6	7.3	-50.4
35	63.6			64.1	130.3		35.0	-46.5	-24.6		-17.3	-74.9
42	-0.5		-64.1		66.2		42.0	-29.2	-7.3	17.3		-57.7
49	-66.7		-130.3	-66.2			49.0	28.4	50.4	74.9	57.7	
Neil							Symon					
	21.0	28.0	35.0	42.0	49.0			21.0	28.0	35.0	42.0	49.0
21		-68.4	-58.9	-59.1	-17.7		21.0		69.7		46.8	97.5
28	68.4		9.5	9.3	50.7		28.0	-69.7			-22.9	27.8
35	58.9	-9.5		-0.2	41.2		35.0					
42	59.1	-9.3	0.2		41.4		42.0	-46.8	22.9			50.7
49	17.7	-50.7	-41.2	-41.4			49.0	-97.5	-27.8		-50.7	
Nick							Tony					
	21.0	28.0	35.0	42.0	49.0			21.0	28.0	35.0	42.0	49.0
21		57.1	122.3	135.8	101.3		21.0		59.5	14.2	87.0	50.8
28	-57.1		65.3	78.7	44.2		28.0	-59.5		-45.2	27.6	-8.7
35	-122.3	-65.3		13.4	-21.0		35.0	-14.2	45.2		72.8	36.6
42	-135.8	-78.7	-13.4		-34.5		42.0	-87.0	-27.6	-72.8		-36.2
49	-101.3	-44.2	21.0	34.5			49.0	-50.8	8.7	-36.6	36.2	

Key (Hz)

-10.00	10.00
-25.00	25.00
-50.00	50.00
-100.00	100.00

4.1.3.3 F2 by vowel

Figure 29 - Mean percentage decrease between 21 and 49 years for each monophthong across all speakers



As with F1, there are observable differences between vowel types in the extent of F2 reduction. The first finding of note is the increase in F2 of /u:/; this would be expected in line with the trend in standard forms of English for /u:/ to become fronted, a sound change in standard English varieties which occurred concurrently with this series (Hawkins & Midgley, 2005). While not all speakers use a standard variety of English, it seems this pattern has shifted into other accents, as most speakers display a very front vowel in words in this GOOSE category throughout the series. The second noticeable pattern is for open vowels to show a slightly more marked reduction in F2 than close vowels. The vowels /ɒ, ʌ, a & ɑ:/ all show a reduction between 6-9%, higher than the other vowels in this analysis. Table 26 confirms that there are fewer significant changes in F2 than for F1, and that the STRUT, TRAP and KIT sets produce the highest number of significant changes in this formant. This is surprising given the relatively small percentage change in F2 frequency of /ɪ/ overall, but two speakers show non-significant *increases* in F2 of KIT (again following changes in RP/SSBE (Hawkins & Midgley, 2005)) which would affect this interpretation.

Table 26 - Showing F2 significance results for each vowel by each speaker by Univariate ANOVA (key in §4.1.2.3)

F2	Andrew	Bruce	Lynn	Neil	Nick	Suzy	Symon	Tony
ʊ					n		n	
ɒ		n	*	n	n	n	n	n
ɑ:		n		n	*	*	***	n
ʌ		*	*	*	***	n	n	***
a	n	*	***	*	***	n	*	*
e	n	n	n	*	***	n	*	n
ɪ	**	**	n	***	n	n	***	**
i:	n	*		n	*	*	n	n
u:		n	n	n	n	n	n	

For F2 in general there is an overall pattern of reduction, much like for F1, but in a less extreme manner and in fewer significant cases. Also, open vowels seem to change (overall) more than other vowels, conversely to F1 changes which occur predominantly and most strongly in close front vowels.

4.1.4 **Third formant (F3)**

4.1.4.1 Predictions

As with the first two formants, F3 is also reported as decreasing across this three decade period. Similarly to F2, Linville and Rens (2001) report that F3 is also more likely to change in women (the reason for this in F3 specifically is not suggested).

4.1.4.2 F3 results

Figure 30 - Mean percentage F3 decrease between 21 and 49 years for each speaker across all monophthongs

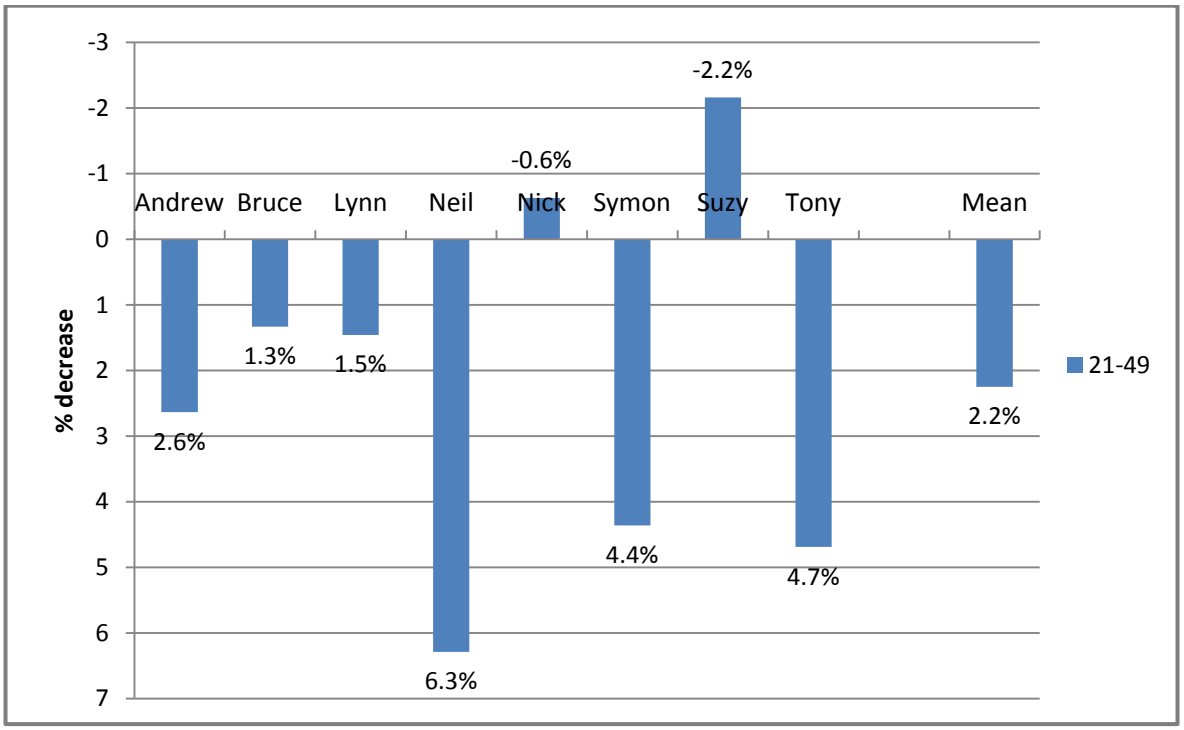


Figure 30 illustrates F3 changes across all speakers. It is apparent that changes are less consistent in F3 than for other formants. Nevertheless, there is still a reduction in six of eight speakers. Although the overall mean decline is only 2.2%, less than both F1 and F2, actually the magnitude of change of those speakers who exhibit a reduction is only slightly smaller than for F2 (3.4% compared with 5.4% for F2). As with F2, two of the speakers display an increase between 21 and 49 years, again Suzy is one of these speakers, contrary to the prediction made that females exhibit greater declines in F3 than males. Lynn also shows only a minor decrease in F3, contrary to this prediction. The most extreme reduction is a 6.3% shift in Neil's overall F3, with two other speakers at around 4.5% and three speakers between 1-3%.

Figure 31 also shows a less consistent pattern for speakers' F3, as well as the limited reductions or increases for female speakers in light green and orange. For the male speakers, Figure 31 demonstrates steady decreases in all cases except for Nick (who shows a minor increase overall, 0.6%).

Figure 31 - Average F3 across all monophthongs at each 7 year interval for each speaker

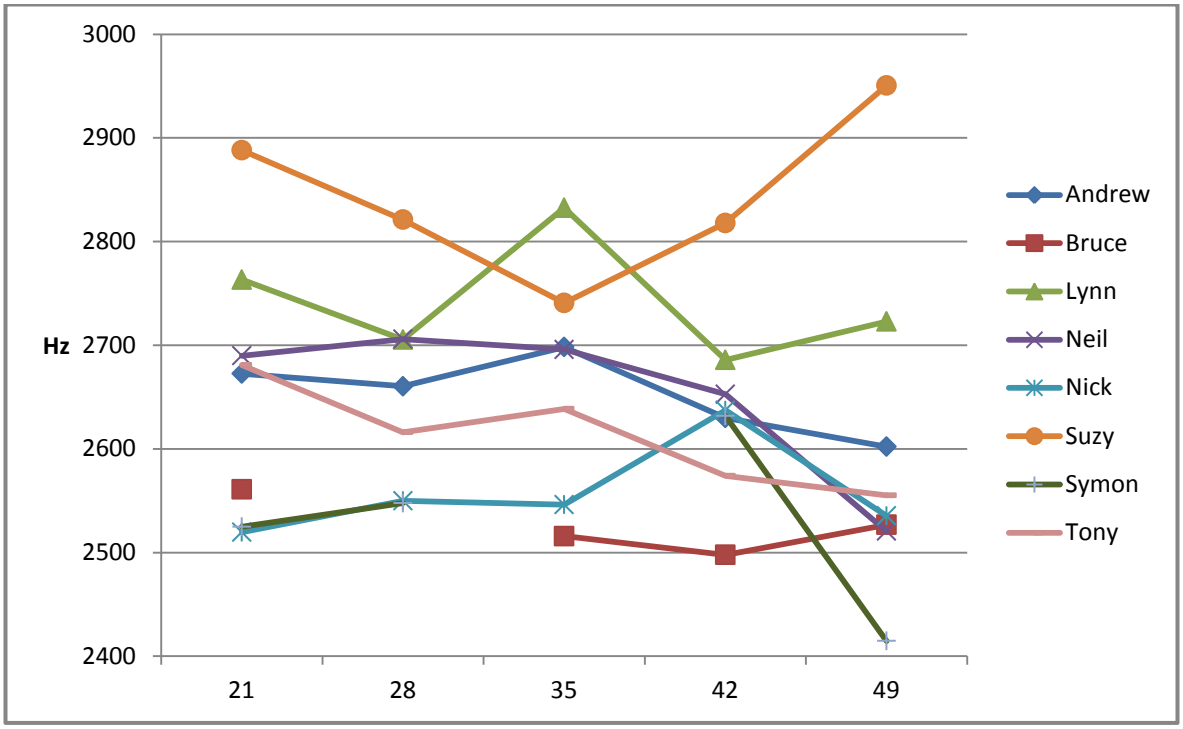


Table 27 below illustrates the changes in F3. Data for Andrew, Bruce and Tony demonstrate fairly consistent and quite steady reductions in formant frequencies. The patterns for Lynn, Suzy and Symon are much more mixed, with many more unexpected changes, corroborating the observations made about the figure above. For the remaining two speakers, Neil and Nick, the data clearly shows a drastic change in a single stage, at 49 and 42 respectively.

Table 27 - Showing changes in average overall F3 frequency (Hz) between each age stage

Andrew							Lynn					
	21.0	28.0	35.0	42.0	49.0			21.0	28.0	35.0	42.0	49.0
21		47.8	-18.6	67.3	94.1		21.0		57.8	-69.3	77.4	40.3
28	-47.8		-66.4	19.5	46.4		28.0	-57.8		-127.1	19.6	-17.5
35	18.6	66.4		85.8	112.7		35.0	69.3	127.1		146.7	109.7
42	-67.3	-19.5	-85.8		26.9		42.0	-77.4	-19.6	-146.7		-37.1
49	-94.1	-46.4	-112.7	-26.9			49.0	-40.3	17.5	-109.7	37.1	
Bruce							Suzy					
	21.0	28.0	35.0	42.0	49.0			21.0	28.0	35.0	42.0	49.0
21			45.2	63.1	76.7		21.0		66.9	147.3	70.3	-62.4
28							28.0	-66.9		80.4	3.3	-129.4
35	-45.2			17.9	31.5		35.0	-147.3	-80.4		-77.1	-209.8
42	-63.1		-17.9		13.6		42.0	-70.3	-3.3	77.1		-132.7
49	-76.7		-31.5	-13.6			49.0	62.4	129.4	209.8	132.7	
Neil							Symon					
	21.0	28.0	35.0	42.0	49.0			21.0	28.0	35.0	42.0	49.0
21		-7.9	6.7	51.4	170.5		21.0		-15.9		-95.5	116.8
28	7.9		14.6	59.4	178.4		28.0	15.9			-79.5	132.8
35	-6.7	-14.6		44.8	163.8		35.0					
42	-51.4	-59.4	-44.8		119.1		42.0	95.5	79.5			212.3
49	-170.5	-178.4	-163.8	-119.1			49.0	-116.8	-132.8		-212.3	
Nick							Tony					
	21.0	28.0	35.0	42.0	49.0			21.0	28.0	35.0	42.0	49.0
21		-30.3	-26.6	-117.5	-15.8		21.0		64.8	42.3	106.9	125.8
28	30.3		3.7	-87.2	14.5		28.0	-64.8		-22.5	42.1	60.9
35	26.6	-3.7		-90.9	10.8		35.0	-42.3	22.5		64.6	83.5
42	117.5	87.2	90.9		101.7		42.0	-106.9	-42.1	-64.6		18.8
49	15.8	-14.5	-10.8	-101.7			49.0	-125.8	-60.9	-83.5	-18.8	

Key (Hz)	
-10.00	10.00
-25.00	25.00
-50.00	50.00
-100.00	100.00

4.1.4.3 F3 by vowel

Figure 32 - Mean percentage F3 decrease between 21 and 49 years for each monophthong across all speakers

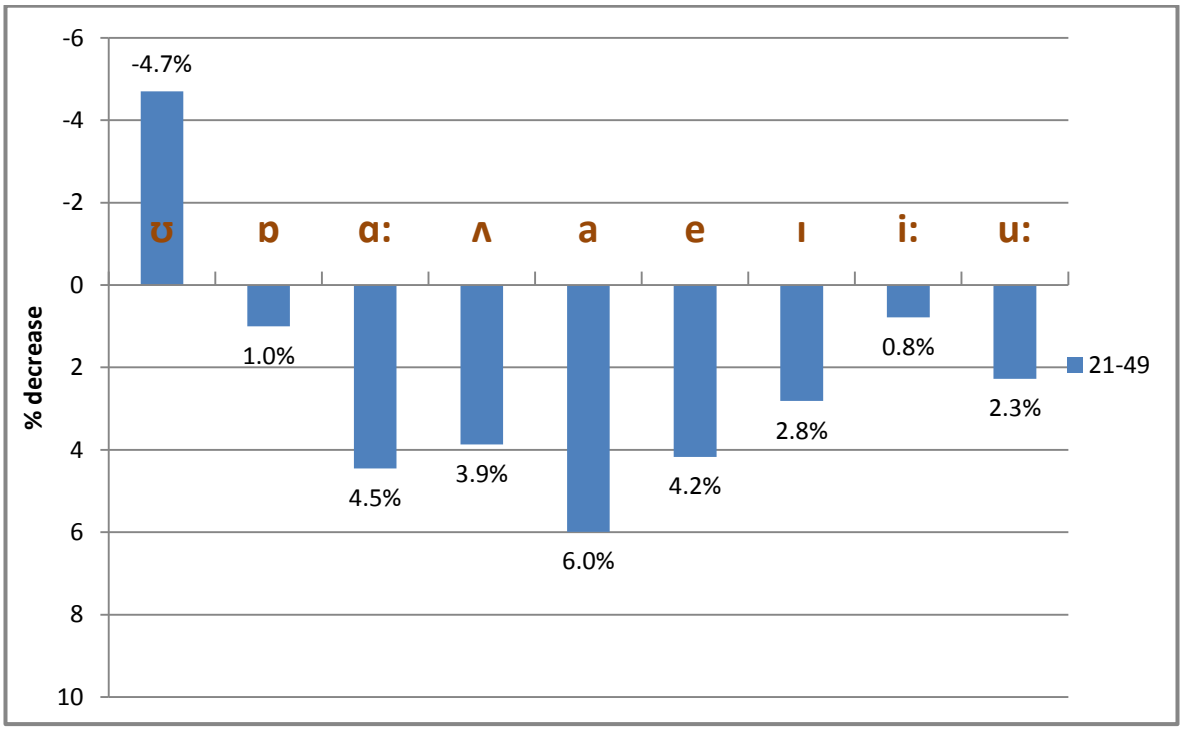


Figure 32 gives a breakdown of F3 reductions by vowel, which confirms that decreases are more limited than in F1 and F2. It also illustrates that F3 is similar to F2, as open vowels are more affected by age-related change than the others in this analysis, particularly /ʌ, ɑ: & a/. One surprising result is in the FOOT lexical set, where overall F3 increased by 4.7%; this increase was apparent in all but two speakers. Standard deviation (SD) for this /ʊ/ vowel was also much higher across speakers than for any other vowel type (152 HZ, with the highest other three vowels at 112, 102 and 102 Hz). As Table 28 shows, however, Ns for this vowel were relatively low, as they appear rarely in spontaneous speech; this may have affected the reliability of this measure, although this would not account for such large SD differences.

Table 28 - Showing F3 significance results for each vowel by each speaker by Univariate ANOVA (key in §4.1.2.3)

F3	Andrew	Bruce	Lynn	Neil	Nick	Suzy	Symon	Tony
ʊ					n		n	
ɒ		n	n	n	***	n	n	*
ɑ:		n		*	n	**	n	n
ʌ		*	n	*	n	***	**	*

a	**	***	n	***	n	**	***	*
e	n	*	n	***	n	n	*	n
ɪ	**	**	n	***	n	***	***	**
i:	n	n		n	n	n	n	n
u:		n	n	*	n	n	*	

What the above table illustrates is that F3 changes are significant in fewer cases than for F1 and display a similar pattern to F2 (although the percentage magnitude of change is lower). The same numbers of tests (26/59) display a significant result in F2 and F3, and in fact more F3 tests show a significant decrease.

For F3 then, there are generally reductions, although they are not to the same extent of F1 and (somewhat) F2, despite there being a similar number of significant age-related changes in F2 and F3 tests. As with F2, open vowels seem to be affected in more tests and for more speakers.

4.2 Exploring explanations using estimation formulae

This section presents results from further testing with formant frequency data in order to provide exploratory hypotheses about age-related vocal changes. They are not designed to account for variation, but to illustrate likely explanations with patterns in the data.

4.2.1 F1 changes as compensatory for F0 decline

In their study of 5 speakers, Reubold et al. (2010) propose that changes to F0 and F1 may be related (note they found no significant changes in F2 and F3). They argue that "age-related changes in F1 may be compensatory to offset a physiologically induced decline in F0 and thereby maintain a relatively constant auditory distance between F0 and F1." (Reubold, Harrington, & Kleber, 2010, p. 638). Although they present explanations for potentially physiologically-induced changes to F1, they reject a principally aging-related cause for this F1 reduction and instead posit that speakers are maintaining the auditory distance (to preserve acoustic-phonetic clues to vowel height) between the first formant and fundamental frequency. Using correlation of the linear regression of logarithms of F0 and F1, they demonstrate that F1 approximately tracked F0 across the years. They support this argument with perception experiments that showed F1 had a negligible

effect on recognition of age by listeners and by testing that F1 changes were not simply the 'acoustic artefact' of F0 changes.

While their study is broadly comparable to the present investigation, and offers useful data and explanations, there are a number of key differences. Their subject base covered a slightly later age group (between 35-40 and 65-75 years old) and a slightly longer delay (about 30-35 years). Furthermore they are all professional voice users, broadcasters, politicians or the Queen. The recordings were taken from the voice in use, across radio and television media and therefore the speech style is different from the interview style in the *Up* series. Moreover, their formant measures came only from LTAS measurements of schwa vowels, as a time-saving measure.

Further to these differences, there are limitations of Reubold et al (2010) which should be mentioned, particularly with reference to their argument for speaker compensation strategies as the cause of decreasing F1. The first issue to note is that their F1 measures are based only on schwa vowels, which may not represent the full range of speakers' variation. Indeed, it can be seen in the present study that different vowels (and groups of vowel types) react differently to aging, and furthermore in this varies within different formants. The study of different vowel types might lead them to revise their decision to reject an aging-dependant formant change, as they do in this paper, and in other papers by the second author (in a separate series of papers (Harrington, Palethorpe, & Watson, 2000a; 2000b; 2005; 2007)) for Queen Elizabeth II. They cite inconclusive literature, and cite two studies on larynx lowering (Flügel & Rohen, 1991) and increased vocal tract length (Xue & Hao, 2003) as reasoning for rejecting this explanation (although findings in the latter study did show lowered formants and extended oral tract length and vocal tract volume in older speakers). In contrast, they agree with the aging literature that extent of jaw opening may be responsible for decreased F1 (studied for the Queen in Beyerlein et al. (2008)), but describe this as speaker-motivated, and not as a physiological result of aging. Although the theory fits their observed data, it is unclear how their study is equipped to differentiate between a compensatory or physiologically-determined change to jaw or mouth opening.

Another concern is that they seem to draw most of their conclusions from data presented for only two of their speakers, Alistair Cooke and Queen Elizabeth II, and neglect to perform the same analyses on the other 3 speakers from their initial analysis.

Furthermore, the correlation of the linear regression coefficients of the F0 and F1 data from the Queen and Cooke are converted to a natural logarithm (F0) and a logarithm as a function of speaker age, because this produced "the most systematic relationship between F0 and F1" (Reubold, Harrington, & Kleber, 2010, p. 642). If testing the relationship between two features, caution should be taken in transforming the data to find the best 'systematic relationship'.

Returning to the present *Up* data, it is difficult to support this compensatory hypothesis. It can be observed below, and has been stated in §4.1.1, that for most speakers there were not clear reductions in F0, as indicated might be likely from the aging literature (although largely from more controlled laboratory or cross-sectional data). The two female subjects, Lynn and Suzy, were the only ones to show a decrease across the period, but for Suzy F1 decreases to 35 then increases after that; despite the reduction in her F0, this does not match the changes in F1.

Figure 33 - Graphs showing mean F0 across samples and mean F1 (in Hz) across all monophthongs, at each age stage, for each speaker

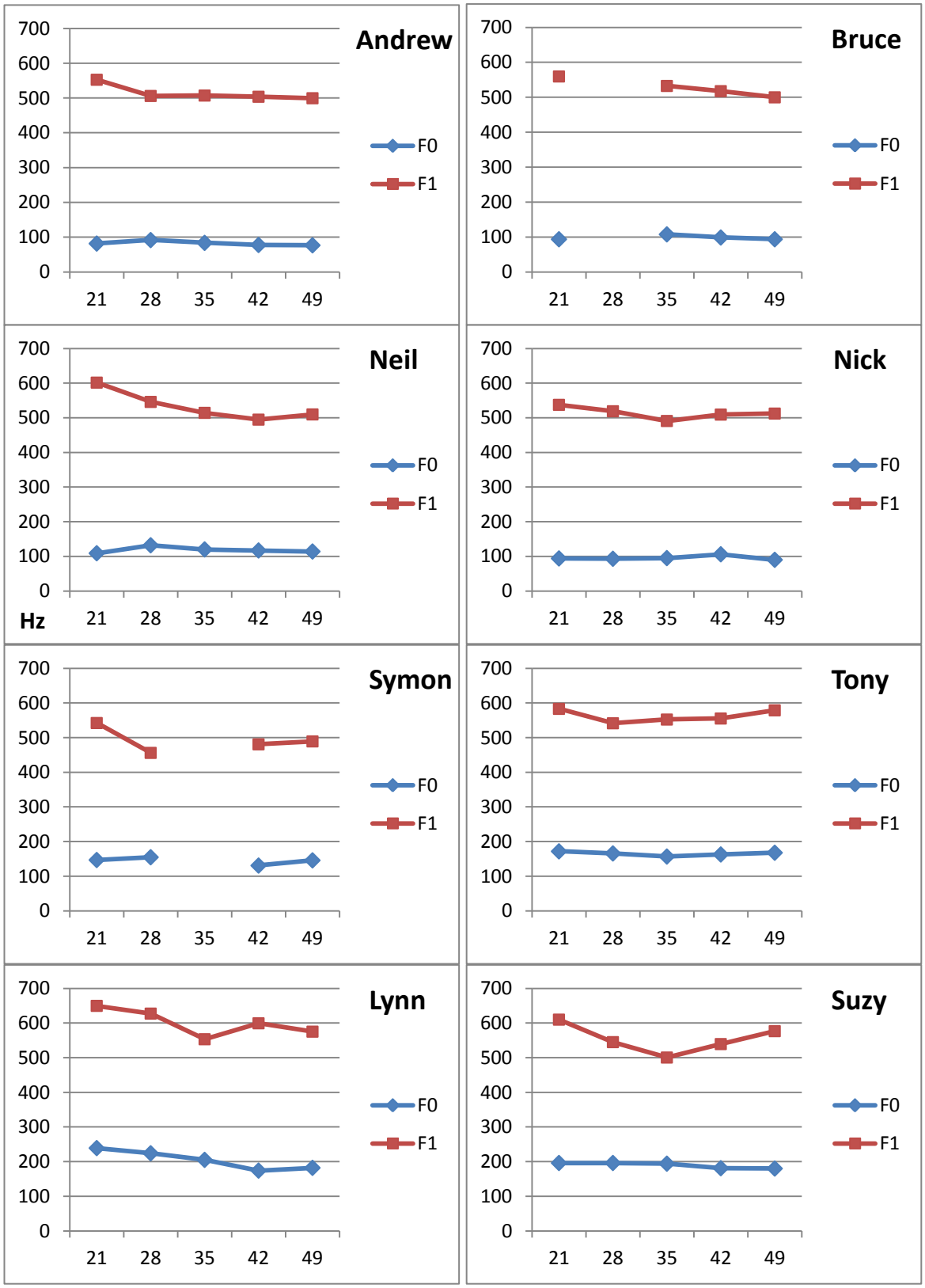


Table 29 below shows the Pearson correlation coefficients for mean F0 and F1 for all speakers. As might be expected from

Figure 33 above, only Lynn shows a positive correlation for reduction with a large effect (r=0.6587) between F0 and F1 across age. (Neil and Nick, in fact, show a small negative correlation effect.) While the data in Reubold et al. (2010) were subjected to a much more complex statistical treatment, it is not necessary to transform these data and correlate their linear regression coefficients, as it is fairly clear that there is not demonstrable relationship between F0 and F1 movements for most speakers.

Table 29 - Table showing mean F0 across samples and mean F1 across all monophthongs, at each age stage, as well as Pearson Correlation coefficients

Subject		21	28	35	42	49	Correlation Coefficient
Andrew	F0	82	92	84	78	77	**0.038**
	F1	553	506	507	504	500	
Bruce	F0	94		108	99	94	**0.043**
	F1	560		533	518	500	
Lynn	F0	239	224	205	174	182	**0.659**
	F1	649	627	553	599	575	
Neil	F0	109	132	120	117	114	**-0.235**
	F1	601	546	514	495	510	
Nick	F0	94	93	95	106	90	**-0.183**
	F1	537	519	491	509	512	
Suzy	F0	196	196	194	181	180	**-0.02**
	F1	610	545	500	539	577	
Symon	F0	147	155		131	146	**-0.096**
	F1	542	456		481	489	
Tony	F0	172	166	157	163	168	**0.666**
	F1	583	542	552	555	579	

Although the analyses and age ranges in Reubold et al. (2010) and the present study are slightly different, in general it might be expected that they would follow a similar pattern if their hypothesis (from the Queen and Alistair Cooke's data) to explain F1 changes were correct. The data from the current study do not seem to support their idea that F1 changes compensate for physiologically-induced F0 changes.

4.2.2 **Vocal tract length estimations (VTLe)**

4.2.2.1 Predictions

One theory put forward is that changes in formants and vocal patterns can be attributed to extension of the vocal tract. This is reportedly caused by a number of factors (see §2.1.4), principally extension of the facial skeleton and lips and weakening of the supporting structures around the larynx and respiratory organs, allowing them to drop

(Beck, 1997). It is known that a larger vocal tract results in lower frequency outputs (Fant, 1960), and we have seen that formants in this sample have shown a tendency to reduce, so we would expect these data to support a tract extension hypothesis.

4.2.2.2 Vocal tract length estimate results

Figure 34 – Scatterplot and line graph showing VTLe for each speaker at each age

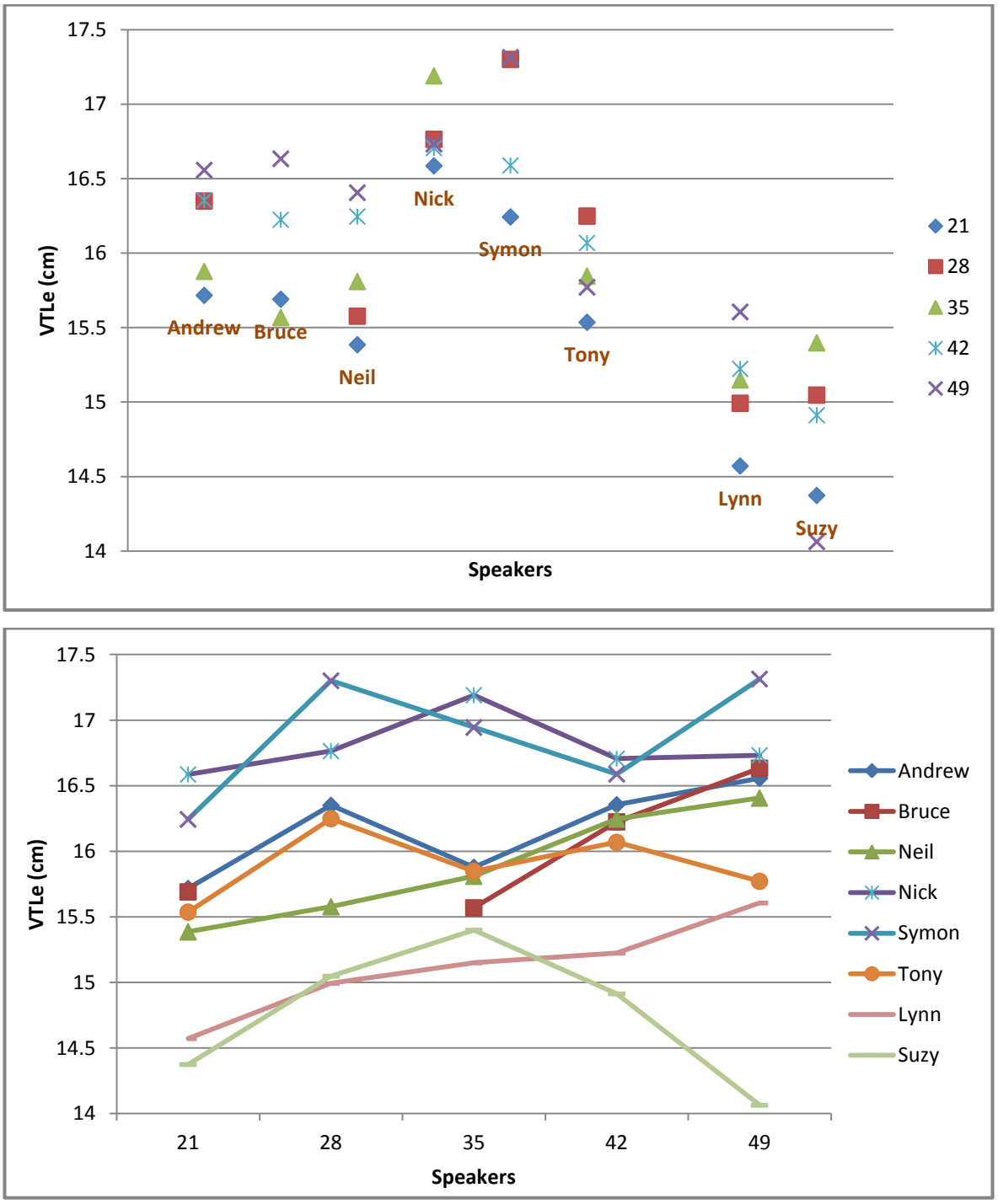
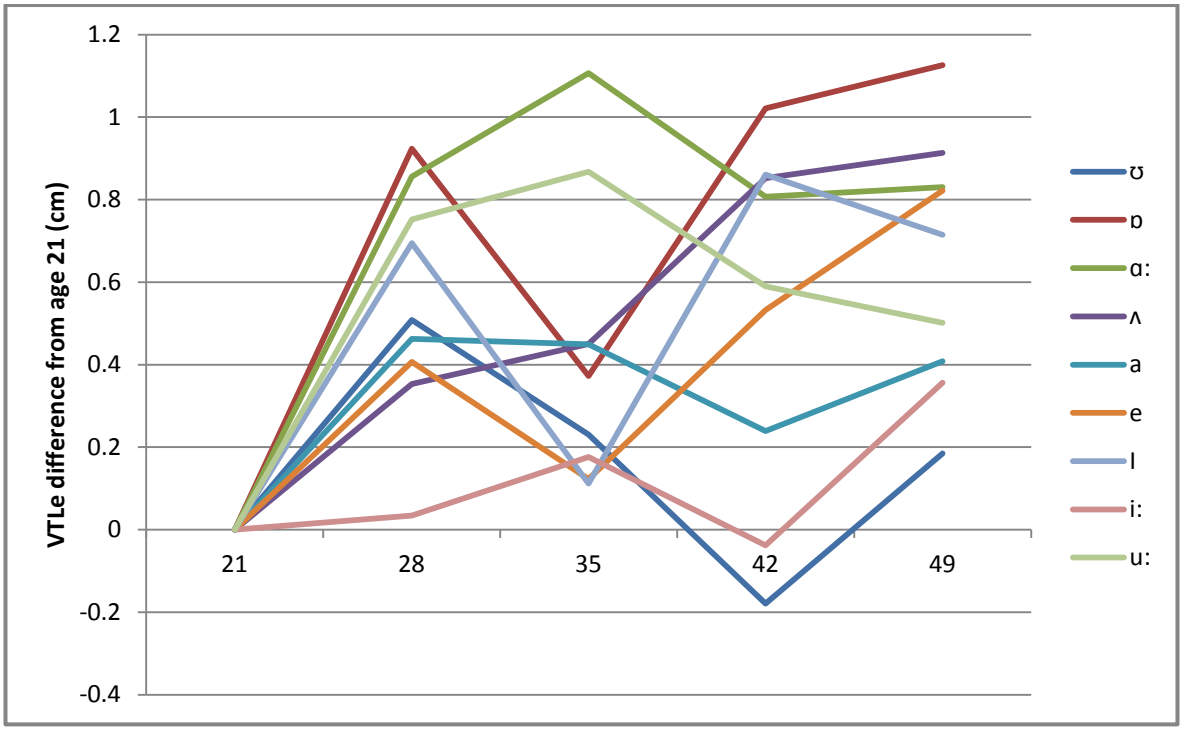


Figure 34 demonstrates that estimations, in general, support the possibility that vocal tract length changes could account for some of the changes in acoustic output. For half the speakers, VTLe increases in a steady linear fashion. Even for those who have a slightly

more mixed pattern, VTLe at 49 (purple x) is higher than at 21 (blue diamond) for all but Suzy. In those cases where increases occurred, VTLe changes by about 7.5mm. This figure is probably an overestimation of the actual physiological change given that other changes, such as reduction in F0, also affect overall formant frequencies. Figures for comparable differences (although over a larger time difference) between younger (mean 19) and older (mean 71) speakers (Xue & Hao, 2003) are around 50-100mm over that period, although this was in relatively small groups of different speakers. The line graph shows that for all but Tony and Suzy, VTLe is showing an increase between 21 and 49, and for most a fairly steady one (it should be expected that Suzy and Tony exhibit different VTLe as their formants showed overall different patterns to the other speakers, particularly in F1 for Tony; see §4.4 for formant summaries by speaker).

Figure 35 below demonstrates the average across all speakers across each monophthong analysed at each stage, illustrated as the difference from the 21 average (hence 21 values at zero). As we would expect, the general pattern is for an increase overall, with all vowels showing a greater VTLe at 49 than at 21. It is also noticeable that those with a more marked increase are the back vowels /ɒ, ɑ: & ʌ/. Of course speakers do not suddenly experience a longer vocal tract when producing these vowels, but perhaps there is an age-related change in physiology which predisposes speakers to producing back vowels with an extension of the vocal tract, via lowered larynx vocal setting perhaps.

Figure 35 - Line graph showing difference (mean difference across all speakers) in VTLe from age 21 for monophthongs



Even though we would expect VTLe measure to increase, reflecting formant frequency reductions, the formula used to calculate VTLe is weighted against the lower formants, which exhibits the most significant and largest formant reductions. Different physiological processes are posited as the cause of this lengthening, discussed in §2.1.4.4 and including lowering of the laryngeal musculature and respiratory organs, and extension of the lips and skull (although the increase in skull size is around 3-5% and probably would not account for the increases in VTLe completely). Again, it must be stressed that this is merely an estimate, and that VTL changes probably do not account for all the acoustic variation, but this does provide an interesting illustration and support for, in theory, a vocal tract lengthening hypothesis.

4.2.3 **Vowel space area estimations (VSAe)**

4.2.3.1 Predictions

As reported in §2.1.4, vowel space area is expected to reduce as speakers use a more centralised system of vowel production. However, this finding seems to be more consistent in elderly speakers, rather than the ages represented in this study. VSAe in this study is a two-dimensional geometric unit determined by plotting average F1xF2 points as a polygon and measuring the area of that polygon.

4.2.3.2 Vowel space results

Figure 36 - Vowel space area for all speakers at each 7 year interval

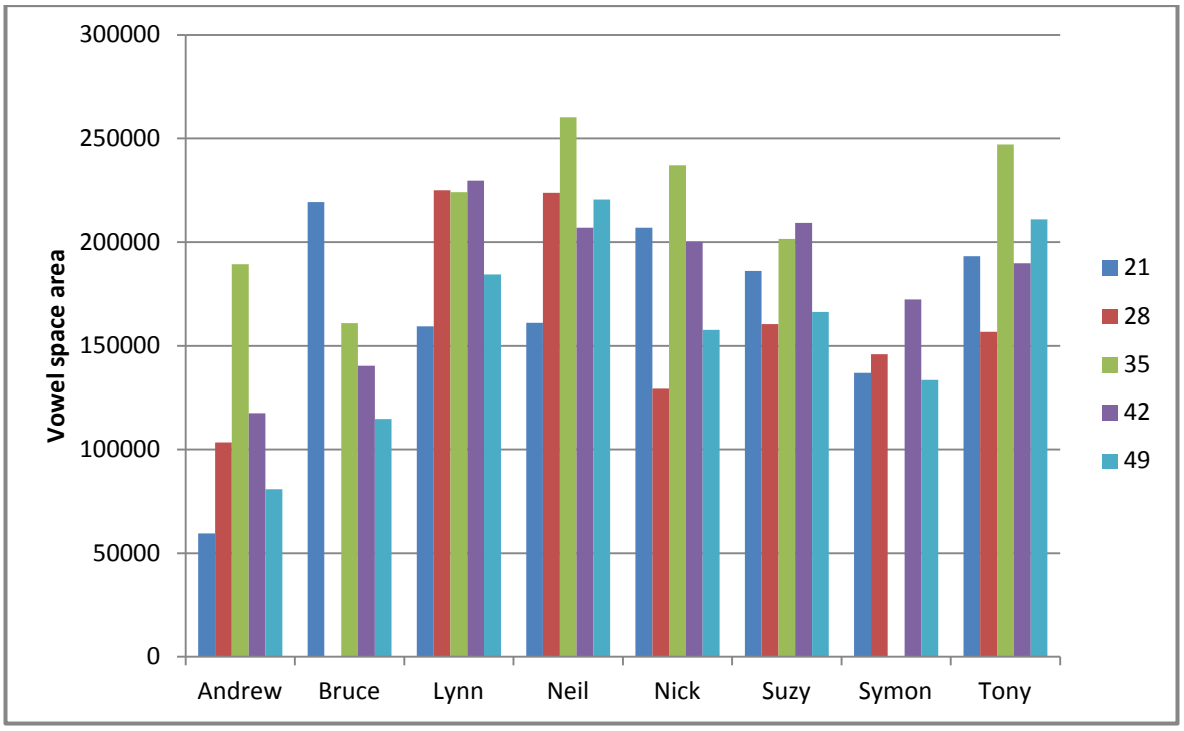
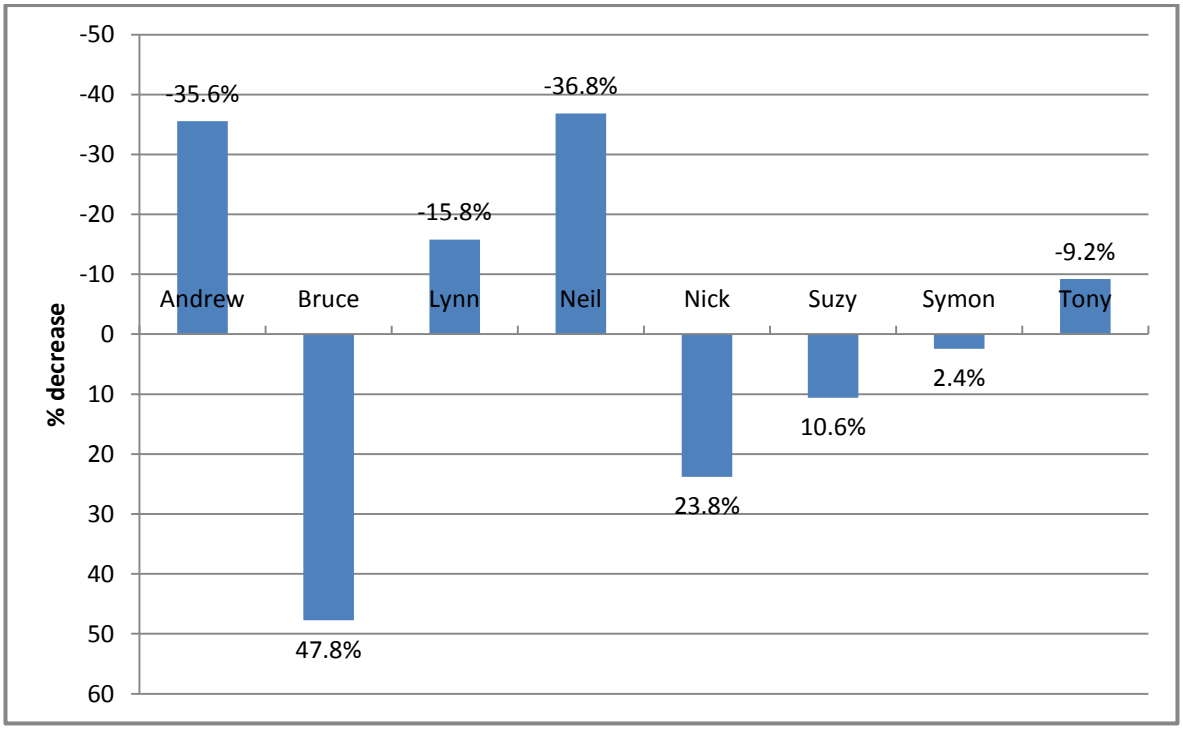


Figure 36 illustrates that there are varied patterns of changes in VSAe. Bruce is the only speaker for whom VSAe reduce consistently over the period as might be predicted from previous research. Figure 37 also shows that there are no consistent changes between 21 and 49 years, with four speakers' VSAe increasing and four speakers' decreasing, to varying extents.

Figure 37 - Percentage VSAe decrease between 21 and 49 years for all speakers (negative % entail an increase)



There is a trend, nevertheless, for VSAe to decrease between 35 years and 49 years, in contrast to predictions from previous studies. This may not be surprising, given that the prediction that VSAe would decrease was expounded from observations about a very small number of vowels, which was perhaps unjustified (although this was in an older sample than that in the present study) (Liss, Weismer, & Rosenbeck, 1990). It is apparent from Figure 36 that for most speakers, VSAe decreases from 35 onwards. Across all speakers (excluding Symon as he has no 35 recording), mean changes from 35 to 49 represents a 47% reduction in VSAe. Although this figure sounds large, it must be remembered that as a geometric measure of area, small increases in one or either dimension will entail large changes. Two of the main explanations for this change are reduced articulatory (lingual) movement and also restricted jaw opening (due to reduced flexibility in the temporomandibular joint) (Linville, 2001), which would affect the height vector.

Caution should be exercised in considering these results, however, as this method for representing a speaker's habitual vowel space is not widely used forensically, and these samples may be too short to give an accurate picture of a speaker's habitual behaviour. What these VSAe results do give us is an exploratory insight, from which it may be possible to make assertions about the causes of observed changes.

4.2.4 **VSAe and VTLe**

In summary then, VTLe in these speakers' increases from 21, this is a steady increase in half the cases, with VTLe nearly always highest in the 49 recordings. For VSA, all speakers exhibit decreases from 35-49 (except Symon, who was omitted from the 35 recording). Physiological explanations for VSA decreases (outlined in §2.1.4) include less use of less 'extreme' gestures, (indicated from formant transitions in Liss et al. (1990), thought they inferred change by comparison with other speakers from a different accent type and different consonant context) reduced flexibility in the movement of the temporomandibular joint (which would contribute to an explanation for larger decreases in F1) and atrophy of the tongue musculature, resulting in reduced lingual movement. Hypotheses which would explain expansion in the length of the vocal tract, or 'laryngeal ptosis' (Ferreri, 1959), include extension of the lips and the facial skeleton, lowering at the larynx, with weakening of the support ligatures and dropping of the thoracic cage.

It is important to remember that these estimated measures only present possible explanations for the formant results presented previously. They were arrived at by (in a sense) reverse-engineering formant data, so we would expect the patterns described above for VTLe (though perhaps not so pervasively given the weighting of formants) and also overestimation of actual differences in length. It is also not easy to separate these factors and their effects, nor try to classify them as either organic or phonetic, as in Laver's (1980) description of variation. VTLe seems an 'organic factor' (Laver, 1980, p. 9), but differences in vowel behaviour suggest a 'phonetic' element; similarly although VSA changes are principally down to phonetic factors, there may be underlying organic causes, such as reduced flexibility in musculature and joints.

The scope of this study is too limited to investigate the effects of these biological processes, and it is not possible to infer from these data the physiological causes that have been suggested. It would be sensible to assume that the average 7.5mm increases in VTL are an over-estimation, where tract extension plays one part in a range of physiological processes which lead to formant reductions. Further research into the causes of these changes within the same individuals might shed more light on this complex process. While some studies (such as Xue and Hao (2003)) have attempted to measure vocal tract changes, it would be useful to attempt to simulate the acoustic effects of these processes. This would be possible using synthetic vocal tract software

such as that developed by Birkholz and Kröger (2007) to determine what possible effect those changes would have on the speech signal and whether they represent a satisfactory hypothesis for observed age-related frequency changes. Furthermore, actual and more comprehensive longitudinal measures of changes to these physical properties and acoustic artefacts would give a more comprehensive overview of the causes of these changes. Again, this is beyond the scope of the current study.

What these estimation measures do indicate, however, is that there appear to be (at least) two age-related processes at work which can account for (at least) some of the acoustic variation, supported by hypotheses from previous research. The finding that this thesis can be more confident of is perhaps the fact that these appear to run on two separate timescales, with vocal tract length changing more steadily from the start of the third decade to the end of the fifth, whilst speakers' habitual vowel space appears to consistently decrease after around age 35.

4.3 **Overall summary**

This section addresses results more comprehensively and discusses how this interacts with existing research and explanations for acoustic changes. It will also elucidate what the data might lead analysts to expect from realistic data in FSC casework. F0 results may not demonstrate a clear age-related pattern for most subjects; however, there are apparent decreases for female speakers. What this might suggest is that observed reductions in formant frequencies are unlikely to be (largely) due to changes in the glottal source (Fant, 1960). The stronger factors in these changes are very likely rooted in changes in the filter (i.e. the vocal tract). Although there are overall reductions in both F0 and all formants (which Reubold et al. (2010) postulate are inter-related in older age), formant frequency changes are of a much greater magnitude than F0 changes. There are clearly other factors at work.

Formant frequencies for the current speakers are shown to reduce over time, in line with the majority of previous research (Endres, Bambach, & Flösser, 1971; Linville & Fisher, 1985; Linville & Rens, 2001; Xue & Hao, 2003; Reubold, Harrington, & Kleber, 2010). One of the most widely held theories for these changes is that the vocal tract is extended throughout the lifespan. This is due to a number of senescent processes, such as skull and lip growth, attenuation of the musculature which suspends the larynx and lowering of

the respiratory apparatus (see §2.1.4.10 for more detail). Estimations based on the current data (see §4.2.4) would seem to support that hypothesis; however, there are differences between formants that also need to be explained.

This thesis contributes to the expanding body of literature on aging as it analyses the different patterns of different vowels and formants, which are largely under-investigated as many studies employ long-term measurements incorporating all speech or single vowels. Explanations for these patterns in formants and vowel types are suggested, but need further investigation with more complex analyses. However, Linville (1996, p. 197) does postulate that:

> A pattern of variations in formant frequency change in different vowels might indicate variations in articulatory positioning with aging.

More research into overall articulatory settings might provide satisfactory explanations for differences between vowels which are apparent in these findings. Linville suggests that more definitive conclusions could be made based on these kinds of tests. It would be useful to observe these changes in line with formant changes, along with a formalised study of voice quality to illustrate interactions and inter-dependencies.

In the current study, F1 reductions are much more marked than for F2 and F3 (a finding that corroborates the results of Linville and Rens (2001)). Moreover, the F1 of close front vowels seem much more affected than other vowels. From these observations it might seem sensible to follow the suggestions from Reubold et al. (2010), that more limited jaw opening between 21 and 49 years has led to more significantly reduced F1, due to reduction of flexibility in the temporomandibular joint (Kahane, 1980). Reubold et al. (2010) simulated the effects of reduced jaw opening and found it affected stronger reductions in F1 than F2 and F3. Beyerlein et al. (2008) show results for the Queen from inverting an articulatory-to-acoustic model (Maeda, 1979; 1990), estimating jaw opening in a similar way to the VTLe estimations in a previous section (although using a different articulatory model). Their estimations show that mouth opening reduces with age, as does the variance in mouth opening measures.

In terms of casework, it is clear that age-related reductions in F1 between the third and fifth decades of life are the most consistent pattern, reported here (average 8.5% in the present study) and elsewhere, and that close front vowels show a more extensive decline in this study. This study does not concur with the findings of Reubold et al. (2010) in

describing a correlation between F0 and F1, as there were not similar distributions or changes for either feature. It does not, therefore, support their notion that speakers are actively lowering F1 to maintain a perceptual distance between that formant and F0. Rather, it seems more sensible to suggest that if jaw opening is more conservative, that this is due to the process of aging.

F2 and F3 behaved rather similarly, exhibiting slight decreases manifested more in open vowels, although the extent of F2 changes are on average slightly higher than for F3. However, for F3 there are fewer cases of significant changes which do not follow the general pattern. This is fairly easy to explain if we look at individual predictions made on sociophonetic bases, as we would expect F2 to shift in different manners due to accent changes, whether generally or due to mobility (there were also striking examples of this for F1). It is not as easy to determine why F3 changes were not of the same magnitude as those for F2, although changes in extent of jaw opening are simulated as having a lesser effect on higher formants (Reubold, Harrington, & Kleber, 2010). It could also be suggested that as F3 is more dependent on the nasal and sinusoidal cavities (Fant, 1960), which are far less plastic than the oral articulators, that F3 would be more resistant to change. It is possible that lesser changes are simply a result of being more independent of articulatory changes. Although it is difficult to tease apart the specific causes of some of these changes, they still provide a basis for a model for FSC exercises concerned with this kind of delay. In these cases we would expect to see changes in F2 and F3 that are less extreme than for F1.

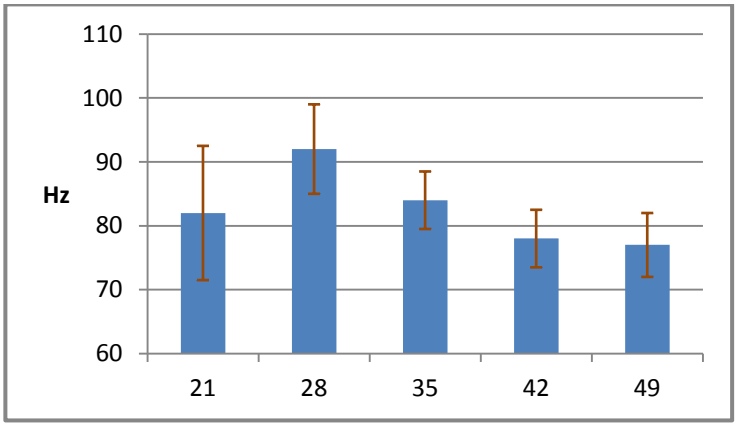
4.4 **Speaker profiles**

This section examines each speaker's data more closely in an effort to make more specific observations about the likely effects of lifestyle factors, such as geographical and social mobility, on measurable speech features and their reaction to aging. In particular it concentrates on observing overarching patterns of change. Where there are counter-examples to this trend, it presents possible explanations for differences. It is important to remember that these eight speakers represent a limited set of case studies with complex social and language contexts; however, they do present interesting findings. Speakers are ordered according to the level of geographical and social mobility they demonstrated throughout the series (tabulated in §3.1). Mean formants are calculated using an average value based on nine monophthongs analysed in the study.

4.4.1 **Andrew (non-mobile)**

4.4.1.1 F0

Figure 38 - Mean F0 for Andrew at each 7 year interval with SD bars



Andrew's F0 shows a typical slight decrease over the 21-49 period. There is a peak at 28, which is present for a few speakers, but probably represents variability of the data rather than a change at this time.

4.4.1.2 Formants

Figure 39 - Mean F1, F2 and F3 for Andrew at each 7 year interval

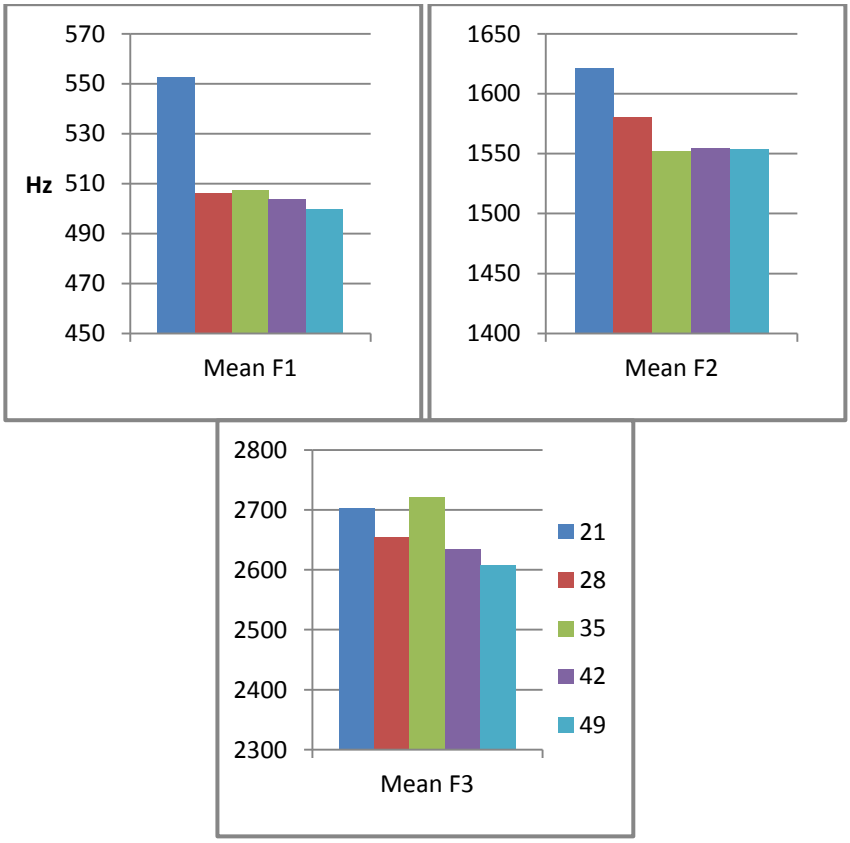


Figure 39 demonstrates the reduction in F1 and F2 over the 28 year period. Reductions in F1 were consistent across most vowels, although F1 was very high in some 21 vowel categories, leading to the high average at 21. Almost all vowels exhibited steady reductions in F2 as well. Significant results were found in a number of F1 tests and in one F2 test. However, Ns were low for Andrew and only four vowels had sufficient tokens at all stages to allow for a satisfactory ANOVA test. F3 also reduces steadily over the period, by around 100Hz between 21 and 49. Andrew, then, seems to follow the predicted pattern of reduction fairly solidly, as we might expect for a non-mobile subject.

Figure 40 - Scatterplot of mean monophthong values at each interval for Andrew, connected to represent vowel space. Histogram of vowel space area at each stage

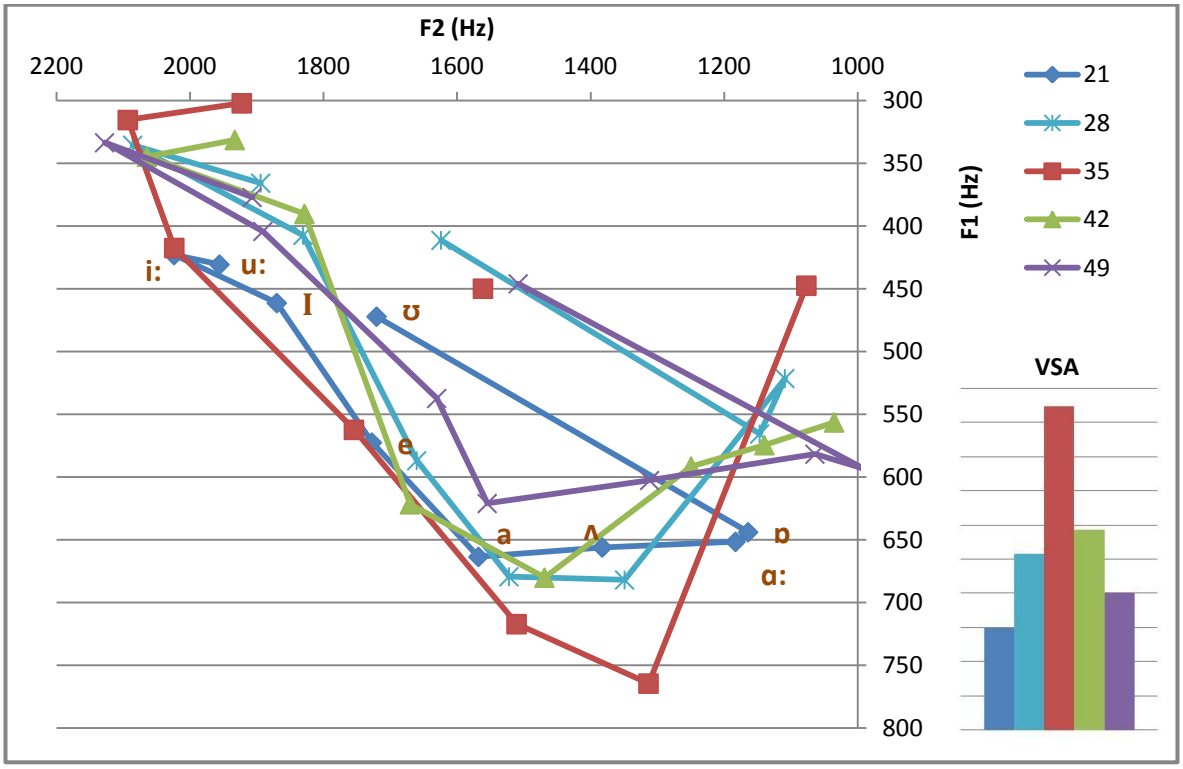


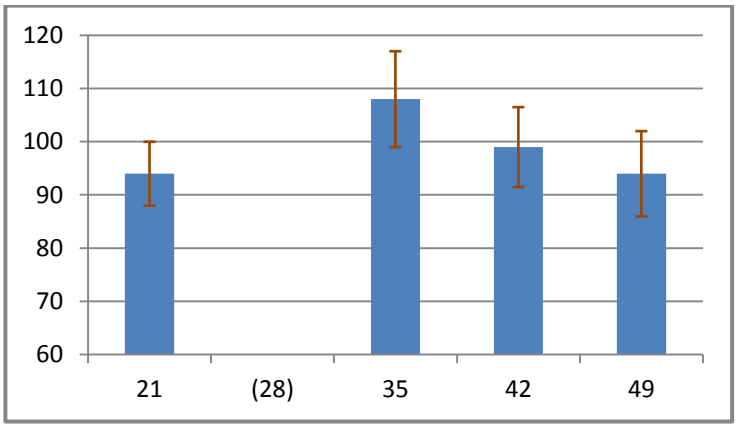
Figure 40 above shows that the vowel space area is contracting from 35 to 49 years. What it also demonstrates is that although there is no reduction in VSAe from 21 to 35, the reductions in F1 and F2 are causing the vowel space to displace towards zero fairly consistently over the whole period. This is clearest if we consider the difference between the two blue lines at 21 and 28 compared with the purple line at 49; even though the vowel spaces are a similar size, the 49 line is much further back and closer. Even for the red line at 35 where VSAe is greatest, there are a number of close front and back vowels which are closer to zero than for the early stages. VSAe show a steady increase from 21 to 35 and an almost parallel decrease from 35 to 49, the latter decrease is observed in most speakers.

4.4.2 **Bruce (non-mobile)**

Due to concerns about the sufficient quality of materials at 28 due to conflicting background noise, data are presented from four stages: 21, 35, 42 and 49.

4.4.2.1 F0

Figure 41 - Mean F0 for Bruce at each 7 year interval with SD bars



Bruce, similarly to Andrew, shows a pattern of reduction in the later stages. Like Andrew there is a peak at 35, and again, like Andrew, this is probably due to the variability within the speaker and the recording situation, where background noise is the likely cause of a 'Lombard effect' raising of F0.

4.4.2.2 Formants

Figure 42 - Mean F1, F2 and F3 for Bruce at each 7 year interval

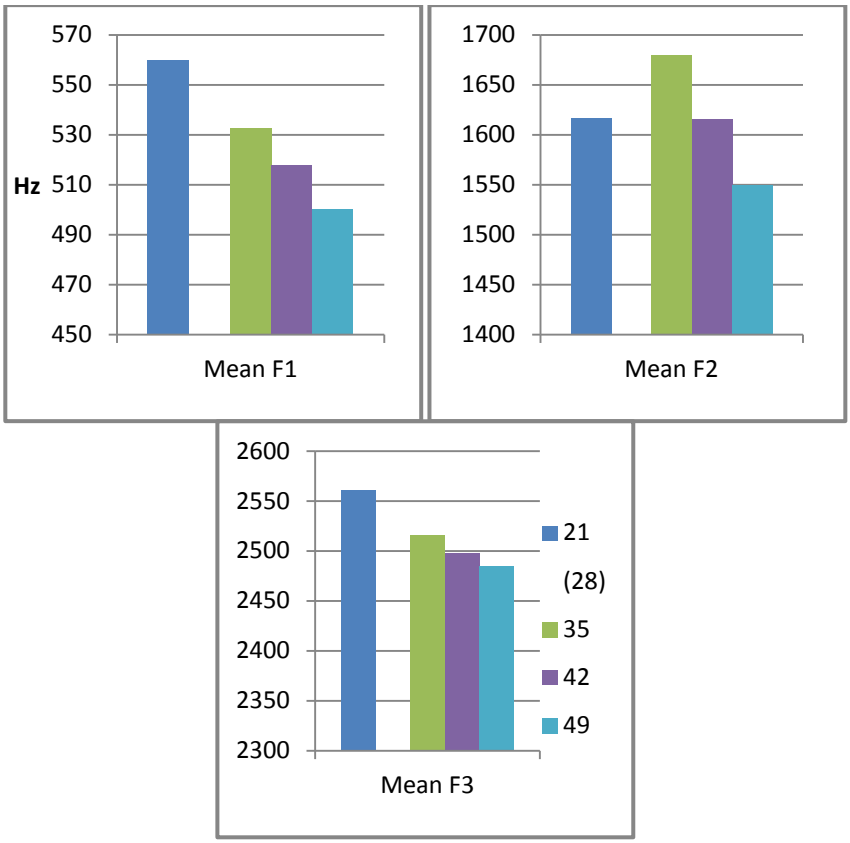
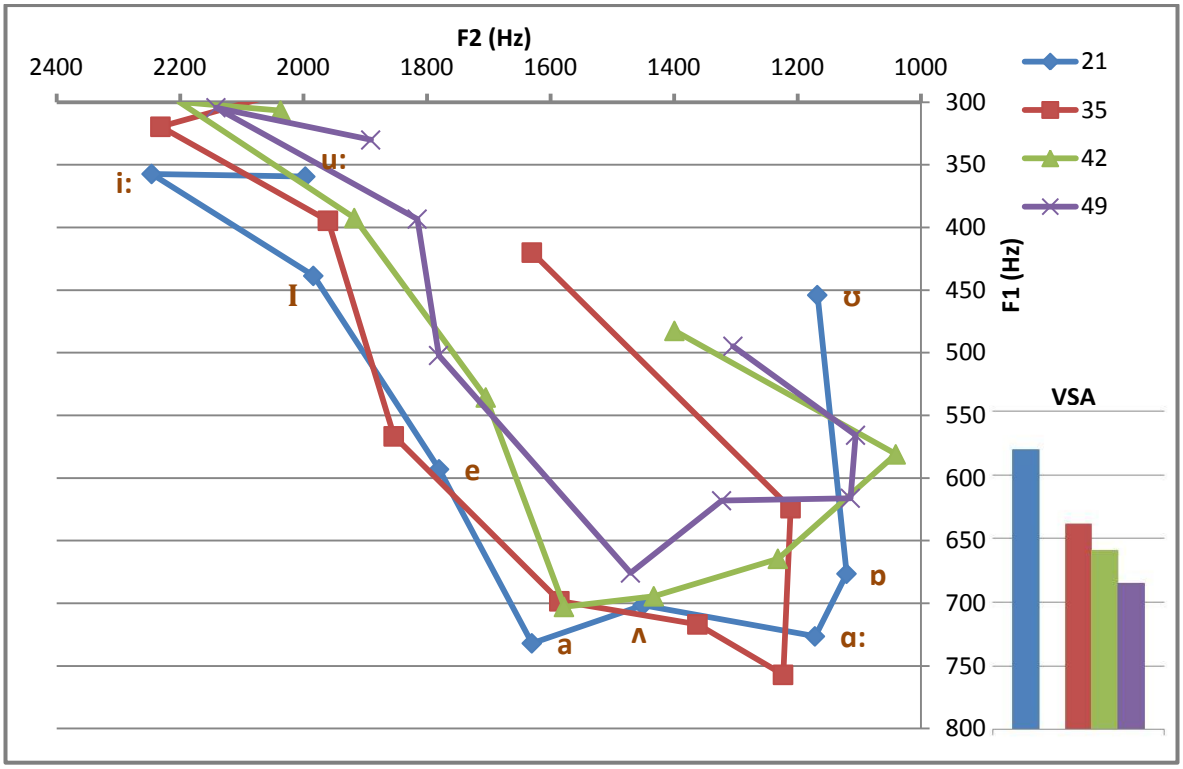


Figure 42 illustrates that Bruce, like Andrew, follows predictions from physiology fairly closely, with only a slight variation in F2. Between 21 and 49, F1 for Bruce decreases by around 60Hz and F2 by 67Hz. These effects are consistent over almost all vowels for both formants, although there are increases for F2 of /ʊ/, which was shown to be prevalent across all speakers in §4.1.3. F1 is significantly different in all but one test (all but /a/). F3 also decreases steadily over the period, by 77Hz overall.

Figure 43 - Scatterplot of mean monophthong values at each interval for Bruce, connected to represent vowel space. Histogram of vowel space area at each stage

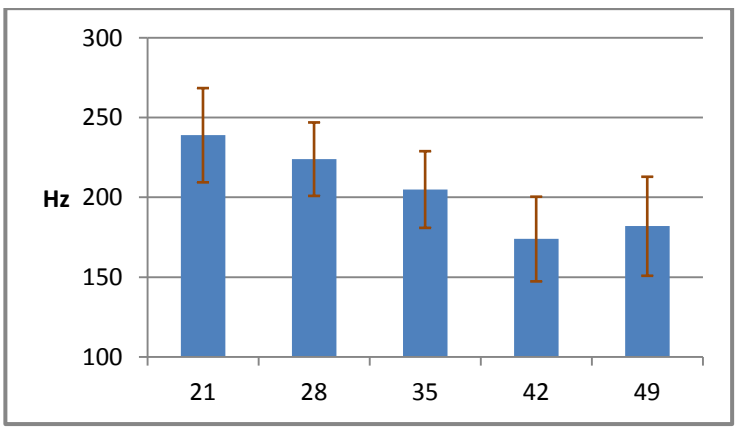


Bruce is the only speaker for whom VSAe matches the steady contraction prediction, as is clear from the histogram in Figure 43, VSAe reduces by almost half at 47.7%. The decrease from 35-49 is consistent with other speakers, however. The connected scatterplot also shows that the vowel space for Bruce, yet again like Andrew, displaces due to reductions in F1 and F2.

4.4.3 **Lynn (non-mobile)**

4.4.3.1 F0

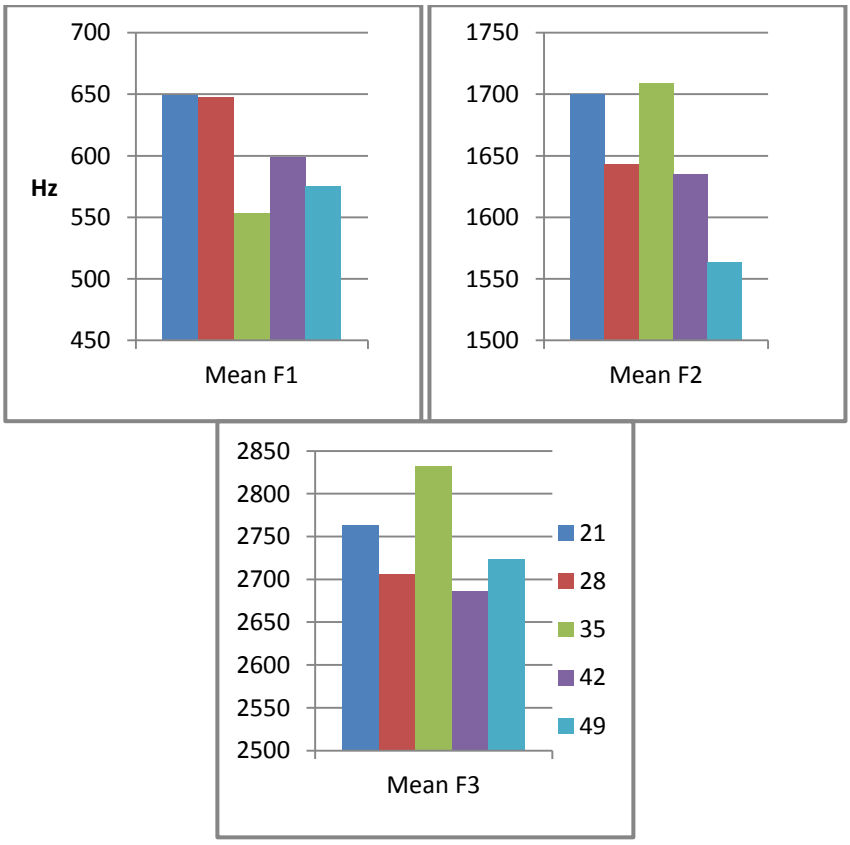
Figure 44 - Mean F0 for Lynn at each 7 year interval with SD bars



Lynn shows the most emphatic decrease in F0, as mentioned in the previous section. She exhibits a consistent reduction, with an overall reduction of 23%. This steep decline is probably due to the fact that she is a habitual smoker. Her F0 is also relatively high at the 21 stage. Lynn also displays a high SD, around 50Hz at most stages.

4.4.3.2 Formants

Figure 45 - Mean F1, F2 and F3 for Lynn at each 7 year interval



For Lynn, the formants are both behaving as we might expect for a non-mobile speaker, showing widespread reductions. This effect is fairly consistent across vowels, with no significant counter-examples. There is a difference at the 35 stage for both F1 (lower) and F2 (greater) and it would be interesting to observe any connected difference in audible voice quality, to see whether the effects are linked. For F1, this is mainly an effect of the open vowels and /u:/ (Ns for /u:/ were very limited at 28 and 35, 1 and 2 respectively). It is also possible there was particularly limited jaw opening in those recording sessions, for instance. F3 changes less drastically, although there is an overall decrease of 100Hz between 21 and 49.

Figure 46 - Scatterplot of mean monophthong values at each interval for Lynn, connected to represent vowel space. Histogram of vowel space area at each stage

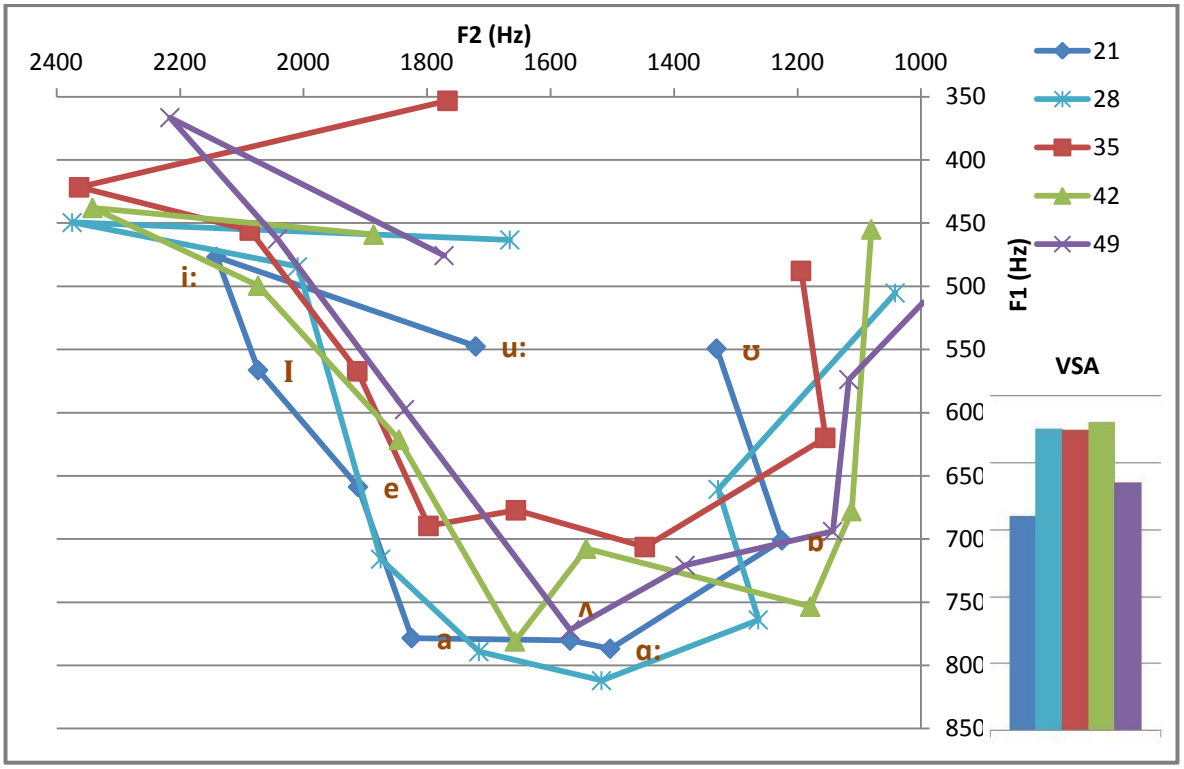


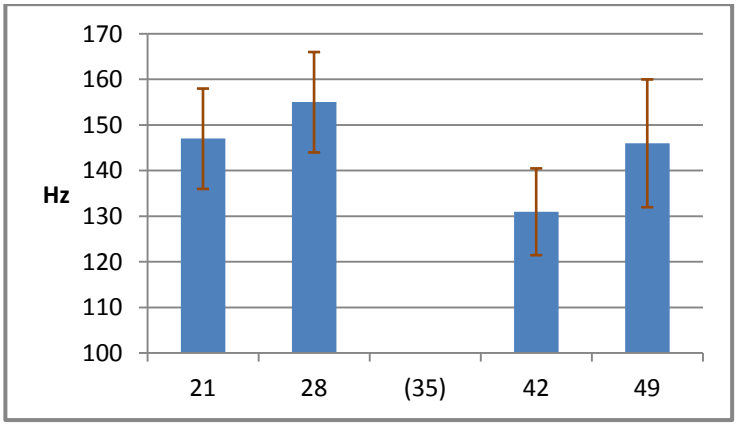
Figure 46 shows that VSAe for Lynn is not contracting overall, but does reduce over the three later stages. The scatterplot confirms what we expect from formant frequencies, a slight displacement towards zero.

4.4.4 **Symon (non-mobile)**

Data are presented at four stages for Symon as he did not participate in the programme '*35 Up*'.

4.4.4.1 F0

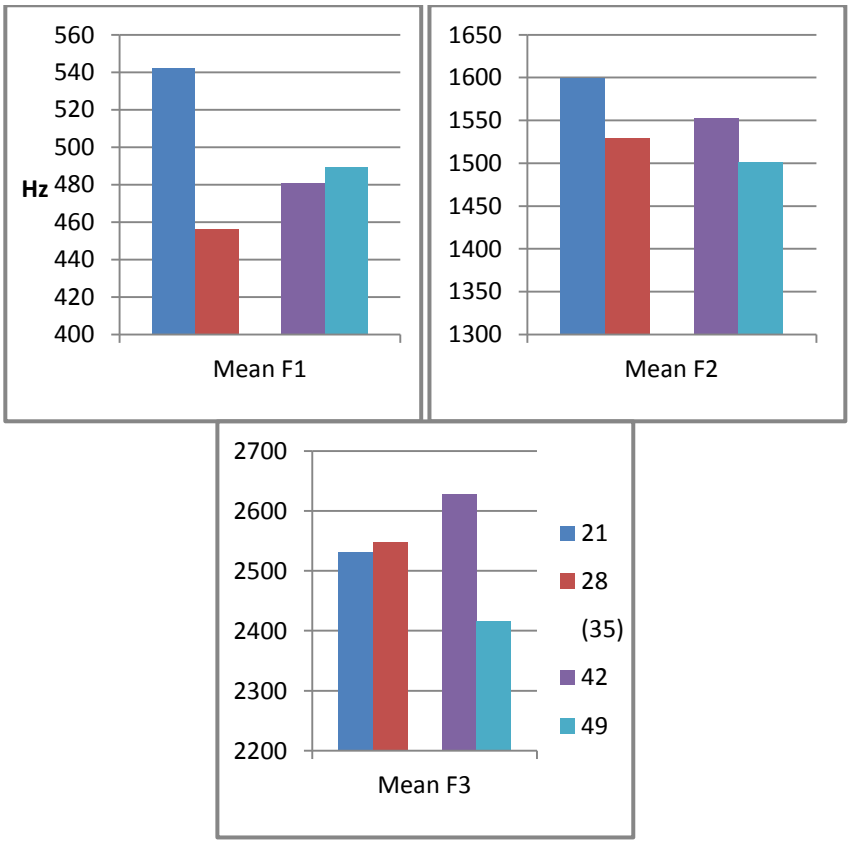
Figure 47 - Mean F0 for Symon for each 7 year interval with SD bars



Symon's F0 is relatively high for a male speaker, and does not decrease as we might expect from early to late stages. There are changes at 28 and 42, but these are not as great as the SD in each sample, which might suggest they are more due to internal variability between samples than steady change.

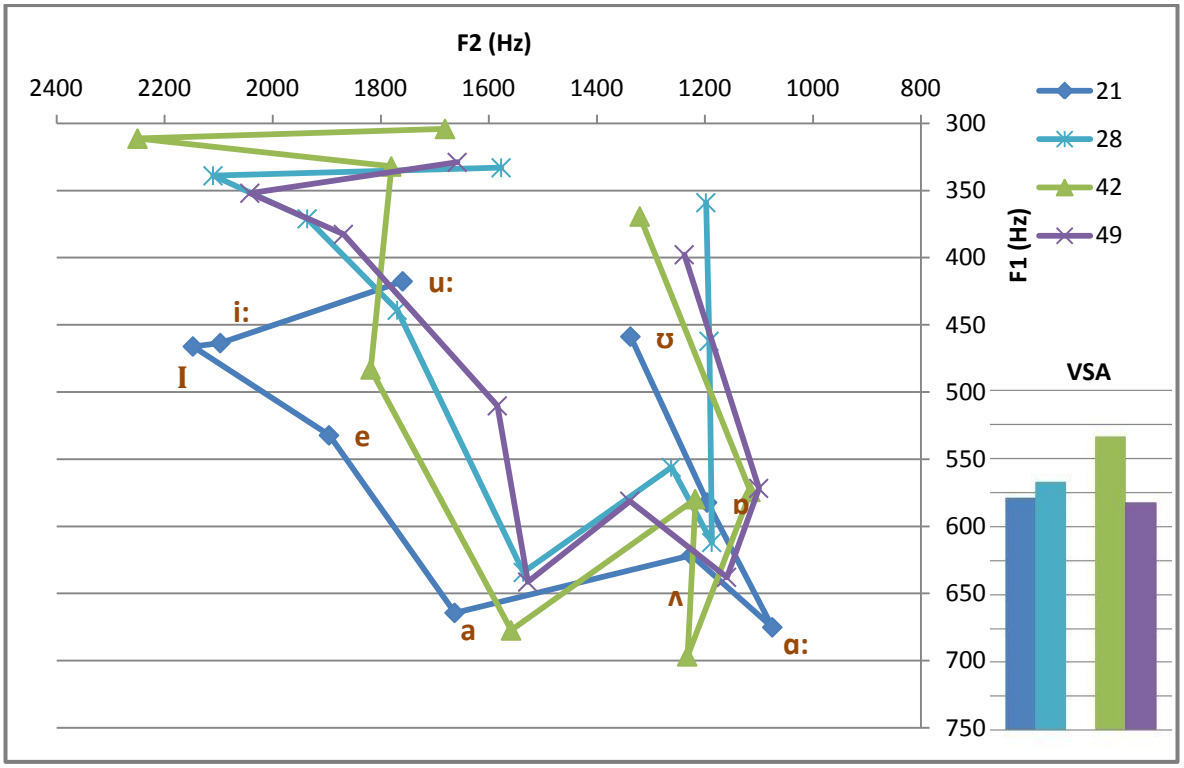
4.4.4.2 Formants

Figure 48 - Mean F1, F2 and F3 for Symon at each 7 year interval



Across all formants, mean formants are relatively low compared with other speakers, especially given Symon's relatively high F0 (compared with other male speakers in this set and Hudson et al. (2007)). For Symon there are decreases in all formants, most strongly in F1, as we would expect. F1 is reduced from 21 to the later stages, although this does not seem steady as there is a sharp decline to 28. There are significant F1 changes in all but two vowels (/a/ and /ʌ/), and four of those seven ANOVA results were highly significant ($p \leq 0.01$). F2 displays a similar trend but to a lesser degree, corroborated by fewer significant tests (3 of 9). The vowel /ɑ:/ showed highly significant increases in F2, which is surprising given the normal pattern for open vowels to reduce more. Mean F3 also reduced by 115Hz overall. Again we see a typical pattern for a non-mobile speaker.

Figure 49 - Scatterplot of mean monophthong values at each interval for Symon, connected to represent vowel space. Histogram of vowel space area at each stage



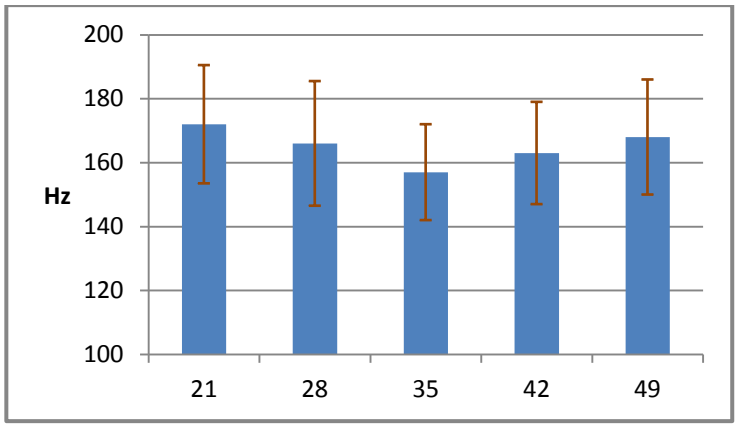
There appears to be a tilting effect in the connected scatterplot due to the sudden and steep decline in F1 between 21 and 28, especially in the close front and high back vowels. We can also see from the histogram in Figure 49 above, like most speakers in this sample, that VSAe is decreasing towards age 49, but the difference from 35 is unclear as that sample is unavailable. The vowel space is still displaced towards zero, combined with the tilting effect. Symon is another speaker who exhibits typical aging-predicted changes across almost all of his data.

4.4.5 **Tony (moderate mobility)**

Tony moves throughout the series, initially from central London (Hackney) to a neighbouring suburban county (Essex). He also purchases a second home in Spain where he stays (in *'49 Up'*), however, from the documentary it is fairly clear that the language community is largely made of ex-patriates with similar accents to Tony's original regions.

4.4.5.1 F0

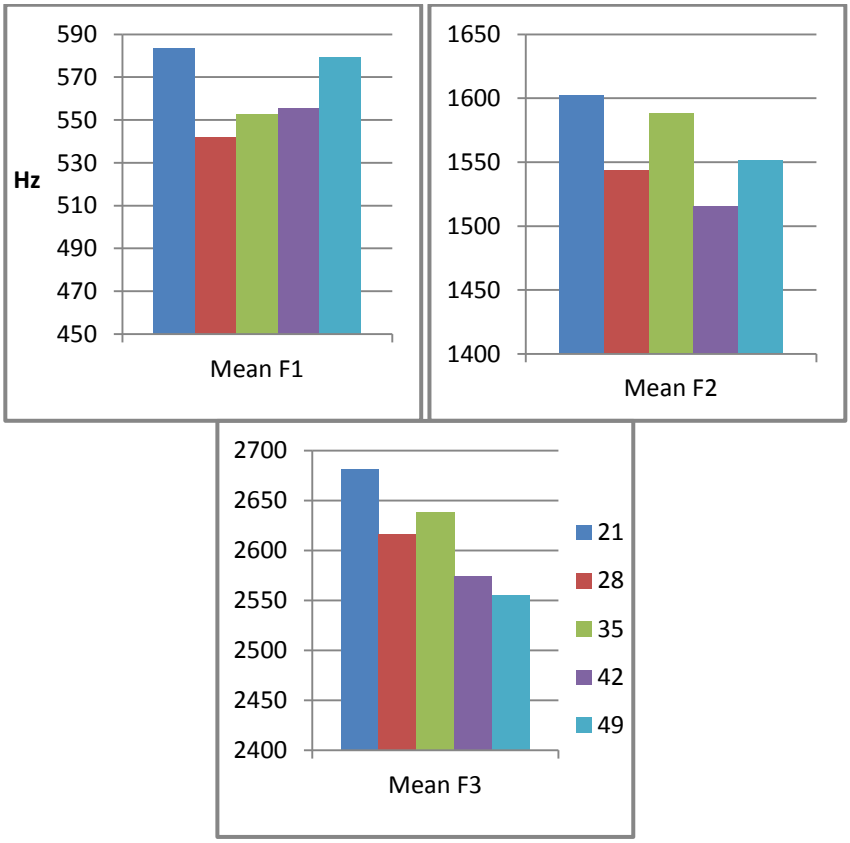
Figure 50 - Mean F0 for Tony at each 7 year interval with SD bars



Tony has a very high F0 for a male speaker, this might be expected given his relatively diminutive stature (he was a jockey for a short time). He also exhibits high SD within each stage. This SD should warn us to treat the results with a degree of caution. There are slight decreases across the period, but also increases and only a 4Hz difference between the 21 and 49 measures. Tony follows what we might expect from the general pattern in this sample, but not from the predictions made.

4.4.5.2 Formants

Figure 51 - Mean F1, F2 and F3 for Tony at each 7 year interval



The overall mean for F1 for Tony is perhaps a little obscured by three open vowels (/a, ʌ, ɑ:/) where the 49 is much higher than all other stages, as F1 is highest in these vowels the average has been upwardly affected quite strongly; the presence in open vowels suggests there may be more extreme jaw openings. However, in general most vowels show a decrease in F1 with aging. This is significant for /e/ and /ɪ/ only, and /ɑ:/ shows a significant increase. F2 displays an overall reduction, as does F3, and both show significant changes in around half of ANOVA tests.

Figure 52 - Scatterplot of mean monophthong values at each interval for Tony, connected to represent vowel space. Histogram of vowel space area at each stage

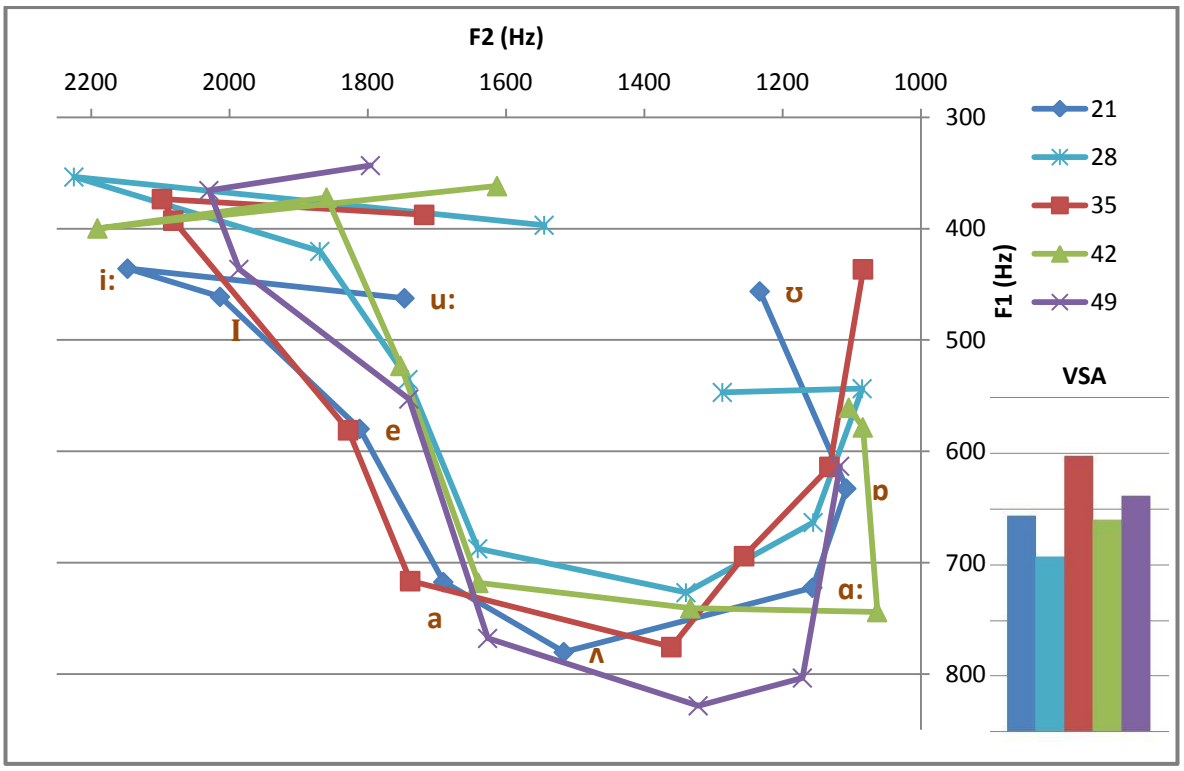


Figure 52 shows that while Tony follows the trend for F1 of close front vowels to decrease, open vowels are actually more open in the 49 stage, suggesting possible differences in jaw opening or articulatory behaviour. VSAe is slightly variable, but follows the general pattern of decreases post-35 years. The overall shape of the vowel space is relatively stable, although there is displacement, particularly in the F2 plane.

Although Tony is somewhat mobile, there are not very significant widespread differences in the accents of the regions and typical (of this sample) age-changes are still observable. Tony is clearly aware of the way he speaks, mentioning it in the documentary, and also talks a lot about a general Cockney or Hackney identity, of which accent plays a part. It would seem sensible to suggest that he would be likely to preserve this accent more than

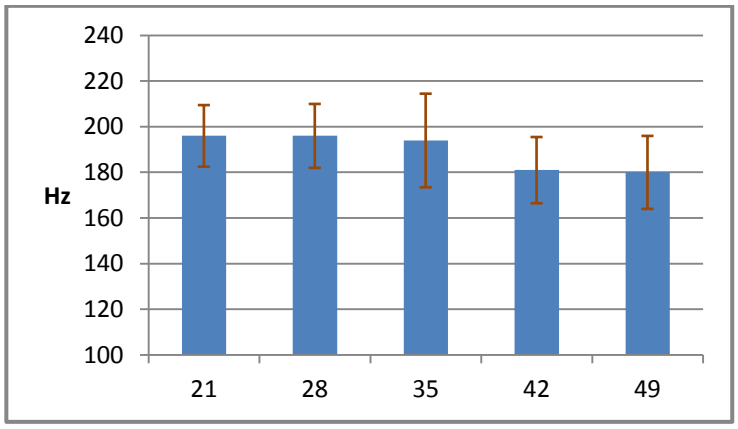
others, for whom this would have no benefit, as part of what he perceives to be his cultural heritage. Although Tony is fairly socially mobile, his sense of geographical identity seems stronger than any aspiration to conform to a more standard accent.

4.4.6 **Suzy (moderate mobility)**

Suzy is fairly mobile geographically, starting the *21 Up* programme living in Scotland (although staying at her parent's 'other' home there). Nevertheless, her linguistic background is clearly rooted in a high-prestige upbringing, being privately educated throughout, sometimes in preparatory schools; even at young ages Suzy speaks in a form of U-RP. She does, however, move to the Bath region in *35 Up.* Although this represents a large geographical move, Bath is still generally a broadly standard accent location and Suzy lives within an upper-middle class family (married to another RP speaker). This upper-middle class status from 28 years onward could be considered a downward class shift, considering Suzy's very privileged education and childhood setting.

4.4.6.1 F0

Figure 53 - Mean F0 for Suzy at each 7 year interval with SD bars



Suzy has a steady decrease in F0, with an overall 8% reduction from 21-49. F0 is fairly steady from 21-35 and shows a 15Hz drop at 42 and 49. SD is a similar proportion of F0 as it is in other speakers. Suzy is a smoker at 21, but there is no further evidence of continued smoking.

4.4.6.2 Formants

Figure 54 - Mean F1, F2 and F3 for Suzy at each 7 year interval

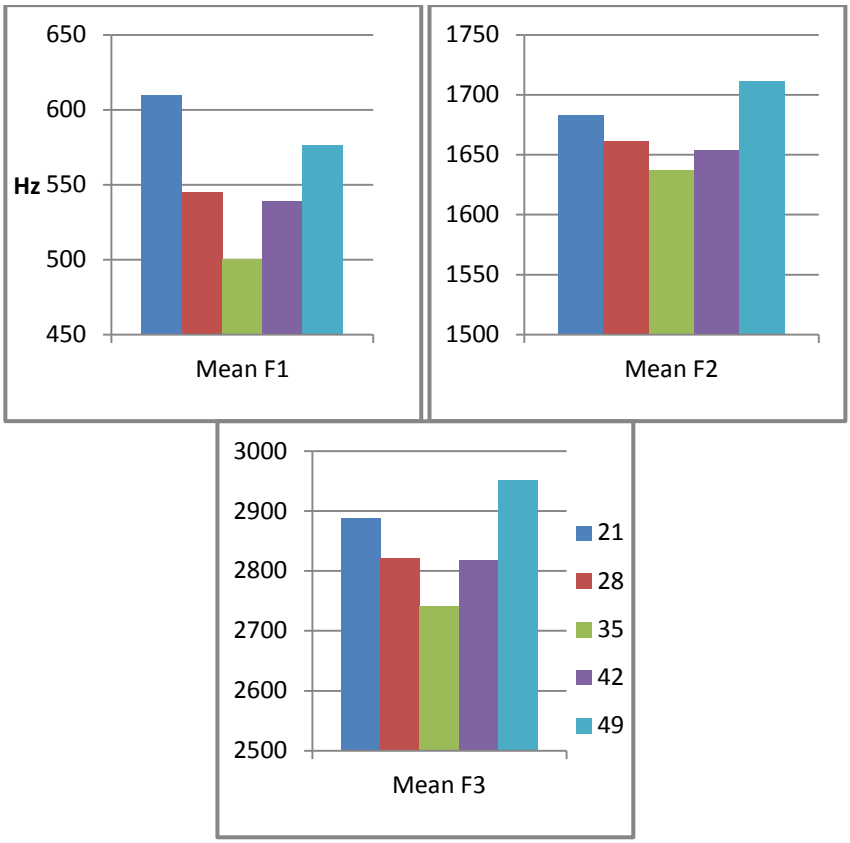


Figure 54 demonstrates a fairly consistent pattern across all formants for Suzy: a decrease from 21-35 followed by an increase in 42 and 49. However, in the cases of F1 and F2 this is rather misleading. For F1, this later increase is highly significant for /a/ and /ɑ:/ only, all other vowels are either decreased significantly (/ɒ/, /i:/ and /ɪ/; we would expect this in the close front vowels) or are non-significant (but still lower in 49 than in 21). This increase is carried by a few vowels for which Hz values are highest, and this pattern is the same for F2, where three close front vowels (/i:/, /ɪ/ and /u:/) i.e. those vowels with the highest F2 frequency are increasing at 42 and 49 and leading the mean formant to show an increase. In fact only /i:/ shows a significant increase effect of age, and only /ɑ:/ shows a significant decrease.

One hypothesis that could explain these changes which do not follow the general pattern of aging is that Suzy is moving from a traditional RP accent to a more widespread variety, either as part of class mobility or, as in the case of the Queen in Harrington et al. (2000a; 2000b), conforming to more widespread varieties of the standard English accent. In reality the process is most likely a combination of both factors, as fewer people use a U-

RP (Wells, 1982) type accent and as SSBE type accents become more widespread, alongside Suzy's own social movement.

Figure 55 shows very broadly Suzy's shift in mean formant frequencies from age 21 to 49, and there are two patterns we observe which do not match a general prediction from a model of aging. Firstly, the two most open vowels increase in F1; although we would expect them to decrease less than closer vowels in F1, we would still expect an overall reduction in F1. Secondly, the close front vowels (and even /e/) increase in F2 over the 28 years; similarly to the previous case, we do not expect close front vowels to reduce in F2 as much as those further back and open, but we would still expect a reduction. There are, however, mainstream changes in the RP accent which match these patterns. Concurrently to the recordings in this sample, Hawkins and Midgley (2005) have shown that F2 of close front vowels has become increased and that TRAP class words are produced with a more open vowel, moving from /æ/ to /a/. There are not, however, matching mainstream changes for /e/ and /ɑ:/ which might explain their increases. It is possible that due to changes in other vowels, Suzy is preserving distances between vowels.

Figure 55 - Scatterplot for mean monophthongs for Suzy, arrows showing shift from 21 to 49

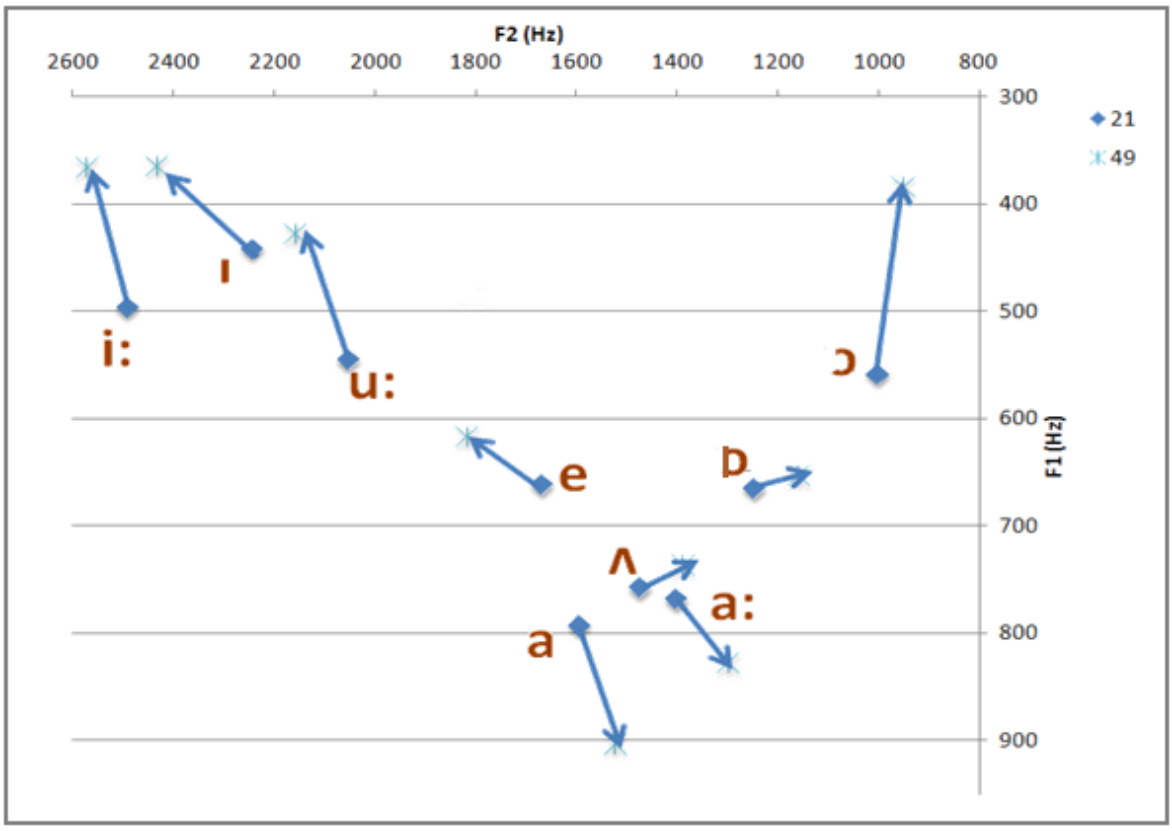
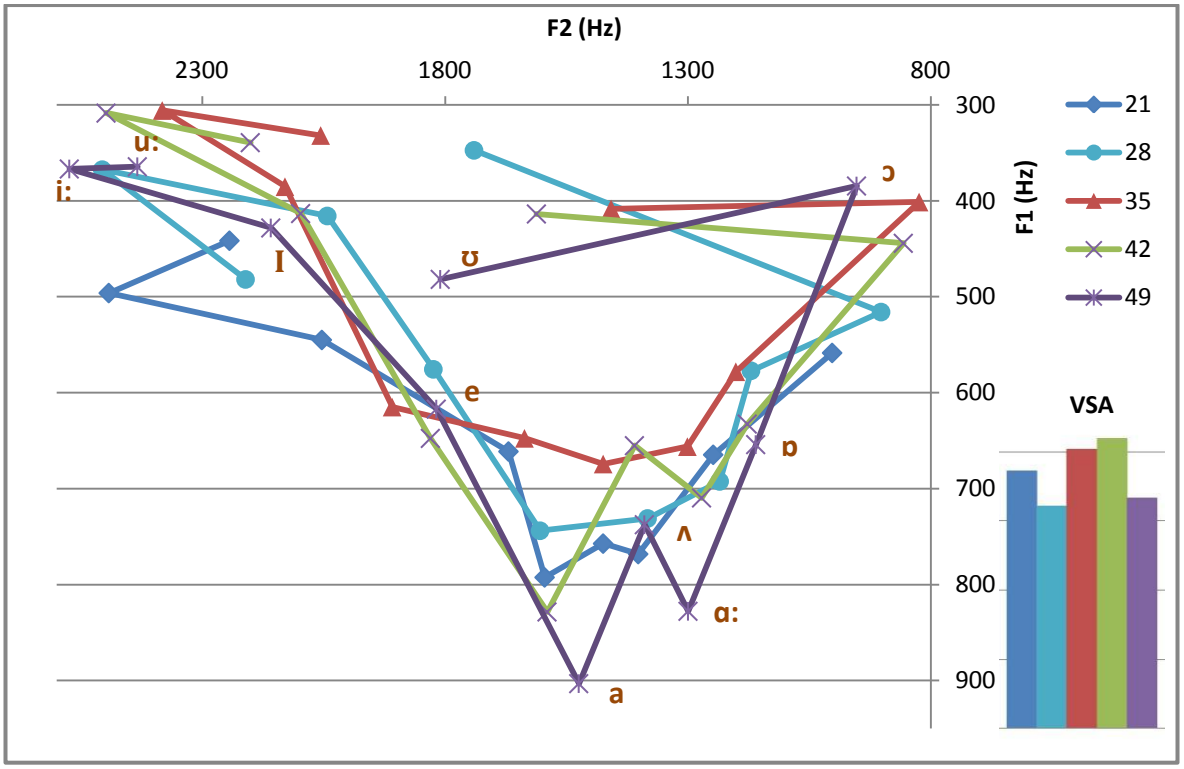


Figure 56 - Scatterplot of mean monophthong values at each interval for Suzy, connected to represent vowel space. Histogram of vowel space area at each stage



Some of the changes mentioned above are observable in Figure 56, especially the increases in F1 of open vowels in the purple (49) line. From the figure above, we can see that VSAe are not quite consistent with other speakers, but still show lower values at 49 than at 35.

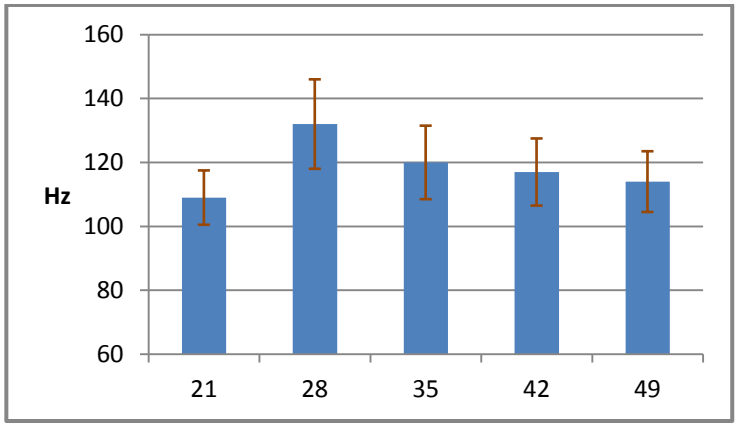
For Suzy, there is a mixed pattern. We can see age-related changes in F1 of close front vowels, for example, but also counter-examples that need explaining using a different prediction, in this case based on mainstream accent changes. This does present the problem that these kinds of predictions have to be based on recorded literature or on information about a speaker that a speech analyst may not have access to, especially in cases of non-cooperative suspects. In many cases, historical changes to accents are not recorded, especially for non-standard varieties. This could present the analyst with a problem in trying to explain changes which do not follow an expected pattern. Similarly if it were speaker information that was important, such as living or educational locations, without this the analyst would not be able to make such predictions or build a satisfactory model of probable speech changes. It is also difficult to determine the roots of accent change, whether this represents a class shift or whether the speaker is conforming to mainstream changes.

4.4.7 **Neil (highly mobile)**

Neil is extremely mobile during the series. He grew up in Liverpool, at 21 he is working in London having left Aberdeen University, at 28 he is living homeless in Scotland, at 35 he has moved to Shetland. By age 42 he has returned to London (Hackney) and at 49 we find him near Carlisle. Neil voices his very strong aspirations to become a lecturer or public figure, getting involved with local politics in both Hackney and Carlisle where he has to speak professionally. From auditory analysis it is clear that most Liverpool accent features have been lost even at 21, instead he is using a broadly SSBE type accent. Neil also suffers from non-specified mental health problems, which he terms 'a nervous complaint', the articulatory/acoustic consequences of which are difficult to predict.

4.4.7.1 F0

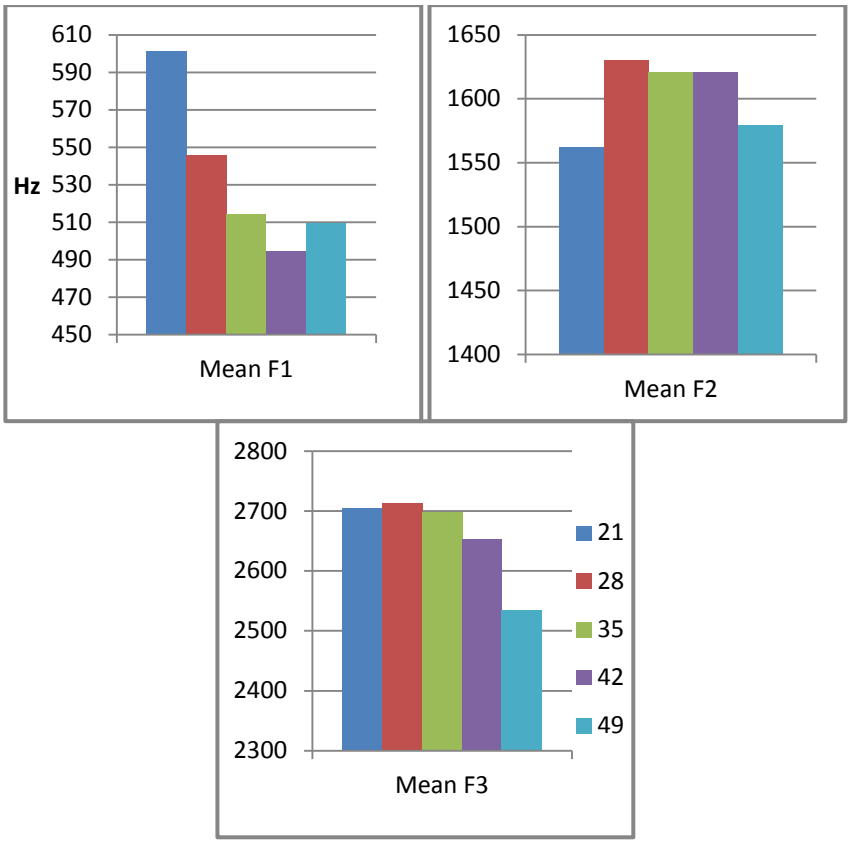
Figure 57 - Mean F0 for Neil at each 7 year interval with SD bars



Neil's F0 shows a steady decline from 28 until 49 years. Like many of the speakers there is actually an increase from 21-28, in Neil's case this is quite substantial; SD is higher at this 28 stage also.

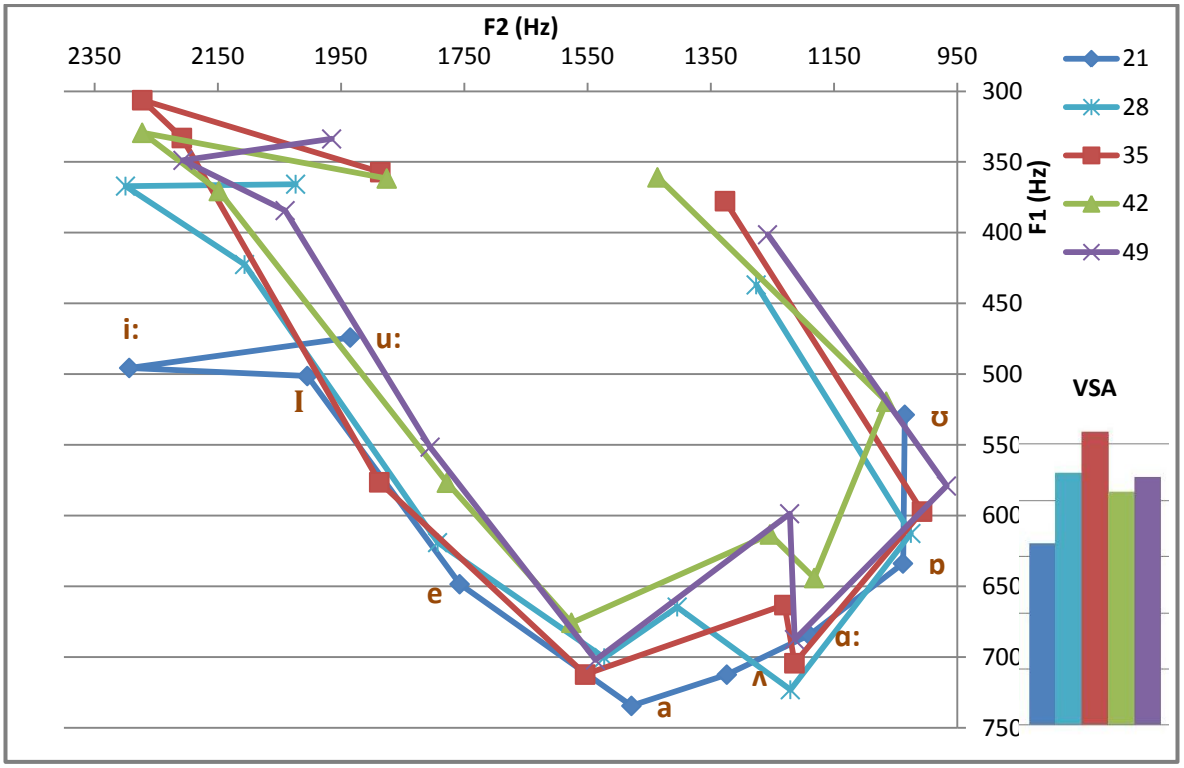
4.4.7.2 Formants

Figure 58 - F1, F2 and F3 for Neil at each 7 year interval



F1 changes are fairly steady and consistent across vowels, with large reductions from 21 to other samples in the three close front vowels and /ʊ/; it is these four vowels which cause the steep difference between 21 and 28 in F1 in Figure 58. There is a highly significant decrease in all but two open vowels (/ɑ:/ is only significant at the $p \leq 0.05$ level and /a/ is not significant), which follows the overall trend. The stark change in F1 is observable in Figure 59, especially in the dark blue line for close front vowels. Overall there are steady decreases, with each stage upward of the last. In terms of F2, there are reductions from 28-42 years, but F2 at 21 is lower than all other averages; this effect is consistent across nearly all vowels. Age is a significant factor in fewer vowel tests, with two increases (/a/ and /e/) and two decreases (/ʌ/ and /ɪ/). It is clear from the vowel space below that Neil's F2 is relatively stable and that contributes to one of the most invariable vowel spaces out of all the speakers (except F1 at 21). This may be surprising given the extent of mobility exhibited by Neil and also his mental health issues. F3 shows a fairly steady decline with age, which we would predict. This effect is consistent across all vowels and is significant in three of nine tests, highly significant in three and not significant for /ɒ/ and /i:/.

Figure 59 - Scatterplot of mean monophthong values at each interval for Neil, connected to represent vowel space. Histogram of vowel space area at each stage



It seems that for Neil, geographical mobility is a much less important factor than class or at least social identity. Neil's vowel realisations are very consistent across the period, especially from 28-49 years; from close auditory analysis Neil retains a broadly SSBE accent type and does not seemingly appropriate accent features from Scotland or Hackney, for example. This consistently could also be partly due to the large number of tokens of each vowel available for Neil.

There are, however, differences which could be explained by changes from his original Liverpool accent at 7 and 14 years to a more standard variety. For example, there are increases in the F2 of vowels in the TRAP set over the three decades, where we would expect to see the greatest reductions in that formant frequency. This could be a result of moving from a system of Liverpool English to SSBE, where in general this would entail an increase in F2 of /a/ as Liverpool TRAP is relatively back (Ferragne & Pellegrino, 2010). VSAe show a decrease after 35 years, and similarly to other male speakers, an increase from 21 to 35. This pattern of losing 'home' features but a reluctance to adopt novel features from a 'host' accent is noted in previous studies (Nycz, 2011).

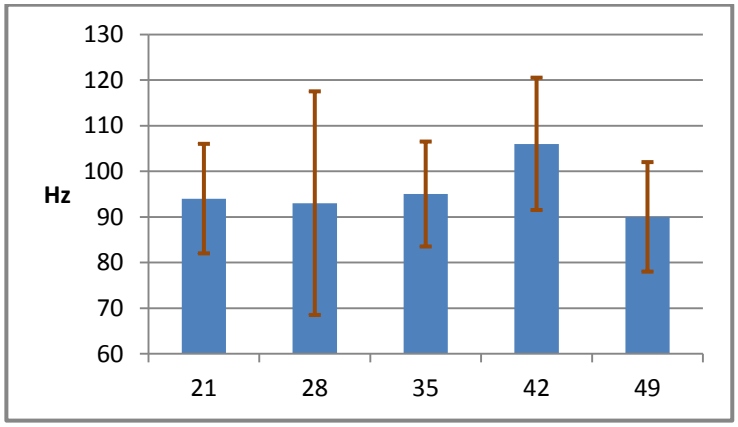
For Neil then, there are predictable reductions in formant frequencies overall, and a surprisingly invariant overall vowel system. There are quite marked changes in the F1 and F2 measurements at 21.

4.4.8 **Nick (highly mobile)**

Nick is originally from the Yorkshire Dales (Arncliffe), but by 21 has a hybrid standard accent, having been to Oxford University. He does retain obvious features of Northern and Yorkshire English, such as some examples of short /a/ in BATH set words, but these are now variable distributions. For example, throughout the samples presented here, Nick has established a full STRUT/FOOT split, and there are fewer and fewer examples of BATH words being produced with /a/ (Beal, 2004) as the series progresses. The more interesting mobility-related changes from Yorkshire to SSBE occur for Nick between 7 and 28, and are outlined (by auditory means) in Sankoff (2004). As this is principally an acoustic study, unfortunately it is limited to adult recordings, and so mobility in this case largely relates to differences between more standard-type English (tempered with some Yorkshire features) and General American type accents. By 28, Nick has moved to America (Madison, Wisconsin) where he stays until 49.

4.4.8.1 F0

Figure 60 - Mean F0 for Nick at each 7 year interval with SD bars



F0 for Nick is relatively stable, except for age 42, where there is a 10Hz increase. Overall, F0 is relatively low for Nick, around 90-95Hz, but with fairly high SD around 25Hz. There is a slight decrease between 21 and 49, but as with most males speakers no consistent and clear reduction in F0 as predicted in previous research.

4.4.8.2 Formants

Figure 61 - F1, F2 and F3 for Nick at each 7 year interval

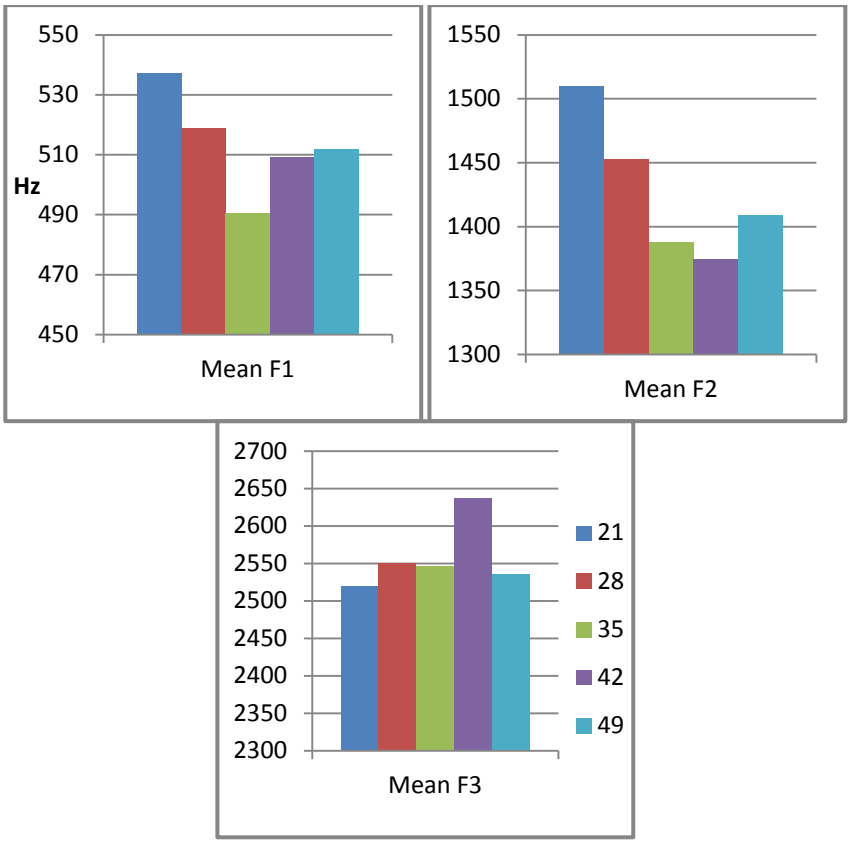


Figure 61 shows that Nick's formants somewhat follow the predicted pattern of reduction. For F1, there is in general a decrease over the three decades; however, there is either a low trough at 35, or a systematic increase from 35 to 42-49. This could be the effect of a particularly low F1 in 35, or changes to the vowel system as part of a transition to a more American type accent. The increase later in F1 is mainly carried by increases in /a/ and /e/ at age 42 and 49, which show strongly within the mean as they have highest F1. Only /ɑ:/ decreased significantly for F1, while /ɪ/ also showed a significant effect of age, but not within a clear increase or decrease. For F2, there is a fairly stable and consistent reduction of around 100Hz across the period. In terms of vowels, Nick conforms to the general pattern of aging with significant F2 reductions to /ɑ:/ and highly significant reductions in /a, ʌ & e/, those open vowels where we are more likely to see a significant change. F2 of /i:/ showed a significant increase and this is clear from Figure 62 below. F3 remains relatively stable, except for a peak at 42 of around 85Hz above the rest of the values. This is not particularly marked or different in any of the vowels.

Figure 62 - Scatterplot of mean monophthong values at each interval for Nick, connected to represent vowel space. Histogram of vowel space area at each stage

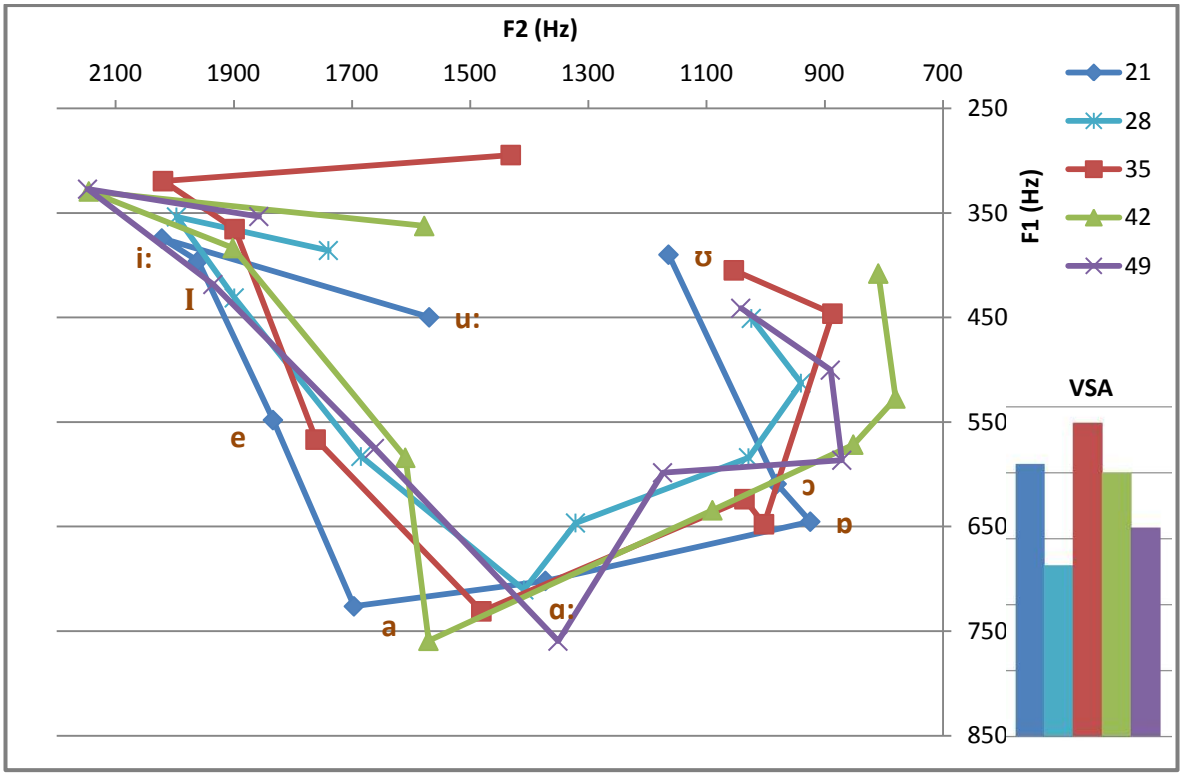


Figure 62 demonstrates these formant changes, with vowel spaces becoming less advanced in F2 at each stage (if we ignore increased /i:/ F2). Increases to F2 of /i:/ and /u:/ are more likely the result of widespread changes in SSBE than a shift to American accents as they occur before Nick moves abroad, and are consistent with changes in other speakers over this time period and with Hawkins and Midgley (2005). F1 is also decreasing over the first three periods, and generally within the close front vowels, as we would expect. However, there are increases in F1 of /a/ and /e/, the former being a surprising change given that we might expect Nick to accommodate to a more close production (Wells, 1982) as he settles in America. It is unclear what the reason for these increases is, although both vowels conform to decreases in F2. VSAe for Nick follow the general pattern, early variability with reductions between 35 and 49 years.

There are a number of changes that might be predicted from differences between Nick's English accent and a General American accent. Indeed, in accommodation settings he starts to use a number of American consonantal features (Windsor-Lewis, 1971; Wells, 1982), including intervocalic t-tapping and also yod-dropping post-/t/ for example in 'stupid' (although this is in teaching to an American classroom). His lexical choices and intonation also reflect an American language context, with phrases such as 'bunch-a stuff' and primary stress in the word 'research' for example. In spite of this, there is little

evidence of systematic changes to the vowel system to appropriate a General American type accent; in general changes are in line with aging patterns with few counter examples. For example, BATH words are not produced with an /æ/-like quality and neither are LOT vowels becoming more open, towards /ɑ:/ (Windsor-Lewis, 1971). What Nick does exhibit quite noticeably are individual examples of phonologically different vowel qualities (usually in accommodation type contexts) or structural differences in consonant realisations that are more obvious markers of an American type accent than systematic vowel changes. This is presumably because these changes are much more obvious markers (or indeed stereotypes (1972)) of the accent to the speaker and listeners, and also are easier to implement (for the speaker) and recognise (for listeners). Form auditory impression, Nick is by no means an American sounding speaker, but these stereotype changes in consonants and a few vowels must have sociolinguistic benefit to Nick, especially given that his profession involves lecturing to American students.

4.5 **Speakers summary**

As we have seen from overall results, F0 does not uniformly reduce in all cases, especially for male speakers. This is surprising given the prediction that changes would be more prevalent in males. F1 reductions are the most consistent finding; it was summarised in §4.1.2 that the magnitude of F1 changes was greatest, and this is also shown in this section to be very consistently realised across speakers. Furthermore, the effect also appears to be strongest in close-front vowels for most of these subjects. F2 also decreases consistently, but to a lesser extent. Individual speaker data show us that this overall difference may also be tempered by limited changes in close front vowels, where a widespread pattern in Standard English varieties is for F2 increases, which may offset likely age related decreases. Like F1, there are counter-examples where changes to certain vowel qualities occur due to widespread changes in a speaker's accent or assimilation of a new accent type. This seems to be more comprehensive when speakers assume a standard type accent, whether this is for personal, social or geographical reasons; for example in the cases of Neil and Nick, and in a different sense, Suzy. F3 is more stable in general, and this might be due to inflexibility of those tracts which are acoustically reflected by F3. It is very likely that F3 shows limited changes as it is not as directly associated with movement of the articulators in producing different phonetic

qualities. For example, a number of speakers show fronting of /u:/ and other close front vowels, which affects overall F2; these kinds of changes would not affect F3 as strongly.

In summary, when we consider individual cases it seems that mobile speakers appropriate a standard variety of English much more readily than a 'host' accent type (in cases of geographical mobility). Although there is some previous literature regarding dialect acquisition, there is little mention of acquisition of a standard variety and the benefits for a speaker that this may hold (this may be as they are focussed on microscopic phonological changes or singular segment changes). It also seems true that speakers more readily appropriate a standard accent in younger life (i.e. before 21-28 years), a finding that is widely corroborated in the literature around mobility and age, cf. Chambers (1988; 1992). Of course it is much too simplistic to assume that speakers simply adopt different accents, and there are of course spectra of styles and assimilation, and gradual processes of adjustment (Nycz, 2011).

In Nick's example, a standard accent is generally retained despite relocating to America. Although in some of the recordings Nick uses stereotypical (Labov, 1972) consonantal and lexical-grammatical features to accommodate whilst teaching American classes and interlocuting with his American wife, in general his phonological system and phonetic realisations remain quite stable. For Neil, assuming a standard accent is much more important than assuming features of his current region, and it seems likely this is closely tied in with the linguistic capital (Bourdieu, 1977; 1991) this variety holds. This is probably exacerbated by Neil's political ambitions, where this accent is highly desirable, especially historically. His conception of a standard accent seems to resonate with Agha's (2003, p. 231) description of standard type English, as: "a status emblem in British society, an emblem of speaker status linked to a specific scheme of cultural values". For Suzy, assuming a standard accent means conforming to more widespread changes, much like the Queen's English in Harrington et al. (2000b). It seems that geographical or social mobility gives speakers the opportunity to change, but there still has to be a benefit or a reason for that change. Presumably assuming a standard (or more widespread) accent is more beneficial for these speakers. It is also important to remember though, that these speakers were interviewed by an RP speaker for a national British television broadcast; it is hard to know what effect if any this had on speaking style. This was, however, a constant throughout all recordings. Linguists (Trudgill, 1989; Johnstone, Andrus, &

Danielson, 2006) argue that mobility (social and geographical) makes speakers more aware of accent differences, and thus more aware of the social 'work' that their language performs. In this case, interviews with the subjects in the *Up* programmes suggest that they are very aware of the social functions of their language, regardless of individual mobility. For Tony, although he is fully aware of the impact of his style of speaking, he retains the desire to express his cockney identity. This variety holds covert prestige (Trudgill, 1972) for him personally.

In general, there are few changes across all the speakers which do not follow the predicted pattern established by aging research and set out in §2.1.5. Especially in those cases where there is little to no evidence of mobility which might be likely to cause a change in accent, the data follow predictions set out for this study very consistently. There are a few specific counter-examples, predominantly from speakers who display one or more kinds of mobility. In most cases we can generate hypotheses about these speakers' linguistic behaviour from the information that is made available from the documentary. Although this is a small sample, these changes are individually variable, and most likely motivated by the personal views and social context of the speaker (and of the speech act). Across all speakers there are widespread and consistent reductions in formant frequencies, and to a lesser extent, reductions of F0.

4.6 **Discussion**

Returning to the research focus of this study, these findings are relevant to questions 1 and 2.

4.6.1 **Research question 1**

1 What is the extent of vocal instability in adulthood and how will this affect the interpretation of results from apparent-time studies?

It is useful to separate 'vocal instability' into two types of change that have been predicted by physiological and sociolinguistic research: firstly there is instability in terms of predictable age-related changes that can be attributed to physiological causes. Secondly there are accent changes whose causes are not physiological, such as those found in speakers who appropriate a different accent type or adjust to mainstream changes in their own accent.

In terms of physiological age-related changes, an AT method assumes that speakers are, in all significant respects, invariant in adulthood. This study shows that speakers are not, in fact, inflexible in terms of widespread changes to frequency output and specific social or personally motivated changes in certain speech features. If we know there are age-related changes in certain acoustic parameters we can also make an assumption (from past and present data) that the population roughly reflects those frequency characteristics, and therefore speakers from different ages naturally produce (on average) different frequencies. If we expand this to an example AT sampling method, imagine an older group, say 50 year olds, exhibit on average significantly lower F1 frequency (perhaps by 8.5%) than a younger group of 20 year olds. This might lead to the conclusion that over the last 30 years a process of F1 lowering (or vowel raising) has occurred. In fact this would be simply a process of aging (not sound change in progress and not age-grading). Results from this study should lead AT studies to view their results with caution, especially if they are in the same direction of change as aging processes. Of course, this largely applies to phonetic-acoustic studies, although these significant differences in realisation and acoustic output might also affect auditory judgements of vowel quality.

A second issue comes with respect to the second type of non-aging related changes detected in this study, i.e. those without a physiological explanation. It may seem that these factors (i.e. mobility) are not an issue in sociophonetics as these kinds of speakers are routinely excluded from studies into a language variety. This is true for those speakers who may have moved or for those, like Neil, who have conformed to a more standard variety. Sociolinguists' focus on traditional non-mobile groups means that their overall samples are not representative of the speech community, which is particularly problematic for the UK, where people are highly and increasingly mobile. This is worrying if analysts are using sociophonetic studies to make informed judgements about phonetic and acoustic features and their distribution in certain speech populations within FSC casework. These samples are not representative of a potential perpetrator population and will likely be more homogenous than the real group. In very simplistic terms a more homogenous reference population would exaggerate the magnitude of an LR. Moreover, if these subjects are ignored, analysts are learning little about the linguistic consequences of these processes, which may be vital for interpreting forensic cases where these factors are apparent.

Even within this small subject set, it is clear that while physiological modulations and changes following mainstream accent trends occurred throughout early adult life, speakers seemed to resist accent changes due to geographical mobility after age 21 (at least in those vowel features presented here). Those changes that were not predicted by a physiological model were highly individual, and although they can be explained after the fact by social and lifestyle factors, it would be hard to make accurate predictions about these acoustic changes.

4.6.2 **Research question 2**

2 What is the magnitude of change in individuals' vocal output during adulthood?

This question is answered in relation to formants of monophthongs within the data presented in §0. The sub-questions are addressed below.

4.6.2.1 Research question 2a

a Which features remain more stable than others?

In response to question 2a, it has already been summarised in §4.3 that there are clear differences in the stability of certain parameters across a delay of this type (between 21 and 49 years). F3 remained the most stable across all speakers. F2 was not quite as stable as F3, with greater magnitude reductions, but both changed significantly in an almost equal number of cases. Within F2, open vowels reduced more than closer vowels. F1 was the most significantly affected in both number of significant tests and also magnitude of change. This resonates with a number of studies that have shown F1 to be heavily affected by a number of situational and channel effects, which have concluded that F1 is not a reliable parameter for FSC in those cases (cf. §1.1.3). Although fundamental frequency was less affected in male subjects across the entire period than formants, it was variable within speakers and has been shown in other studies to be an unreliable parameter.

Two predictions from previous research were borne out by exploratory estimations using the formant data. Speakers were shown to be generally using smaller habitual vowel spaces from age 35 onwards in this sample. Estimations of vocal tract length also indicated that the vocal tract extension hypothesis might be another likely explanation for age-related acoustic changes. Although these were just estimation formulae, they

indicate two processes which ran on different timescales, both potentially affecting measurable speech properties.

4.6.2.2 Research question 2b

b To what extent are changes predictable from a model of sociolinguistics or gero-physiology?

In terms of this question, a large majority of the data met predictions that were made from physiology and also sociolinguistics. Reported changes with physiological aging are very general and do not assess individual vowel types, making comparisons difficult. Formant frequencies are not stable across adulthood, in the majority of cases there was a predictable pattern of reductions between 21 and 49 years. For the first formant, for example, there were similar reductions in speakers (mean 8.5%) as well as predictable changes in certain types of vowels (close front showing greater and more significant reductions). F0 did reduce, but it was more marked in female speakers, contradicting the predicted pattern (however, this is a small dataset and one female subject smoked throughout). VSAe shows that, as previous research hypothesises, speakers' habitual vowel spaces decrease in overall area, and generally displace toward zero due to overall formant reductions.

In terms of sociophonetic predictions, these were more suitable applied as post-hoc hypotheses for certain counter-examples. It would be very difficult to tease apart sociophonetic variation and age-related changes when they would lead to differences in a similar direction (i.e. aging and vowel lowering). In casework a sensible approach would have to be taken to assess which data are reliable based on the knowledge of the speaker's situation and the effects this may have on their speech. It may be sensible, in some cases, to avoid sociolinguistically sensitive variables in these analyses.

4.6.2.3 Research question 2c

c What effect should this have on how we evaluate forensic speech evidence?

It is important to consider question 2c carefully in forensic cases where there is a clear delay between evidential and suspect samples. The broadest recommendation is to make predictions on the likely changes based on reliable information about the speaker. This may include background information about origin and a social or geographical history.

In general, given mean 8.5% reductions in F1, it would seem wise to avoid using F1 measures as speaker-characterising unless it could be shown that there was an expected decrease between young and older samples. This follows a consensus of research into the reliability of F1, and is the recommendation of studies with similar F1 findings in predictable directions (e.g. for the telephone effect (Künzel, 2001)). It would also be wise to exercise caution in using F2 and F3 measurements, as they show significant changes across this three decade period. F3 does, however, provide the most stable formant within this study, which complements findings in other studies that the higher formants are better speaker discriminators (Nolan, 1983; Künzel, 1995; Loakes, 2004).

5 Acoustic analysis of diphthongs

This chapter presents acoustic analyses of frequency and slope of diphthongs, including a discussion of changes in light of dynamic approaches to capturing speaker identifying information based on McDougall (2005; 2006). Strength of evidence estimates for polynomial coefficients from diphthong data are presented in chapter 6, but before these are investigated it is useful to provide a descriptive account of changes in underlying frequency and extent of gestures. This chapter addresses some portions of research question 3, laid out in §2.2.3.

5.1 Diphthong data

This section analyses the effects of age on formant frequencies of diphthongs and compare these findings with monophthong results from the previous section. The method for these analyses are presented in §3.4.2.4. One concern is that the number of appropriate tokens for any diphthong was low and very variable between samples: only /aɪ/ and /eɪ/ had sufficient tokens to allow for analysis, and for some speakers or age stages even this yielded few tokens. Moreover, token frequencies were to an extent exaggerated by or limited to certain contexts by the subject matter. As participants spend a lot of time talking about themselves and their lives recordings yield a large number of 'I' and 'life/lives' tokens. For some participants, most contexts came from habitual use of the filler 'like' (there is evidence this can differs from lexical 'like' and that these differences can be socially conditioned (Drager, 2006). Generally speakers had more /aɪ/ or PRICE tokens than /eɪ/ or FACE vowels. A number of recordings (for Andrew, Nick, Lynn and Bruce for/eɪ/) did not have sufficient tokens to be analysed at all, and most analyses in this section had very unequal sample sizes (cf. Table 30 below).

Table 30 - Token numbers for the subjects included in diphthong analysis

Speaker	Vowel	21	28	35	42	49
Bruce	**aɪ**	18	-	7	4	8
Neil	**aɪ**	19	17	16	18	32
	eɪ	22	9	8	5	13
Suzy	**aɪ**	11	10	16	25	14
	eɪ	3	9	9	9	6
Symon	**aɪ**	8	7	-	5	9
	eɪ	9	4	-	4	5
Tony	**aɪ**	7	15	4	0	12
	eɪ	18	13	5	7	6

In terms of forensic applications, it may be unlikely that a forensic recording features sufficient data for diphthongs to provide meaningful acoustic and numerical analysis (recordings in this study were relatively long compared with the forensic context). Certainly the distribution of tokens suggests that any diphthong-based measure would not satisfy Nolan's (1983) fourth criterion of *availability*. Although this section is able to provide insight into formant behaviour, the sample sizes limit the strength of these conclusions. Furthermore the data violate the necessary assumptions for inclusion in both ANOVA (assumption of sphericity) and MANOVA (assumption of homogeneity of covariances) tests. This means, especially given the small and sometimes unequal sample sizes, that these tests cannot properly assess significance for the data. Therefore this section presents descriptive statistics and discusses the results of this statistical testing in an exploratory way (i.e. with very little statistical power).

This study provides a first-step analysis of spontaneous diphthong data from varying phonetic contexts which contributes to the recent literature on formant dynamic measures of non-elicited material (Atkinson, 2009; Hughes, McDougall, & Foulkes, 2009; Rhodes, 2009). Although it should be noted that not all measures in this section are 'dynamic', descriptive data in this chapter underpins understanding of aging effects on diphthongs. This can then be used to contribute to explanations of the performance of FD measures.

5.1.1 **/aɪ/**

For most speakers, the PRICE vowel is produced with a similar phonetic realisation to SSBE (Wells, 1982), with the first target an open (somewhat) front vowel (for most speakers, realisations of /a/ and /ɑ/ are discriminated by height as well as frontness). For Tony, a Cockney English speaker, the PRICE set is realised with [ɑɪ] or even [ɒɪ], which is sometimes found in Cockney (Wells, 1982). Both pronunciations move from this open vowel to a front close KIT type vowel. This close-front movement, or off-glide, is marked by a decrease in F1 and a steep increase in F2. For both diphthongs, there was no steady-state second target of the vowel in most tokens. The term glide is better suited for this movement (rather than a steady target).

5.1.2 **/eɪ/**

Production is similar for the FACE set, where most speakers have a SSBE type realisation (initial [e]), but Tony's production is fairly typical of a Cockney English accent: a more open and less front initial vowel. Wells' (1982) description for Cockney English FACE is /ʌɪ/, acoustically Tony's nuclear target not is as back as this, it is slightly more open and less front than /e/, usually somewhere around [a], [ɛ] and schwa. Similarly to the PRICE vowel, the second portion of this diphthong is best described as an off-glide, with no steady state.

5.2 **Results**

This is the first study that has examined the effects of aging on formant transitions (using formant transition data) for this age range for speakers other than the Queen (Harrington, Palethorpe, & Watson, 2005). These sections examine age-related changes in frequencies of diphthongs, responding to predictions from monophthong formant data in the previous section and previous research, and address the overall research question:

> What are the effects of aging on [diphthong] formant transitions?

Methodology for these analyses is presented in §3.4.2.4. Consonantal contexts were only controlled for in a limited way, using the same criteria as monophthong analyses (removing those which would have a large influence on the formant transitions), so there would feasibly be high variability in the initial and final parts of the diphthongs due to influences from articulator movements for consonant production (Clermont, 2007; 2009; 2011).

5.2.1 **Formant frequencies**

5.2.1.1 Predictions

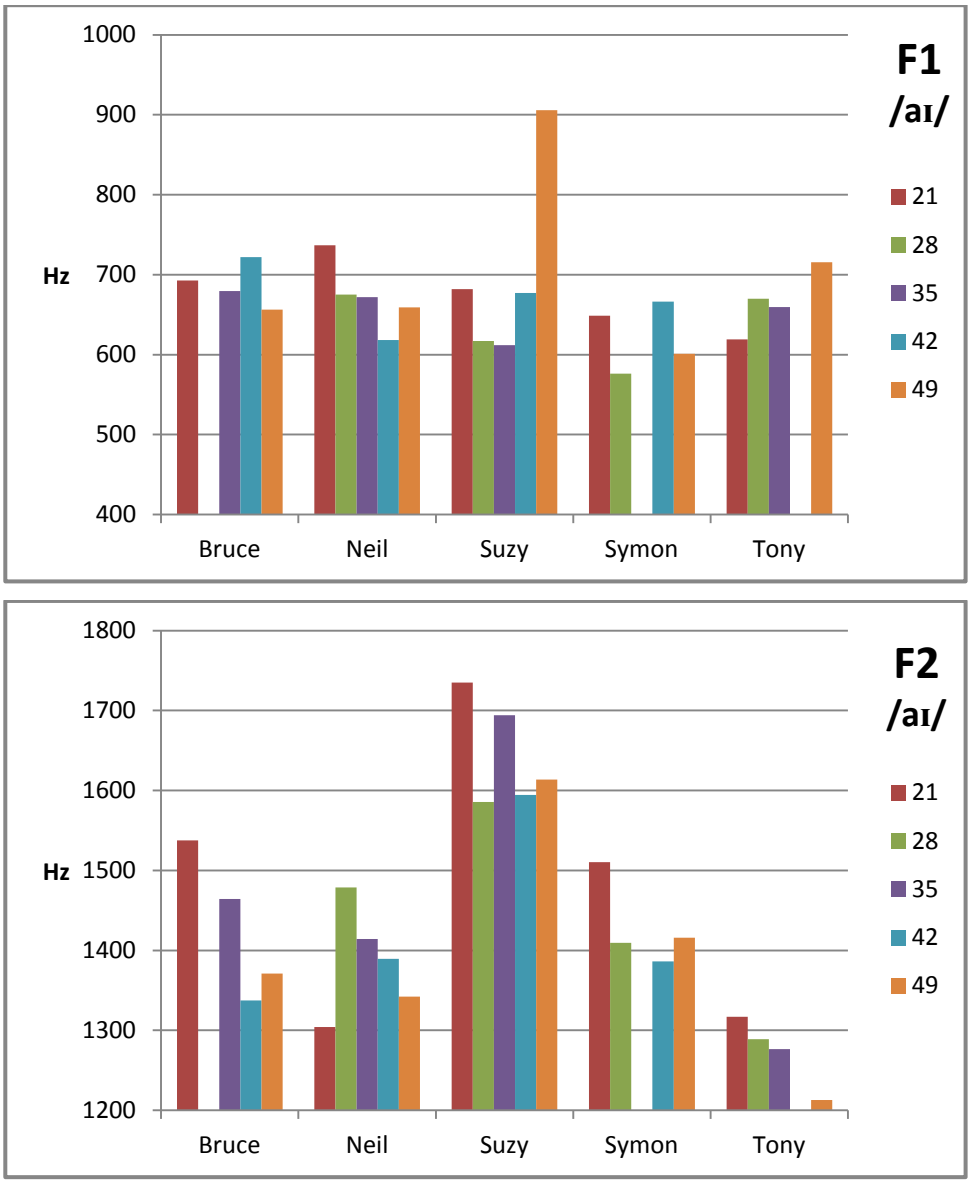
Monophthong data from the current study suggest that (unless sociolinguistic behaviour contradicts it) formant frequencies, in general, decrease. This corroborates previous research, summarised in §2.1.4.10. For monophthongs, this change differed by formant and by vowel type, and the extent was variable between speakers. F1 was most affected, and this was highest in close front vowels, for example. For diphthongs, we would expect, given the universality of hypothesised vocal tract lengthening and reduced articulator flexibility, formant frequencies to behave in the same way. Following the

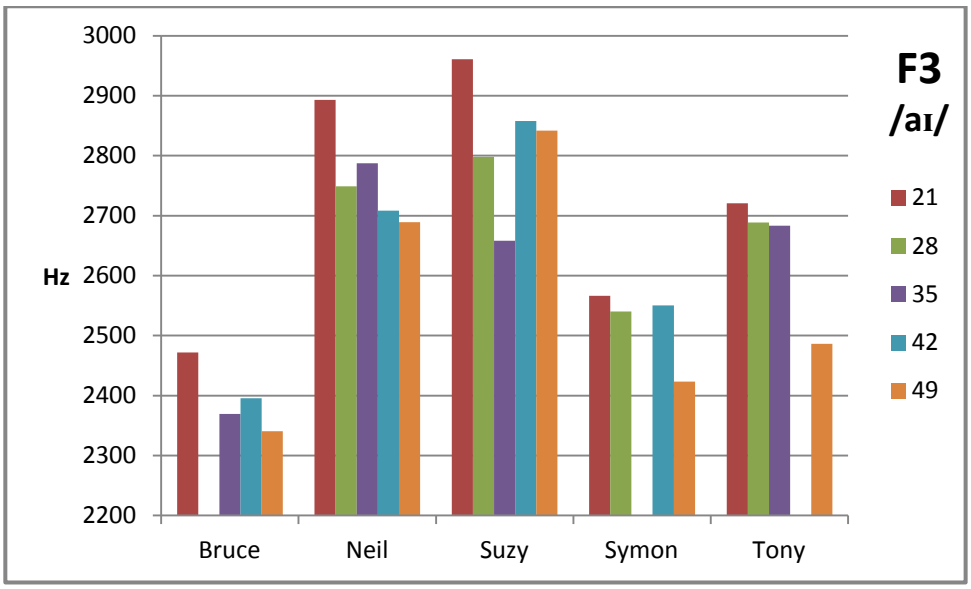
vowel-variable aging changes for monophthongs, there may be a pattern for the two components within the diphthong to change to exhibit differing extents of change.

5.2.1.2 Overall mean formant frequency - /aɪ/

Although it generalises across the transition data and across two separate phonological entities, an average across all intervals gives an idea of the overall behaviour of each formant, and reflects universal changes to the speech production process that are posited as the most likely cause of age-related frequency changes.

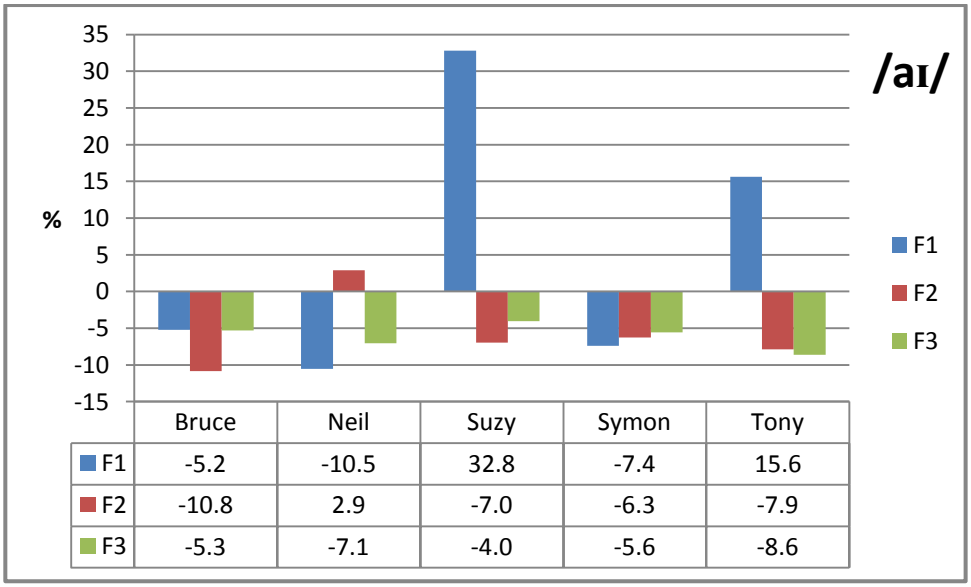
Figure 63 - Mean formant frequency across all intervals for tokens of /aɪ/ at each age stage





It is clear from Figure 63 that formant frequencies reduce in most cases. There are some speakers whose F1 actually increases by age 49 (quite significantly for Suzy), but in general, especially for F2 and F3 overall, formant frequencies reduce across the period. In Suzy's case, F1 is relatively stable until age 49 where there is a massive increase of nearly 250Hz. This pattern is also apparent for /a/ and /ɑ/ in the monophthong data, with evidence from changing trends in SSBE showing that Suzy might, at least for /a/, be conforming to wider trends (similarly to the Queen in Harrington et al.'s (2000b; 2005; 2007) studies). Data for Suzy's /eɪ/ in §5.2.1.3 do not show a similar pattern, nor do other monophthongs, suggesting this is a socially motivated change and not a universal/physiological one.

Figure 64 - Percentage formant frequency difference between 21 and 49 years for /aɪ/

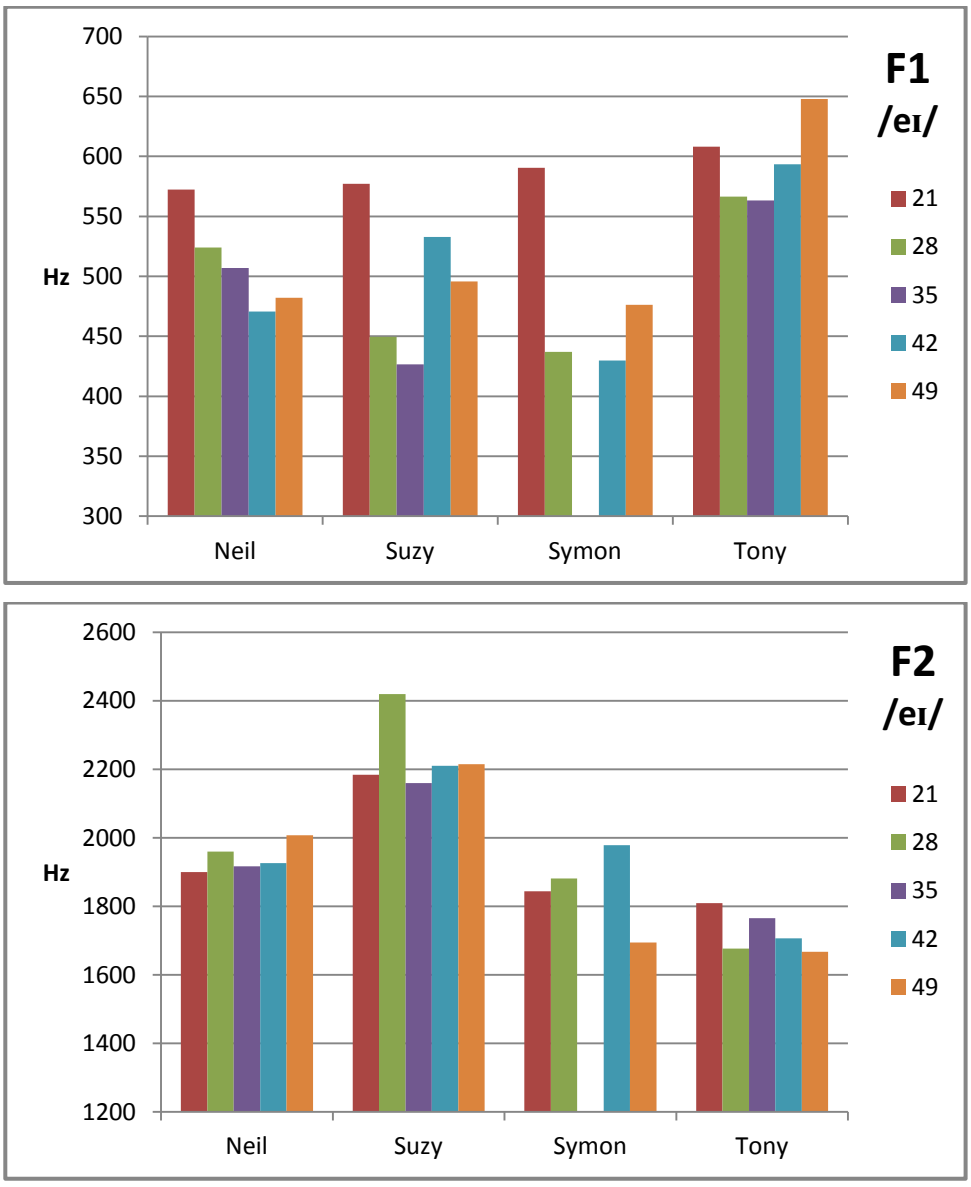


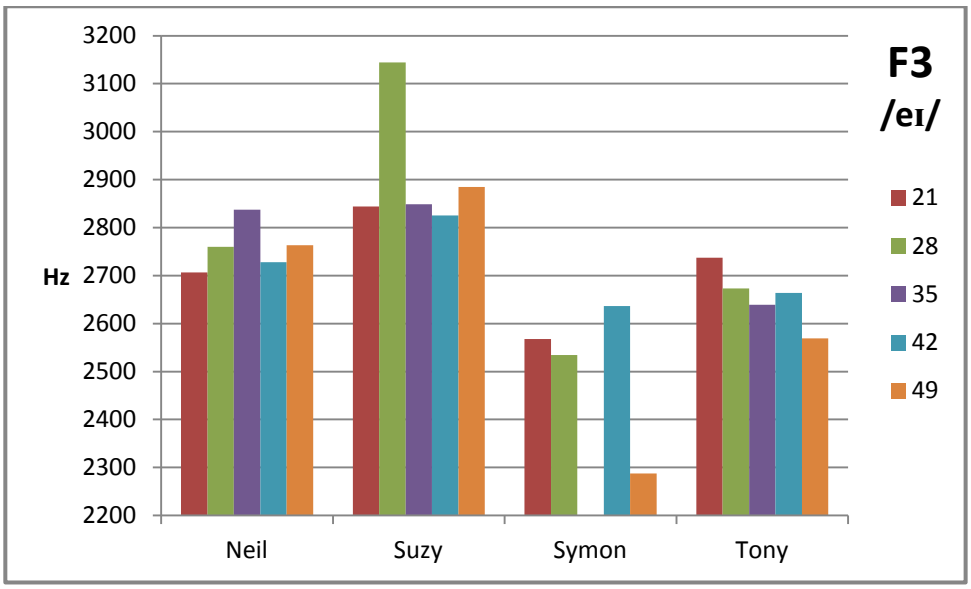
	Bruce	Neil	Suzy	Symon	Tony
F1	-5.2	-10.5	32.8	-7.4	15.6
F2	-10.8	2.9	-7.0	-6.3	-7.9
F3	-5.3	-7.1	-4.0	-5.6	-8.6

Figure 64 demonstrates the percentage change between ages 21 and 49 with respect to the formant means displayed in Figure 63. It illustrates that changes are largely negative, despite two extreme counter examples in F1 and one small increase for F2. Where there were decreases in F1, these were between 5-10% (mean 7.7%); F2 decreases were average 8% and F3 decreases were on average 6.1%. The diphthong findings are not as clear and the sample is not as large as for the monophthong data in the present study. Nevertheless, F1 changes are comparable with monophthong data, whilst F2 and F3 changes for /aɪ/ are more marked in F2 and F3 than for all monophthongs.

5.2.1.3 Overall mean formant frequency - /eɪ/

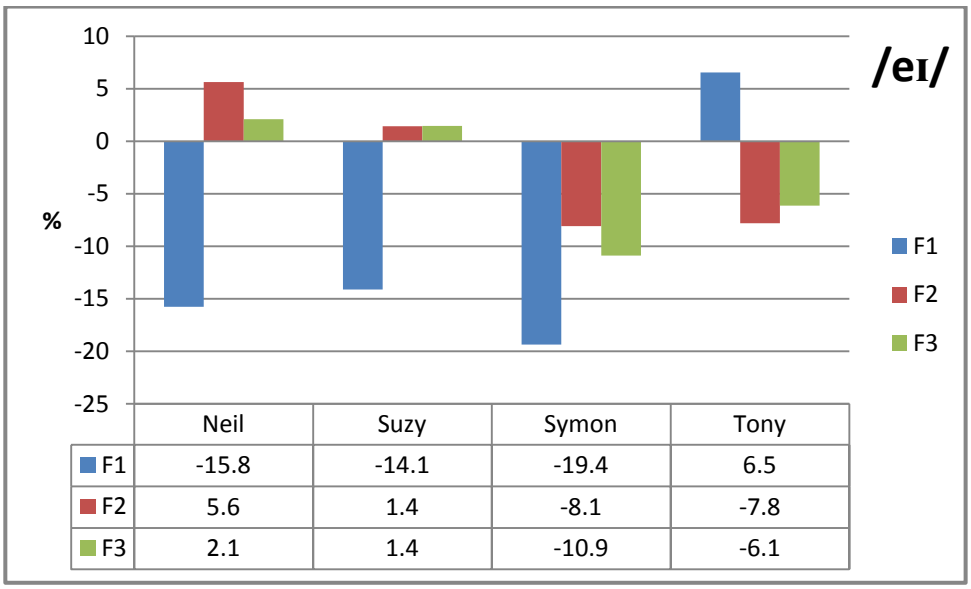
Figure 65 - Mean formant frequency across all intervals for tokens of /eɪ/ at each age stage





The data above are more variable than for /aɪ/, in most cases there are patterns of decrease in F1, but again Tony shows a steady increase. To an extent, this mirrors increases for F1 in Tony's mean monophthongs, suggesting a more universal explanation for this contrary pattern of increasing F1. Suggested reasons include overall changes in voice quality or articulatory setting. Tony's F0 also increases post 35 and this could contribute to F1 increases. For F2 and F3, there are limited changes, two of four speakers (Symon and Tony) in each case show a marked decrease (over 5%) during the period, where Neil and Suzy are relatively stable.

Figure 66 - Percentage formant frequency difference between 21 and 49 years for /eɪ/



	Neil	Suzy	Symon	Tony
F1	-15.8	-14.1	-19.4	6.5
F2	5.6	1.4	-8.1	-7.8
F3	2.1	1.4	-10.9	-6.1

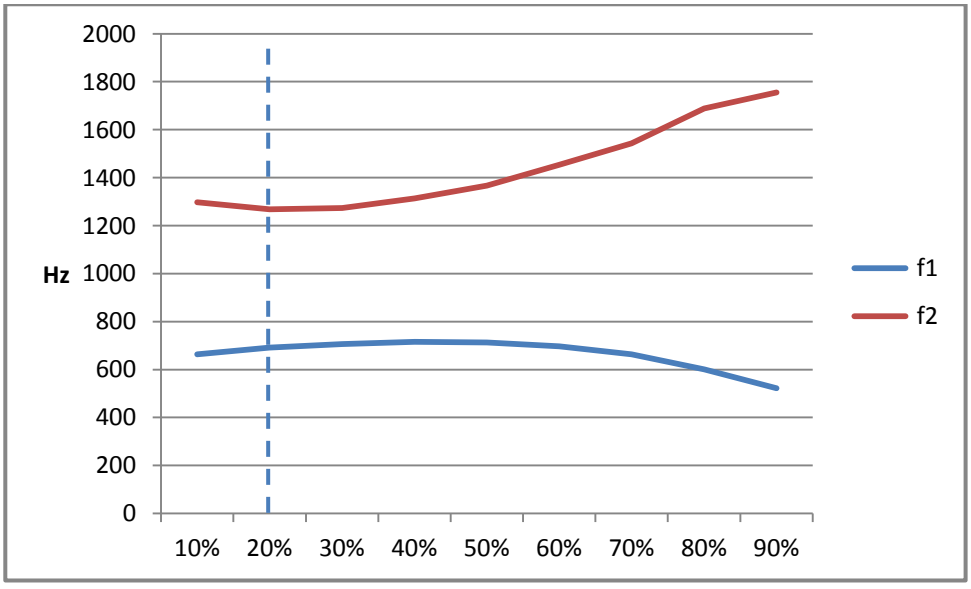
The difference between 21 and 49 here shows this more variable pattern. F1 seems to be changing more markedly (mean 16.4% across three speakers), but this actually reflects relatively high F1 for 21 years in most cases. Two speakers show decreases in F2 and F3

(of mean 7.9% and 8.5% respectively), whilst the other speakers have minimal increases, but appear to be fairly stable. Despite the more variable data, both diphthongs seem to exhibit the expected decreases that are present also in the monophthong data.

5.2.1.4 Frequency at first vowel target and glide

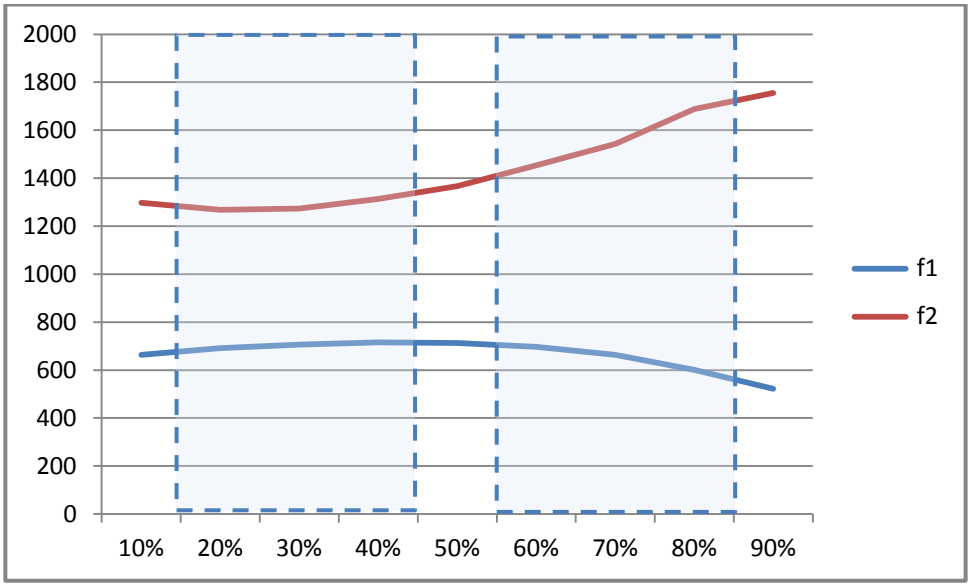
Given the difference in vowel behaviour in changes to F1 and F2, it would be sensible to observe any differing behaviour between the initial target and the glide. Previous studies have used 25 and 75% duration point measurements to encapsulate acoustic activity in a diphthong. However, these are not reliable in this case given the nucleus/glide structure for these diphthongs (much like in McDougall (2005)) and especially in spontaneous speech, where it is unlikely that the phonological targets are realised at these time points. In fact in most cases, the first vowel target portion was slightly later than 25% duration (see the figure below).

Figure 67 - Example of a typical diphthong F1 and F2 with 25% marked



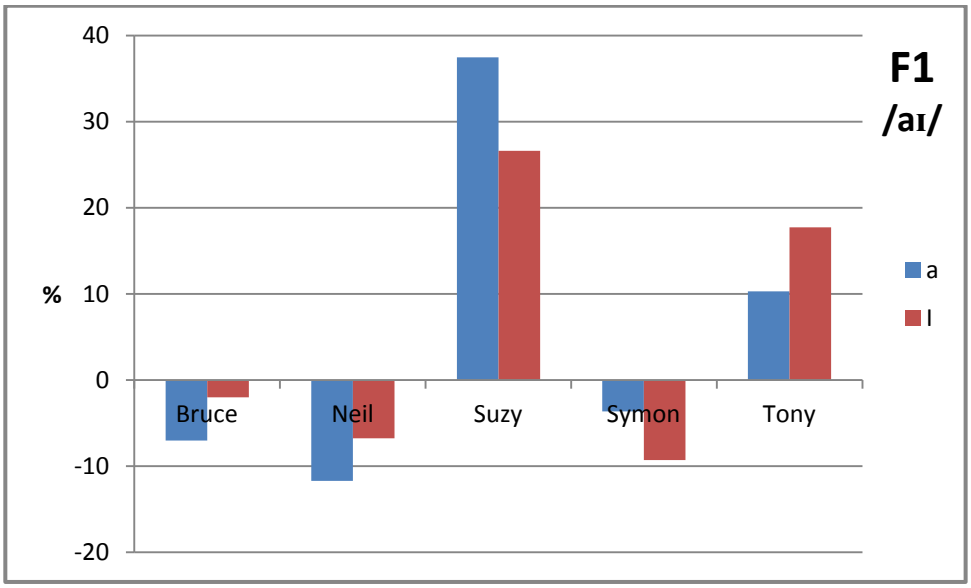
In the present data, 10% intervals were measured across the entire duration of the vowel. Two averages were taken from each token to capture frequency at the *target* and *glide* sections; these were three mean frequency measures at intervals 20%, 30% and 40% and at 60%, 70% and 80% respectively.

Figure 68 - Example diphthong with first and second part average areas marked



If changes behave similarly to long-term changes to monophthongs, we would expect close front vowels (i.e. the glide) to be more affected in F1 and open vowels (i.e. the target, particularly /a/) to show more change in F2. In reality, especially in these diphthongs, the realisation is only 'towards' a target, and certainly in these data, the glide portion is rarely acoustically front and close. Given that the glide rarely displays a steady or static state, this second portion was deemed a better way of representing an average value.

Figure 69 - Percentage difference between 21 and 49 years between mean frequency of *target* (a) and *glide* (ɪ) sections of /aɪ/



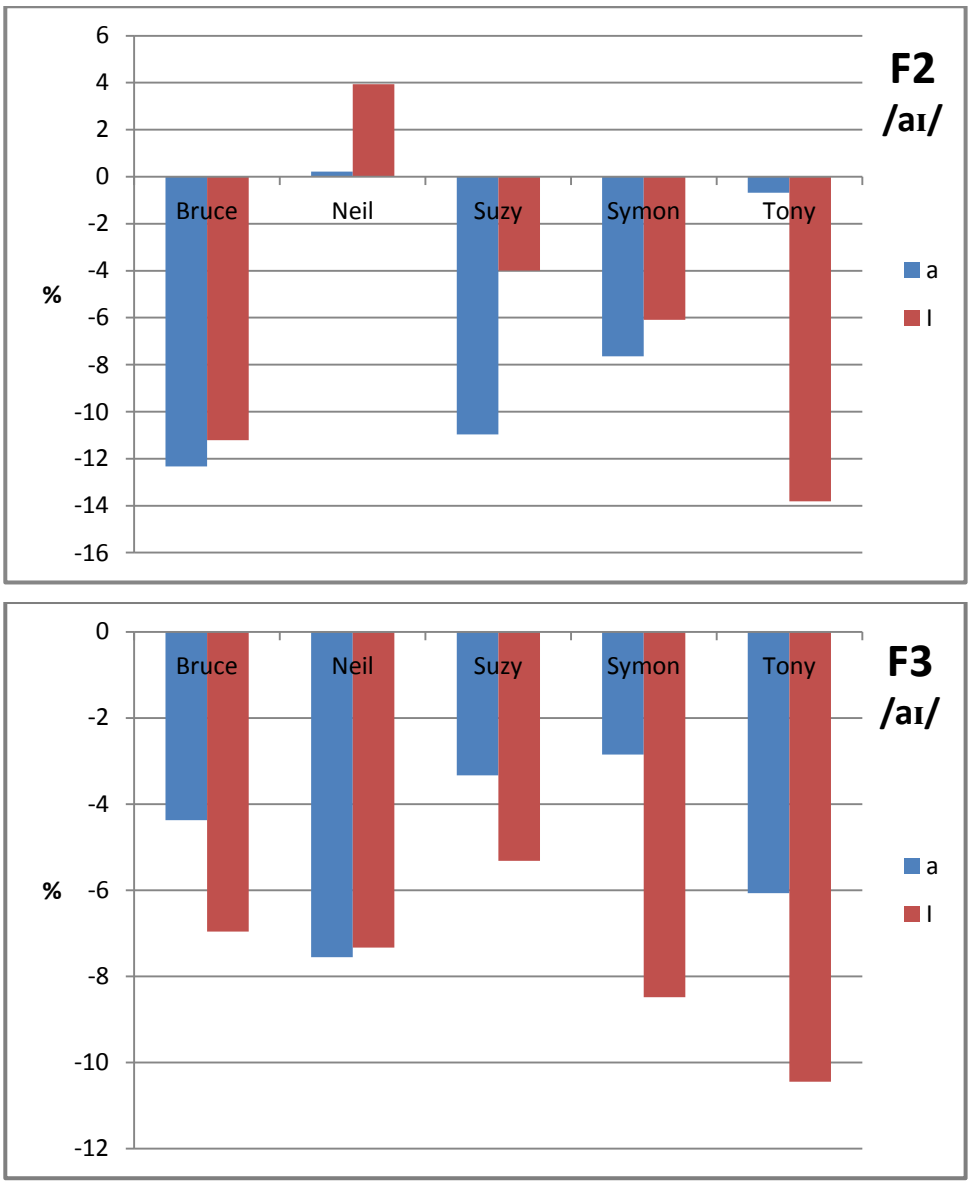
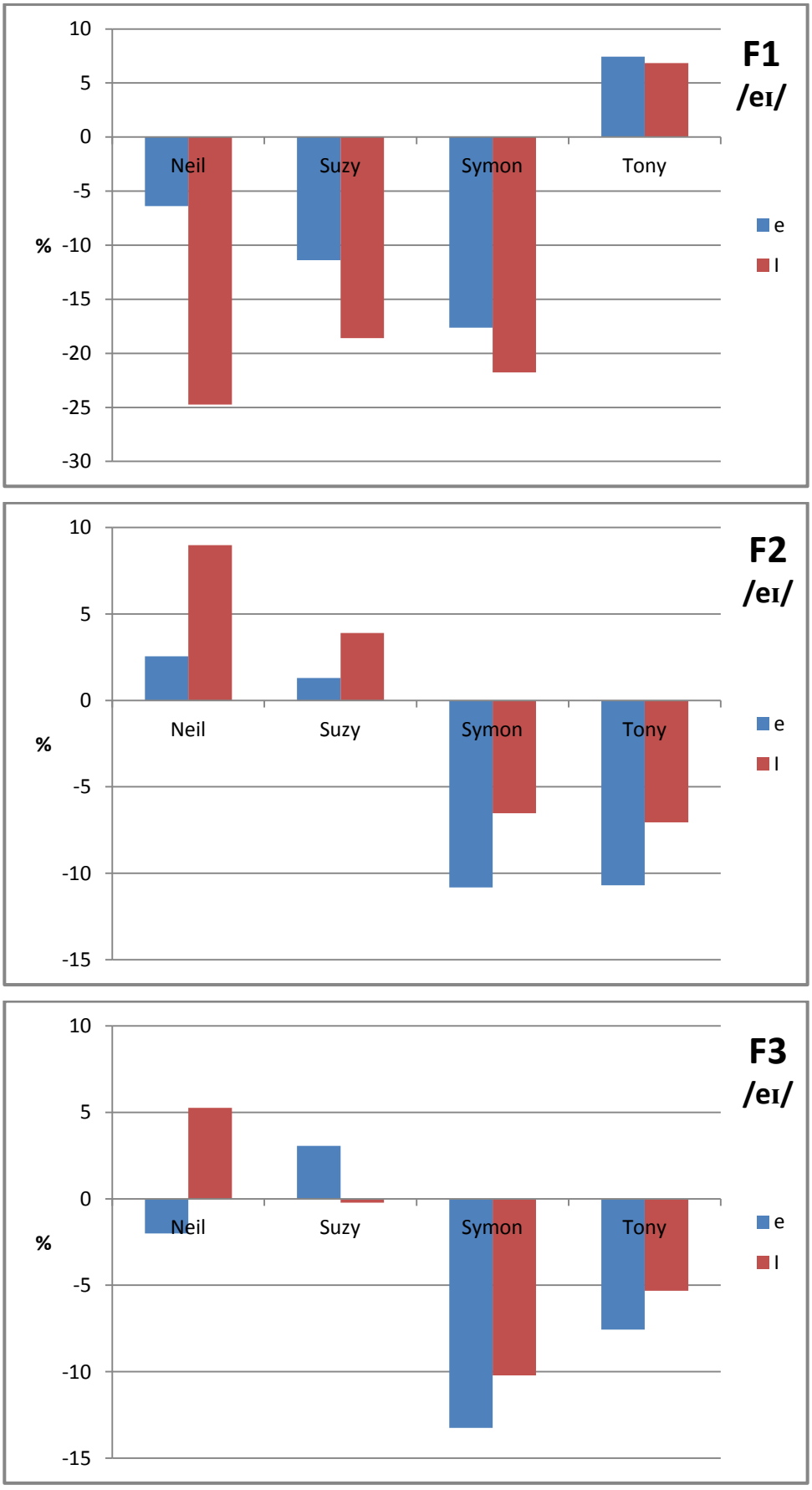


Figure 69 above demonstrates that there are consistent patterns between the measures at the *target* and *glide* sections of the diphthongs. In general F3 decreases were more marked between 21 and 49 for the *glide* section, and F2 changes were, for the most part, more marked in the *target* section. There is not a clear pattern for F1. This F2 trend concurs with monophthong findings for different vowels and could warrant further investigation. It is worth remembering the size of the dataset and more data would yield more generalisable results.

Figure 70 - Percentage difference between 21 and 49 years between mean frequency of *target* (e) and *glide (ɪ)* sections of /eɪ/



The picture is somewhat different for /eɪ/, given that decreases are less apparent overall. However, where there are decreases, the *target* and *glide* sections do behave slightly differently. Reductions in F1 are more marked in *glide* portions, whereas for F2 and F3, in the cases of decreases, these were larger in *target* sections. This follows the pattern of vowel behaviour observed in the monophthong data. Of course we must take into account the sample sizes and the fairly unsophisticated method used to estimate mean frequency for the sections, but it could be that there are links between these data and the monophthong patterns.

5.2.2 **Extent of gestural movements**

Previous research (Liss, Weismer, & Rosenbeck, 1990) shows that formant transitions become less extreme with extremely old age, due to reduced physiological flexibility and reduced motor-neurone function and coordination. These physiological processes are present to a lesser extent across all of adulthood, so we might expect limited reduction in the extremity of diphthong gestures. This was measured as 'slope' in the aforementioned study as the difference between maximal and minimal formant frequency. Evidence from the previous sections show that formant transitions are most likely displaced towards zero (i.e. overall formant frequencies decrease), but does age affect the steepness of transitions? Findings from the monophthong data set and VSAe measures in the present study might suggest, tentatively, that this process might be more prevalent after 35 years.

There are problems with slope as a measure, which are discussed later in this section. The figure below shows results based on this measure of range or 'slope', that is, the maximum formant value minus the minimum; limitations of 'slope' are discussed later, but as the formant curves are generally moving in one direction, they generally capture the steepness of frequency change (this would not be accurate for a sigmoidal curve, for instance).

Figure 71 - Mean Min-Max range in tokens of /aɪ/ at each age stage

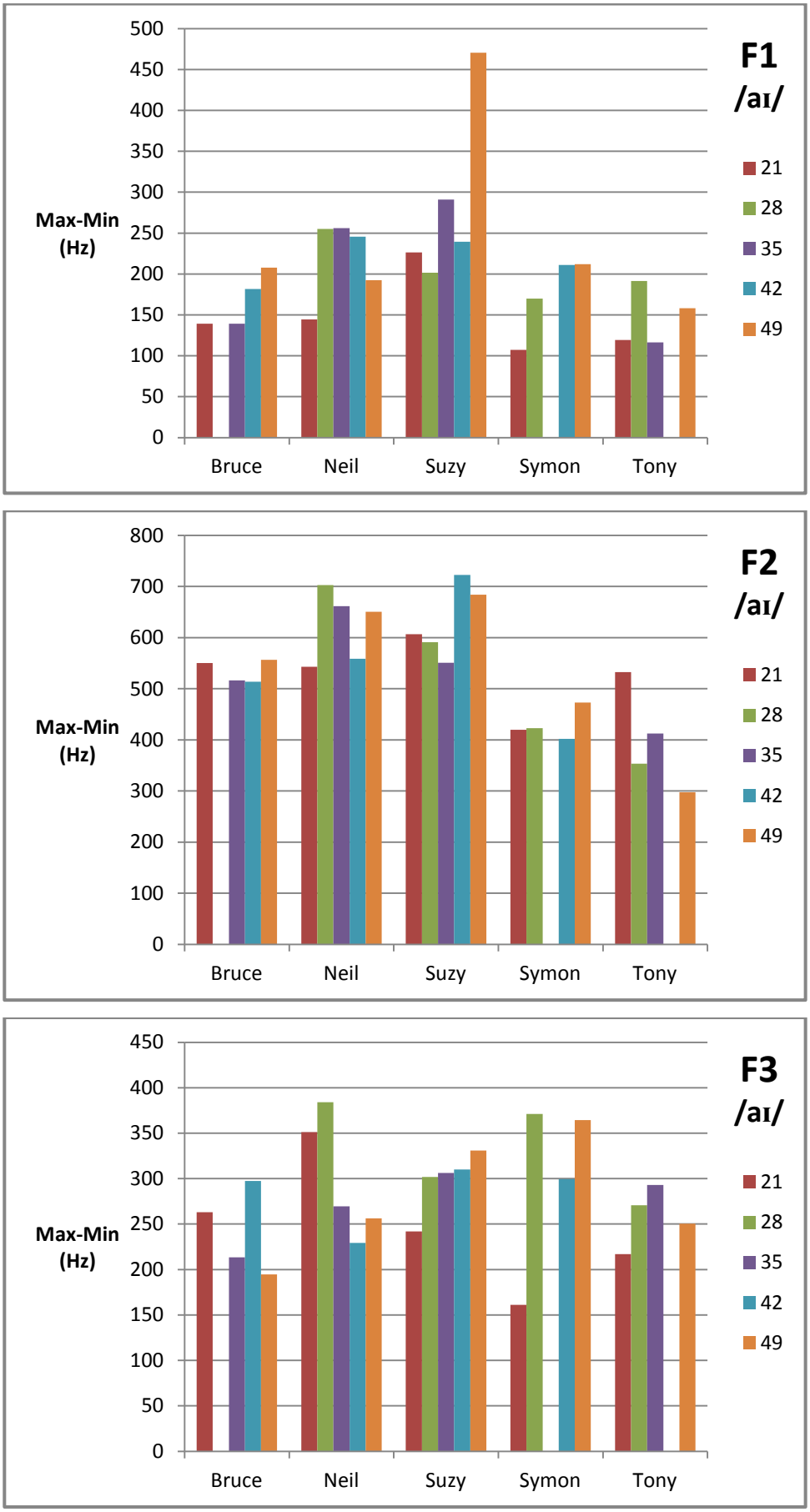
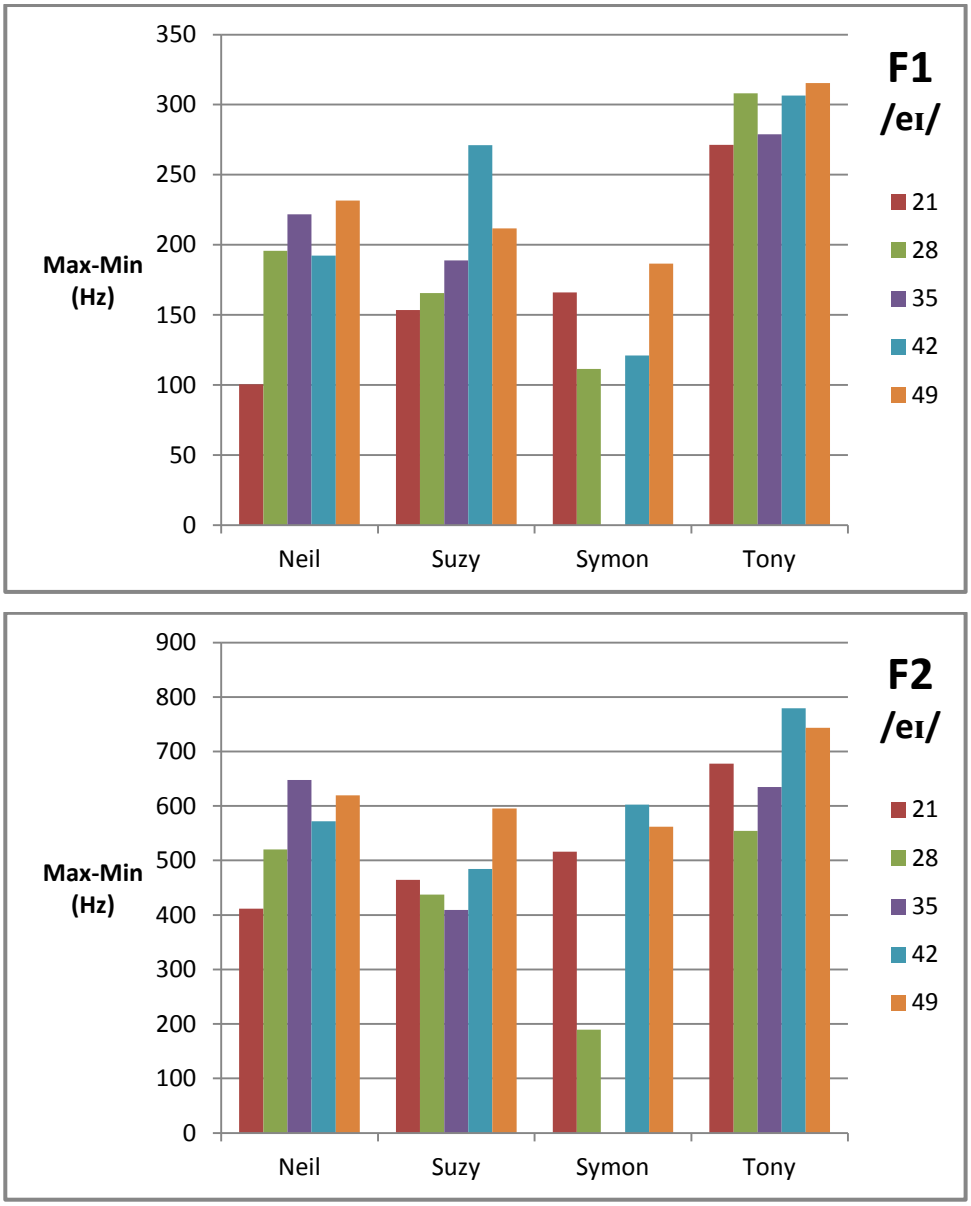
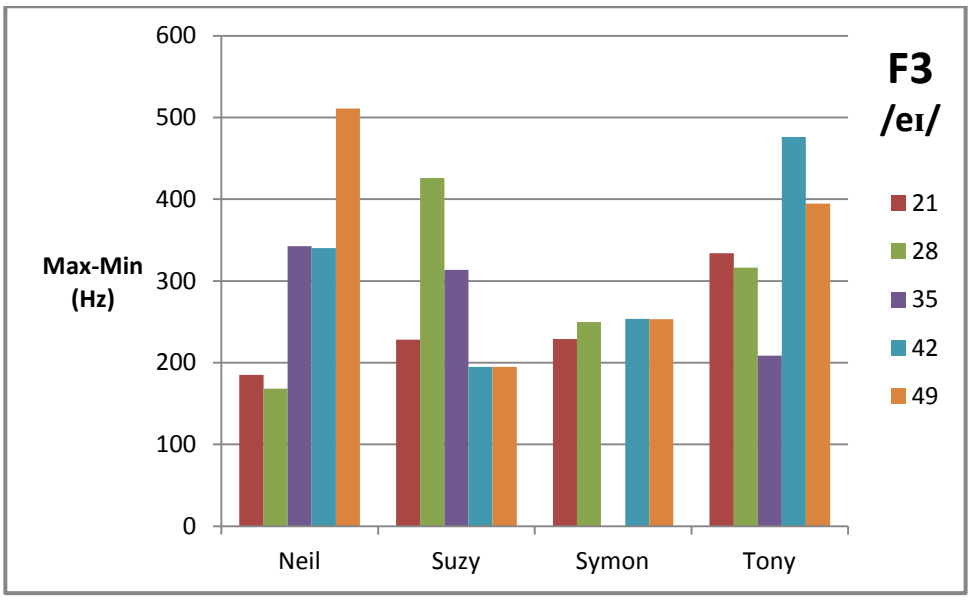


Figure 71 for /aɪ/ above demonstrates that for most speakers, min-max range does not decrease across the period, although there is variance between stages. Most speakers' range is actually increased quite significantly by 49.

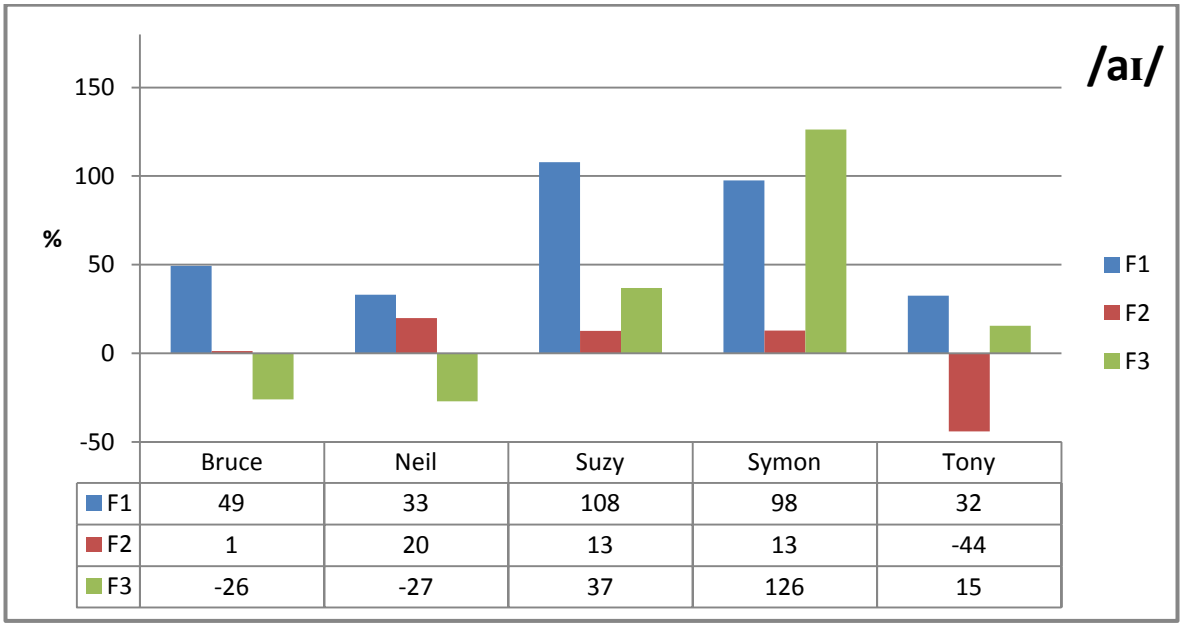
Figure 72 - Mean Min-Max range in tokens of /eɪ/ at each age stage



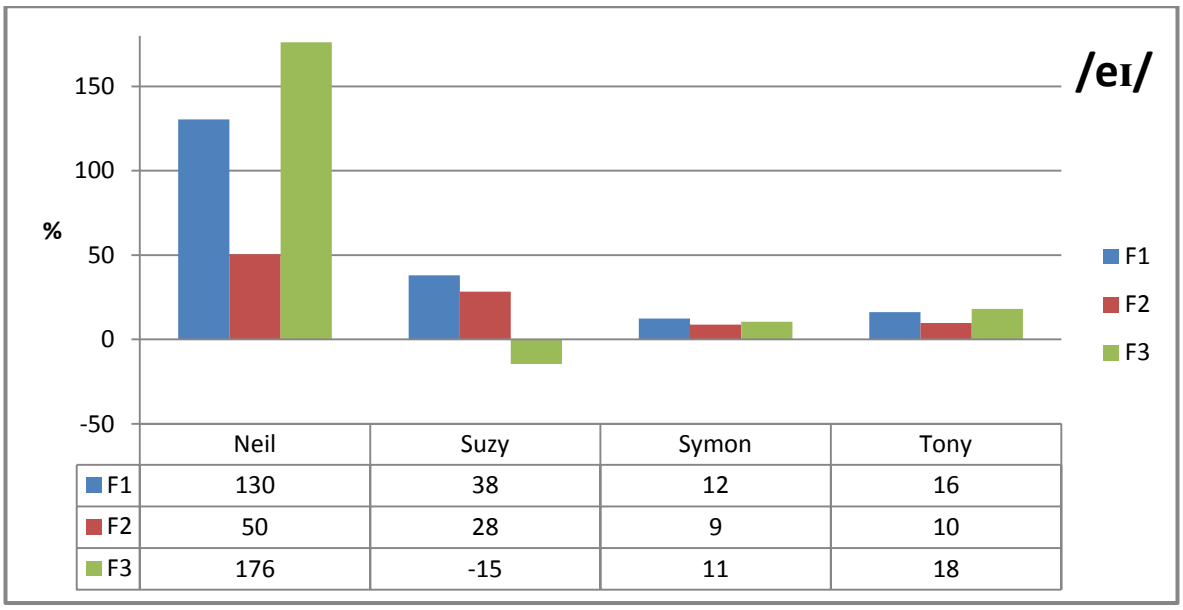


The findings are much clearer for /eɪ/, where all but one case in all formants exhibits an increase, most of them a fairly consistent pattern towards an increase in range. The overall change is perhaps easier to observe in the percentage difference between 21 and 49 years, displayed below:

Figure 73 - Percentage difference between 21 and 49 years for mean min-max range across diphthong tokens



	Bruce	Neil	Suzy	Symon	Tony
F1	49	33	108	98	32
F2	1	20	13	13	-44
F3	-26	-27	37	126	15



	Neil	Suzy	Symon	Tony
F1	130	38	12	16
F2	50	28	9	10
F3	176	-15	11	18

It is clear that for both diphthongs, there are generally increases in range or 'slope' of formant transitions (in all cases for F1, 8/9 for F2 and 6/9 for F3). These increases are generally larger in F1, especially for PRICE vowels. In fact, changes were larger for nearly all increases in range (though we might expect large ranges for Suzy's F1, given the extreme increases in F1 that were observed). It also seems that different speakers behaved very differently from others, and in different diphthongs. In most of these cases, where there is a large change over the period, it is partially due to the speaker having quite a conservative F1 range at the 21 stage. Many speakers exhibit more limited formant movements in the early age stages. There does not appear to be any hypothesis given in the literature to explain this pattern, especially given the larger habitual vowel spaces of younger adults. It also seems that where there are massive changes in F1 range (i.e. over 50% different over the period), that F3 also changes to a large extent, though the limits of the dataset and the fact this occurred in a handful of cases prevent us from making this assertion too strongly.

This 'slope' measure is comparable with findings from Liss et al. (1990), although for a much younger age range. However, it is probably not entirely suitable to apply to data from such a long period, as it is not possible to separate age-related changes in articulator flexibility from changes based on phonologically motivated sound changes within speakers (apart from in a very rough ad hoc way). The findings show consistent increases nevertheless; this might be surprising given the vowel space area findings in the previous section. It is clear that the 'extreme' production of diphthongs and the reduction (post 35) of the habitual vowel space, for these data, are not linked. It is unclear what the

cause of this range increase is; there is no previous research on age-related changes to steepness of transitions for anything other than extreme elderly subjects.

5.2.3 **Significance – targets and transitions**

It might be interesting to examine the differences, if any, between 'target sections' and transitions between them. However, there are limitations to the dataset in the present study. This section presents results of MANOVA analysis *in a purely exploratory way*. For these data, Box's M test was significant, which means the assumption of homogeneity of covariance was violated. This is problematic, especially for such small sample sizes, so these tests should be viewed as exploratory and having little analytical power (individual ANOVA for each interval were also not appropriate as data violated the sphericity assumption) (Field, 2005). To an extent, data that violate the covariance assumption can indicate results which are not significant (Field, 2005), although given the small and/or unequal sizes in these cases, even this is unreliable. These results should not be used to make any direct statements about significant findings, but might provide an insight into directions for later sections or future work.

Table 31 - MANOVA significance results for overall formants for both diphthongs (greyed names indicate small or very unequal sample sizes)

	/aɪ/						/eɪ/			
	Bruce	**Neil**	**Suzy**	**Symon**	**Tony**		**Neil**	**Suzy**	**Symon**	**Tony**
F1	n	**	**	*	n		**	*	n	n
F2	**	**	n	n	*		**	n	*	*
F3	**	*	**	n	n		**	n	n	n

n	not sig.
*	$p \leq 0.05$
**	$p \leq 0.01$

The table above demonstrates an overall picture of changes in formants. Clearly there is a mixed pattern, where some speakers are much more likely to exhibit differences. It has already been presented that speakers are largely showing reductions in formant frequencies, all the significant results here represent decreases, apart from Suzy's F1 in /aɪ/ and Neil's F2 and F3 in /eɪ/. What might be more interesting is (especially in those cases where there are not small/unequal samples) the behaviour at different intervals, where it might be possible to see the effects of context, and differences between target sections, glides and transitions between these.

Table 32 - MANOVA significance results for each interval for both diphthongs (greyed names indicate small or very unequal sample sizes)

		/aɪ/						/eɪ/			
		Bruce	Neil	Suzy	Symon	Tony		Neil	Suzy	Symon	Tony
F1	10%	n	**	**	n	n		*	**	**	n
	20%	*	**	**	n	n		*	*	**	n
	30%	n	**	**	*	n		*	*	**	n
	40%	n	**	**	*	n		*	*	**	*
	50%	*	**	**	*	n		**	*	**	*
	60%	*	**	**	*	n		**	*	*	n
	70%	n	**	**	*	n		**	**	**	n
	80%	n	**	**	*	n		**	*	*	n
	90%	n	**	**	*	n		**	n	*	n
F2	10%	n	n	*	n	*		n	**	**	*
	20%	n	*	**	n	n		n	**	**	*
	30%	**	**	**	n	n		n	**	*	*
	40%	**	**	**	n	n		n	**	*	*
	50%	**	**	**	n	n		n	**	n	n
	60%	**	**	*	n	n		*	**	n	n
	70%	**	**	n	n	n		*	**	n	n
	80%	**	**	n	n	n		**	**	*	n
	90%	n	**	n	n	n		*	*	*	n
F3	10%	n	*	n	n	*		*	*	**	*
	20%	n	**	**	n	*		**	**	**	*
	30%	*	**	**	n	**		**	**	**	n
	40%	*	**	**	n	**		*	**	**	*
	50%	**	**	**	n	**		n	**	**	n
	60%	**	**	**	*	**		n	**	**	n
	70%	**	**	*	**	**		*	**	**	n
	80%	n	**	n	*	**		**	**	n	n
	90%	n	n	n	n	**		**	**	n	n

There are a few observations we can make about these data, even though the analytical power of these tests is limited. Firstly, in many cases where the majority of intervals show a significant change, the first and last interval (and to a lesser extent those adjacent intervals) are less likely to show significant changes. It is most likely that this is not due to lesser frequency changes in those sections, but that they exhibit much more intra-speaker variation, due to differing consonantal contexts. Clermont (2007; 2009; 2011) shows that consonants have significant and predictable effects on vowel transitions; in this study consonant context was only partially controlled and there are a large range of neighbouring sounds which would lead to high internal variability in these intervals. In fact, some speakers will have more similar contexts than others, due to high frequency of 'like' fillers and narratives about 'life' and 'lives', given the content of the recordings. This

is where controlled elicitation has an advantage over spontaneous speech, (but sacrifices ecological validity). In forensic terms then, it might be sensible (if using a dynamic type approach with formant transitions) to exclude initial and final sections; or to use a normalisation technique like that mentioned in the studies by Clermont above. However, if the aim of a forensic exercise is to characterise a speaker, analysts should be wary of [normalisation] techniques that enhance the similarity of frequency data and should certainly be required to state these processes in any reports.

Secondly, the speakers sometimes show changes in one formant and not others, which might suggest changes other than simply universal changes in vocal tract length; sociolinguistic changes, extent of jaw movement or laryngeal setting might explain these differences, although again more data is needed to make any firm statements about this. Finally, as was mentioned before in this chapter, changes in one diphthong do not always imply changes in both, which probably entails either sociolinguistic change, an effect of the inherent quality of one of the individual targets, or both. More data and data from other formants might shed light on this in a more extensive way.

Results of post-hoc tests show that most of these significant changes show stronger p values for the 21-49 comparison, which is logical if we are expecting a steady process, such as vocal tract lengthening or other physiological changes already mentioned. Longer time periods also logically infer more opportunity for speakers to undergo social changes as well. The limitations of this analysis prevent any clear conclusions, but the data do allow for tentative observations about formant behaviour at different portions of the diphthongs. Given that research on the effects of aging is limited, and virtually non-existent for this age range, it is disappointing that the recordings yielded so few appropriate tokens. This does point to a weakness of using these kinds of data given their availability in free speech (many studies have examined these using elicited tokens).

5.3 **Discussion**

This section addresses the research questions set out below in light of the acoustic diphthong results in this chapter. The limitations of the diphthong dataset have been acknowledged in this chapter and consequently the results should be viewed with prudence. These data are one of the only longitudinal perspectives on formant transitions for this age range (perhaps with the only exception of Harrington et al.'s

(2005) study of the Queen), so it is difficult to draw comparisons with existing research. Nevertheless, findings are discussed in light of other formant changes and those which might be expected following ideas around formant dynamics.

5.3.1 **Research question 3**

3 What are the effects of aging on formant transitions?

In summary, average diphthong formants show decreases in frequency which are comparable to those for the monophthong dataset, although changes in F2 and F3 were more marked; F1 changes were still the greatest. This is perhaps expected given that the literature (summarised in §2.1.4), and findings from this study, suggest that vocal tract extension and reduced articulator flexibility are two main drivers for aging change. These processes affect vocal resonances fairly universally, and there is no reason to suggest that diphthongs should behave differently to monophthongs in this respect.

Minimum-maximum frequency range (token internally) or 'slope' showed quite substantial increases across the period, with most speakers' ranges increasing (anywhere between 10 and 150%) in most formants. Comparable conclusions in Liss et al. (1990) suggest reductions in slope in elderly groups, but these are based on LTAS measures and relate to much older speakers than those in the current study. It is, however, plausible that the increases in this study are not incompatible with reductions in slope for elderly speakers, who are shown to exhibit limitations in more complex articulatory gestures due to stronger degradation in innervation.

These changes were both highly variable between speakers and also somewhat variable between diphthongs and formants within speakers. Although these changes seem to be consistent and similar to monophthongs overall, observing frequency changes and slope independently entails that this analysis cannot really be considered dynamic. In the following chapter, polynomial regression will be used to calculate coefficients for these curves, which capture both these types of changes. This will illustrate to what extent this approach can function in the face of the changes observed in the current chapter and comment on the usefulness of dynamic parameters in that respect.

5.3.2 **Comparison with monophthong data**

A question with practical implications for forensic casework is:

a. How do these effects compare with monophthong formants?

In general, age-related change displayed similar effects with diphthong data as that illustrated for monophthongs. F1 changes were of the greatest magnitude, as for monophthongs. F2 and F3 changes for diphthongs showed changes of larger magnitude, though there are differences in how measurements were made and the number of different vowel classes in the analysis. For example although overall F2 changes for monophthongs (mean 3.7%) were less than those for /aɪ/ (mean 7.7%), for /a/, F2 changes were mean 7.2% across all speakers; perhaps this is a more suitable comparison given the vowel-quality related differences in acoustic changes shown for monophthongs. For /eɪ/, F2 and F3 changes (mean 7.9% and 8.5% respectively) were still larger than for all monophthongs and for /e, i: & ɪ/ changes.

Furthermore, the diphthong data were taken from averages across whole tokens, whereas monophthong tokens were measured across a short central portion of the vowel. Perhaps a truly 'dynamic' measurement approach with monophthongs would have yielded more variability i.e. at consonant transition points, as it did for diphthongs data. Presumably monophthong formants are more speaker-internally stable at central point than across the whole duration.

These analyses are not designed to assess directly the performance of monophthongs and diphthongs in discriminating between speakers, but if frequencies for diphthongs seem to be changing at a greater magnitude than those for monophthongs, this might suggest that they would perform worse at characterising a speaker in a long-term non-contemporaneous exercise. Both sets of parameters will be used to calculate LRs over increasing delays in the following chapter to address this issue more closely. While more comprehensive diphthong data would be better suited to this kind of discriminant testing, it is important that spontaneous speech is used, given the findings of studies into the performance of dynamic formant data with non-elicited speaking styles (Atkinson, 2009; Hughes, McDougall, & Foulkes, 2009; Rhodes, 2009).

5.3.3 **Dynamic questions: targets, glides and transitions**

b. Are speakers more stable in 'target' sections of a diphthong, or in their movements between targets?

The data are not extensive enough to allow for appropriate testing to answer this question, though the first and last portions are much more internally variable. This is most likely due to consonantal coarticulations effects. There were consistent findings

regarding age-related change between the initial target vowel and glide. This includes a tendency for glide or /ɪ/ sections to reduce more in F1 and less in other formants in /eɪ/, which mirrored vowel quality based differences found in monophthongs, though this was based on two of four speakers for whom formants decreased in /eɪ/.

Individual differences between extent of change in target and glide portions might also suggest that, as for monophthongs, there may be sociolinguistic adjustments by speakers which have to be taken into account. Some speakers showed different types of frequency change in the target or glide sections which were not in line with predicted aging patterns. These changes in vowel quality might be difficult to differentiate from age-related changes in casework.

5.3.4 **Dynamic diphthong data in cases of long-term non-contemporaneity**

c. Is making dynamic measures of vowel formants worthwhile and/or reliable in cases of long-term non-contemporaneity?

The first point to make is that frequency of tokens of diphthongs is very low. Only PRICE and FACE vowels had sufficient availability to include in this study (and only in four or five speakers); even those numbers were too low to complete comprehensive analyses. Other diphthongs had vanishingly small measurable tokens. Of course, if there are longer forensic recordings with sufficient tokens, these measures should be used, but it seems for the present dataset (which was relatively long) there are too few tokens to satisfy Nolan's (1983) availability criterion.

5.3.4.1 Worthwhile?

Where tokens are available, what benefits do dynamic measures have over taking single or two-point measurements? Other studies have shown that for spontaneous speech, these 'non-dynamic' measures are effective at discriminating between speakers for diphthongs (Hughes, McDougall, & Foulkes, 2009; Rhodes, 2009) and for monophthongs (Atkinson, 2009). These were fairly small scale studies which did not use regression to process transition curves, but they demonstrated that including frequency data from the whole duration of formant transitions did not greatly improve discrimination rates. It has been mentioned in the previous section that peripheral intervals may have an obfuscatory effect given consonantal transitions, and it would be interesting to see the effects of putting into place a consonant normalisation technique. It is likely that a major difference between the 'laboratory' and 'spontaneous' samples' performance is due to

controlled (articulatorily transparent) phonetic context in the former and a different consonant contexts in the latter. Although these wider questions are interesting, in terms of the specific question of non-contemporaneity, the analyses in this chapter are unfortunately not sufficient to assess efficacy of speaker discrimination for diphthongs. Further LR analyses in the following chapter yield more useful results.

Findings from the current chapter that are of interest, however, are the very substantial differences (between speakers, formants and age stages) in slope of transitions, which are most effectively captured by dynamic measures. Certainly being able to capture this kind of information (whether simplistically or as part of representing transition curves via regression) is an advantage of a dynamic approach. In general, extracting measures across the whole duration of a formant transition as a matter of course presents more information (duration, range), and also flexibility in analyses (you can extract two single points from a dynamic measure, but you cannot infer dynamic movements from two measures). Whether this extra information is always useful in practical contexts is a difficult question which still remains to be fully explored.

5.3.4.2 Reliable?

With respect to data in this study, the question of reliability concerns the underlying age-related changes to diphthong formant frequencies. It might be useful to refine this part of the research question, and modify a question this thesis presented in reference to monophthongs:

> What effect should this [diphthong data] have on how we evaluate forensic speech evidence?

Given that the diphthong formant data show similar processes of age-related decreases to monophthong data, the recommendations given in the previous chapter (§4.6.2) should be extended to address diphthongs. Analysts should practice caution when dealing with formant data which are likely to feature these kinds of age effects, and of the importance of knowing as full a context of suspect or reference speakers as possible in order to be able to make sound predictions about these processes. Diphthongs are shown to experience similar shifts in formant frequencies to monophthongs. Moreover, extremity of gestures also increases within this set. The effects of aging vary by different factors (including vowel type) in very individual ways for different speakers.

5.3.4.3 The Queen's Speech

This evidence, and evidence from the monophthong data, addresses assertions from another longitudinal case study. Harrington et al. (2005) reject age-related changes as an explanation for two of the Queen's diphthongs which exhibited lowered F1 in their initial vowel (/eɪ/ and /oʊ/). In fact, they also reject an age explanation for similar changes in monophthongs. They reject age-related changes on three grounds: that there is not enough research into aging, that the Queen's changes were of a larger magnitude than in Ratstatter et al. (1997) (for elderly speakers in controlled reading contexts) and thirdly that the reference speakers they used to characterise widespread SSBE had lower F1 for some vowels than the Queen's later recordings but were not older (judged auditorily). Therefore, they argue that the Queen's decreases in F1 represent a movement towards SSBE and not an age-related change. However, the arguments they put forward are flawed and based on a number of assumptions that cannot be supported by data.

There are studies into aging which show similar patterns to many of the changes shown for the Queen which are addressed as conforming to widespread changes. In fact, the Queen seems to share a similar pattern with Suzy in the present study (with respect to monophthongs and diphthongs), where the general pattern is for a decrease, with few individual vowel changes (most noticeable F1 of open vowels) towards widespread accent norms (perhaps). It would not seem sensible to completely discount age-related change but to accept that both processes are probably at work; two processes which are impossible to tease apart in an ad-hoc way, especially if they are in the same direction. The only research which could take full account of these processes would combine a study of the kinds of widespread accent-changes in Harrington et al.'s research and measures of physiological changes, such as vocal tract length and habitual tongue/jaw movement, which could be used to estimate likely acoustic changes based on these physiological changes.

6 Likelihood Ratio estimation

This chapter investigates what effect age-related acoustic changes can have on likelihood ratio estimation, and uses data from the present study to elucidate research question 4:

4 What effect does age mismatch between criminal and suspect recordings have on LR estimation?

This analysis also illustrates the effects of age-related changes in formant frequencies on strength of speech evidence and explores the resilience of dynamic diphthong measures to aging.

6.1 Motivation for LR estimation

The principal motivation for illustrating the effects of aging within an LR approach is that the magnitude of within-speaker changes can be measured against the distribution of these parameters within the population, not just in isolation (as in the previous chapters). Using this approach also allows the data to be framed in a more practical way, assessing the evidential value of the parameters as well as their raw frequency. This should give an interesting perspective on aging and the relative resilience to aging of different vowels (vowel classes and monophthongs vs. diphthongs) and to what extent larger changes in frequency over time result in changes to strength of evidence. It is important to bear in mind that these estimates are based on the best available reference population, and that there is currently insufficient research into how closely a reference population needs to be matched to the potential perpetrator population. There are further questions regarding how the reference population should be collated and whether this is based on other factors in the case or by using other speakers which sound similar to the evidential speaker (Morrison, Ochoa, & Thiruvaran, 2012). In terms of assessing aging change, nevertheless, the reference population is stable throughout each analysis and demonstrates the effects of age.

There are questions about long-term delays and aging, and other forms of mismatch that are not addressed in the LR literature. The reference population is typically matched with the evidential sample (following the model of a 'potential perpetrator population'). In most cases it would be expected that the reference population would match both samples fairly well. But how well does the LR analysis estimate strength of evidence if

there are inconsistencies between suspect and evidential samples that the analyst would logically expect? In some cases, such as those with non-contemporaneous samples, this would not cause us to exclude the prosecution hypothesis without analysis. In channel or telephone effects this mismatch can be countered with pre-analysis filtering. But with factors such as aging, and other factors which are more likely to be present in cases with a long delay (for example, mobility, lifestyle changes) it is unclear how to proceed within the LR approach where this kind of mismatch is known or expected.

6.2 LR estimation testing

This chapter presents LR estimation scores for formant frequency data from monophthongs from Bruce and Neil and PRICE diphthong data from Bruce, Neil, Symon and Tony. These speakers and parameters were selected as they offered the largest quantities of data and, for monophthongs, those which matched available reference databases for accent. Using these examples as case studies, it investigates the relative discriminatory power of the formant parameters in light of aging.

Data extractions and analysis procedures for the LR estimation are presented in §3.5.3. All tests in this chapter are between the same speakers, with differing levels of delay between suspect and offender samples. Sufficient data were available (in samples and reference populations) to assess 6 monophthongs: /iː ɪ e a ʌ ɒ/. Data from the earliest recording (21) was used as evidential-type data from which later samples were tested. The reference population for these tests were 25 (23 for FLEECE) speakers for monophthongs, and 100 speakers for diphthongs, of SSBE from the DyViS database (Nolan, McDougall, de Jong, & Hudson, 2009). In the case of the monophthong testing, data from Deterding (1997) were used to pilot LR testing, with very similar results (differences in the magnitude of LLR were apparent, but changes were similar). However, there were only 12 speakers from this dataset, and Hughes (2012) shows that LRs are not reliably estimated in tests with reference populations as small as 12, particularly for negative, support-for-defence scores. Results presented below are all from tests with DyViS reference data. LRs are presented in $\log_{10}$ format unless otherwise stated (LLRs); positive LLR scores represent support for the same-speaker hypothesis while negative LLR scores represent (in these cases false) support for the defence hypothesis. It might be useful to revisit the following table from §2.3.5.5, which shows the verbal alternatives for

Log_{10} LRs, to ground the results in a practical application of the strength of evidence assessments:

Table 33 - Scale of LRs and strength of verbal support for the evidence

Likelihood Ratio	Log_{10} LR	Verbal expression
>10000	5	Very strong evidence for the prosecution hypothesis
1000-10000	4	Strong evidence for the prosecution hypothesis
100-1000	3	Moderately strong evidence for the prosecution hypothesis
10-100	2	Moderate evidence for the prosecution hypothesis
1-10	1	Limited evidence for the prosecution hypothesis
1-0.1	-1	Limited evidence for the defence hypothesis
0.1-0.01	-2	Moderate evidence for the defence hypothesis
0.01-0.001	-3	Moderately strong evidence for the defence hypothesis
0.001-0.0001	-4	Strong evidence for the defence hypothesis
<0.0001	-5	Very strong evidence for the defence hypothesis

Source: Champod and Evett (2000)

In theory a 'good' parameter would be one that is resistant to age-related changes and which continues to present (correct) positive LLRs.

6.3 LR estimation results

This section presents results from monophthong and diphthong data. This analysis illustrates issues concerning, principally age-related, mismatch between suspect and evidential (or offender) data, and test and reference population data.

6.3.1 Monophthongs

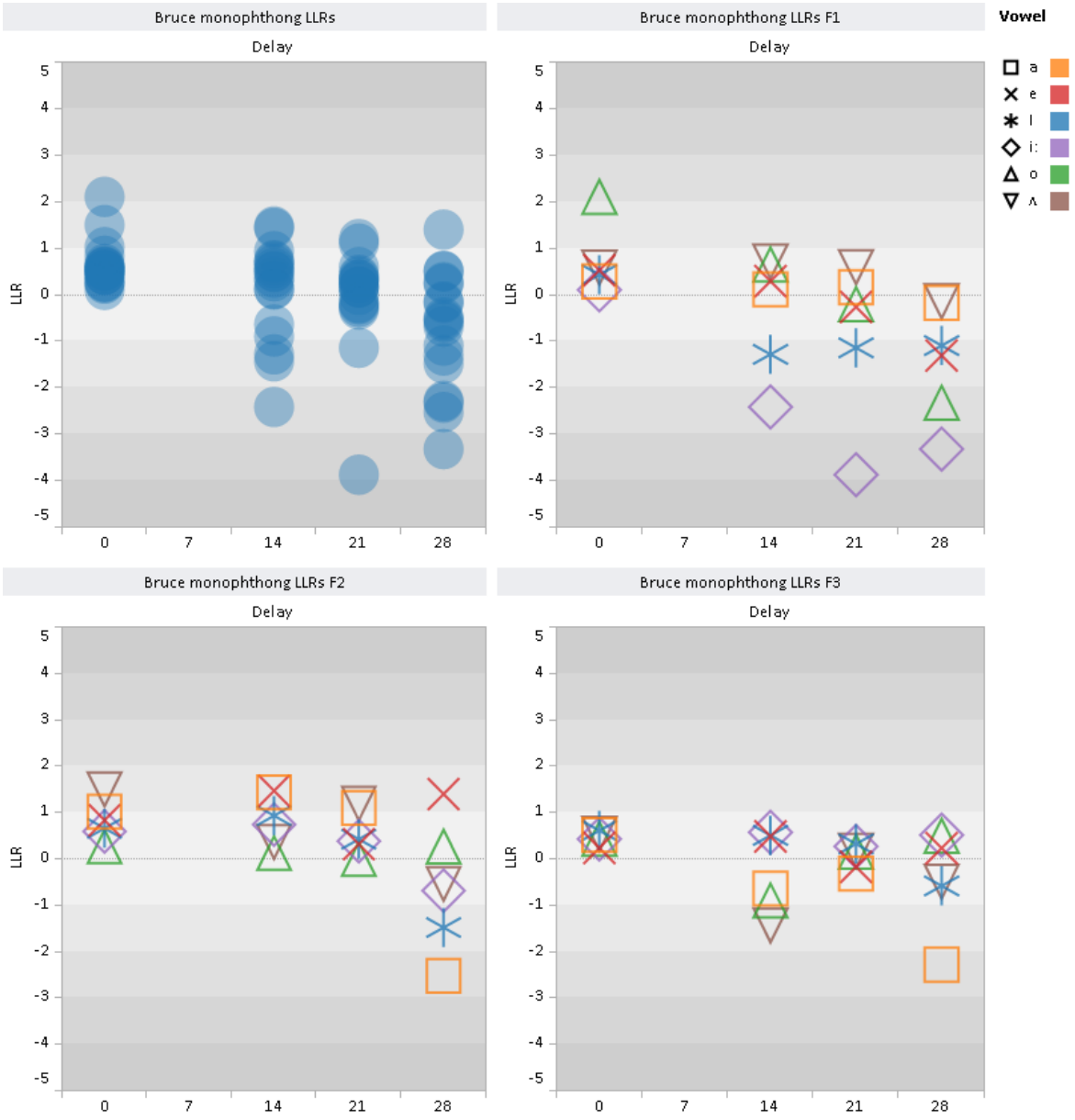
6.3.1.1 Predictions

It is sensible to assume that the largest changes in LLR scores are likely to be in those monophthongs which exhibited largest changes across the 30 year delay. This would entail larger LLR score changes in, for example, F1 overall, and specifically in close front vowels in F1, and open vowels in F2 (see §4.1 for specific results). The distance between the evidential sample and the reference population will also affect the LLR score, however. It may be that even though formants are changing, those that move closer towards reference distributions may not exhibit such large LLR differences (and *vice versa*).

6.3.1.2 Bruce

Bruce exhibits the most consistent and steady reductions in formant frequency, so a comparable decline in strength of evidence is predicted. Figure 74 below shows all tests (top left, blue) and all three formants. All monophthongs tests show positive LLRs in the 21-21 comparison, correctly estimating support for the same-speaker hypothesis. Although the LLRs are generally individually weak, the majority somewhere between 0 and +1, if they were combined, fused or presented together it would lead to a larger overall impression of the strength of evidence (issues around best practice for presenting multiple LRs is outside the scope of this study, but in need of attention).

Figure 74 - Plots showing log_{10} LR scores for Bruce's monophthong data with increasing delay (years). Greyed bars represent stages of the verbal scale for presenting LRs. Reference population – DyViS, N = 25



In the 14 and 21 year delay tests, a small number of tests present negative LLRs, but the majority are still above 0, thus correctly providing support for the same-speaker hypothesis. In the 28 year delay condition, the results are spread fairly evenly between -3 and +1, with 11/21 tests showing (false) support for the different speaker hypothesis. After this longest delay, 13 of 21 tests support this incorrect hypothesis; these are not evenly spread, with all six for F1, four for F2 and three for F3. This finding was predicted based on the descriptive data reported in chapter 4, with more marked changes in F1 than F2, and F2 than F3.

In terms of F1, KIT and FLEECE show consistently negative LLR scores after 14 and 21 years, unlike most other vowels. This should probably be expected given that we know these

vowels are most likely to show greater and more significant reductions in F1. After 28 years, all vowel calculations result in negative LLRs between about -1 and -2.

For F2, LLRs stay consistently above or around 0 for all conditions apart from the 28 year delay, where FLEECE, KIT, TRAP and STRUT all show negative LLRs between 0 and -3. It is certainly expected that the open vowels would show large LLR changes, based on observation of all speakers' data. Bruce shows steady reductions in F2 for close front vowels, unlike most speakers, which means we might also expect these results for FLEECE and KIT for his particular data.

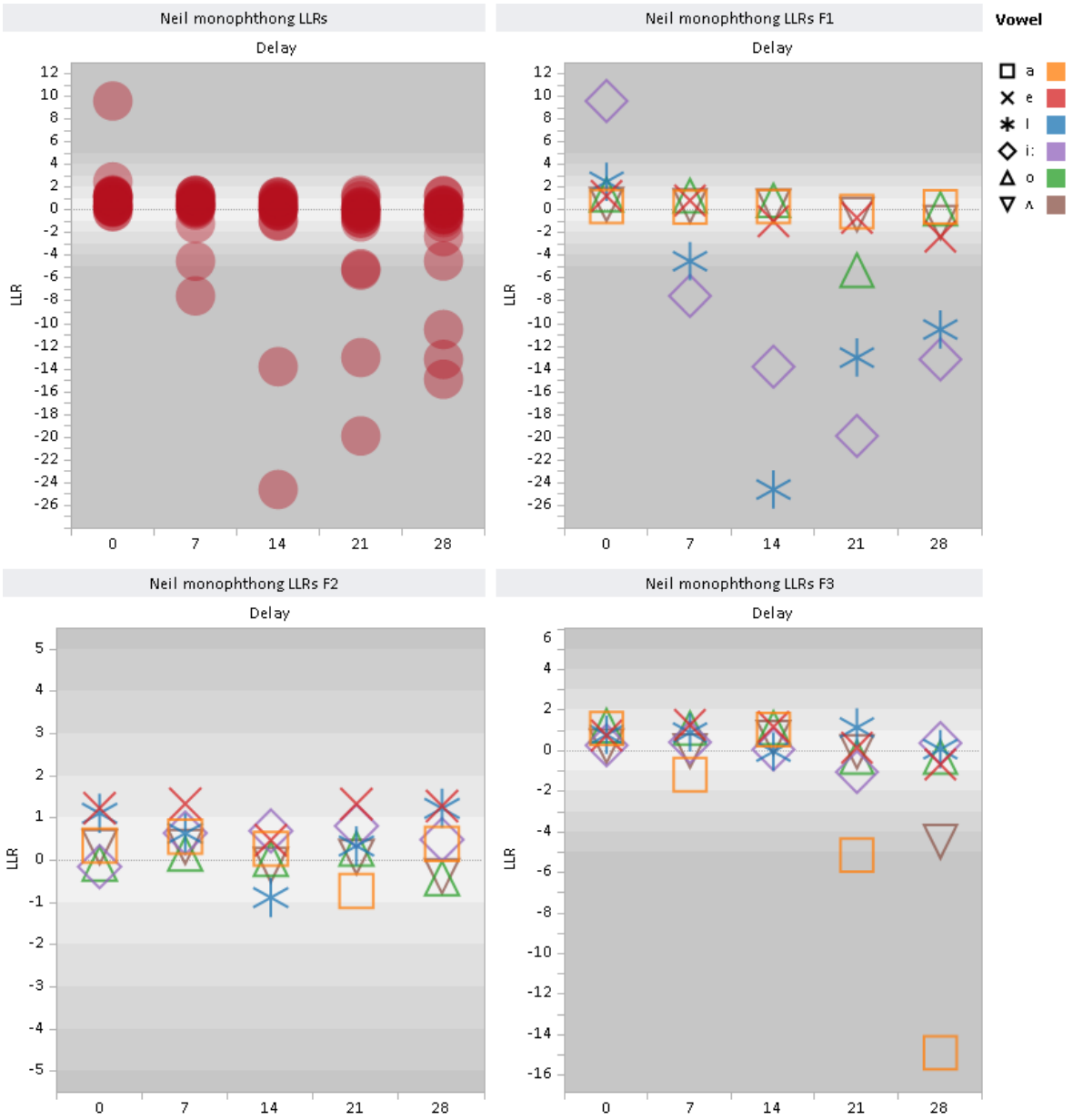
For F3, the results are fairly similar, with increasingly negative LLR scores across the period. TRAP vowels show negative LLRs across all delay conditions and half of the tests show negative LLRs between about -1 and -2 in the 28 year delay (as in F1).

6.3.1.3 Neil

Figure 75 below shows results for Neil's monophthong data. Principally of note, especially in the overall graphic (red circles), is the much larger magnitude of some of the LLR scores. Although most of the tests are within the -5 to +5 range, there are a number of very large scores, up to +10 or -25 LLR. These come from KIT and FLEECE in F1 and TRAP in F3.

This section first focuses on these larger results and subsequently discusses results in the -5 to +5 range (as with Bruce) for a more direct comparison.

Figure 75 - Plots showing $\log_{10}$ LR scores for Neil's monophthong data with increasing delay (years). Greyed bar area represent stages of a verbal scale for presenting LRs. Reference population – DyViS, N = 25



In Figure 75 above, it is clear that a small number of (largely negative LLR) tests have far greater magnitude LLRs than most of the others. This is probably due to a combination of factors, amplified by the relatively poorly matched reference population. The first factor is that for these tests (F1 of FLEECE and KIT and F3 of TRAP) there were some of the largest changes in formant frequency (over 100Hz in F1). In light of this, we would indeed expect the LLRs to be negative and perhaps stronger, as this reflects the between (suspect and offender) sample variability, therefore the samples are less similar. The second factor is slightly more complex, and is probably the cause of the very high magnitude of some of the tests, which reached up to LLR -27. These scores were not only highly different, but

would be placed towards the extremities of the distribution of data in the reference population. Average F1 in the reference population speakers was 375Hz for KIT and 291Hz for FLEECE, these averages were 501 and 495Hz respectively for evidential data for Neil at 21. Again, the LLR logically shows larger magnitude scores where the test samples are so far from the reference population, meaning they are atypical of the general population. Both of these factors are evinced in Figure 76 and Figure 77 below:

Figure 76 - Scatterplot showing F1 and F2 means for FLEECE for Neil at each stage and the DyViS reference speakers

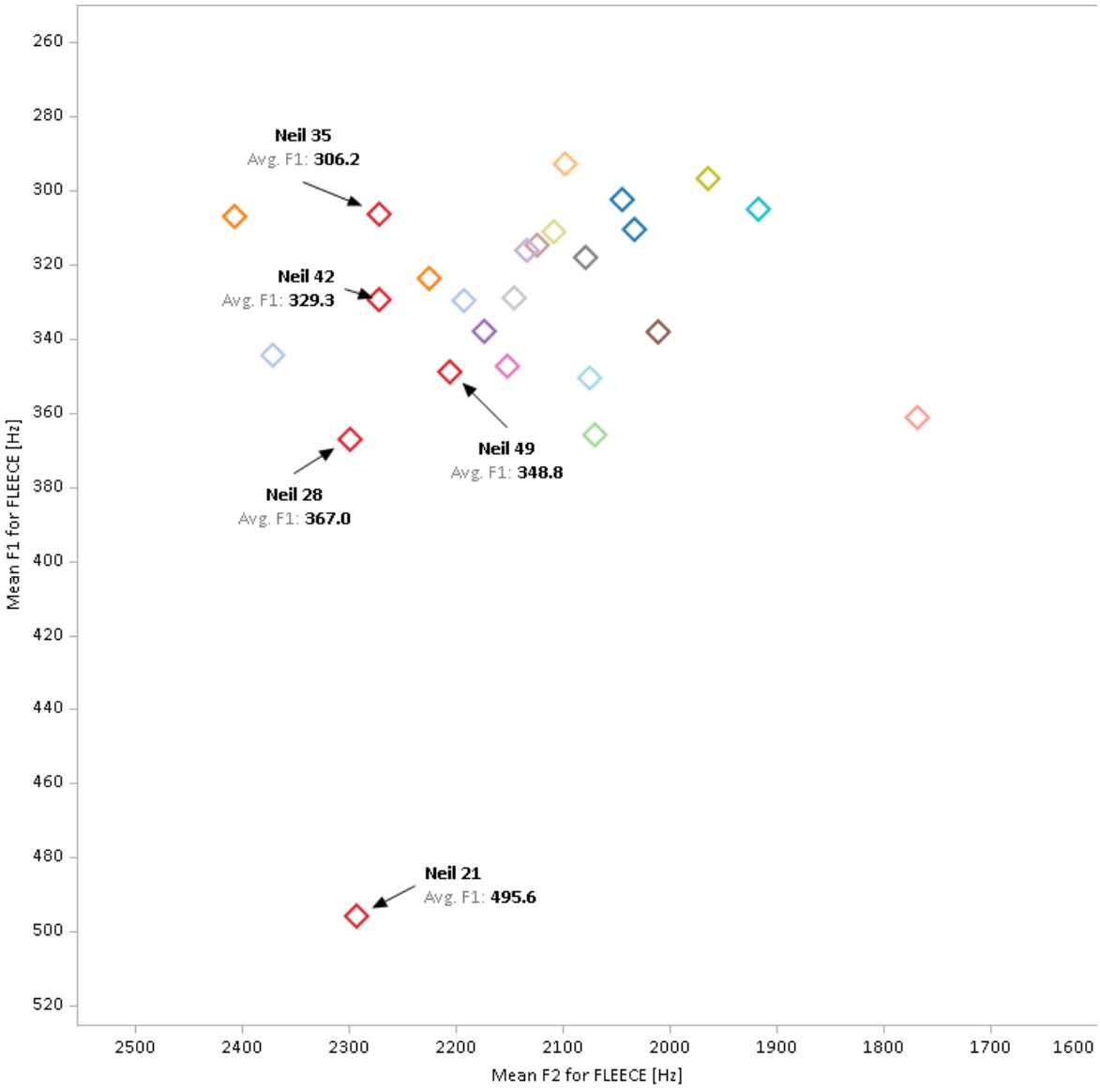
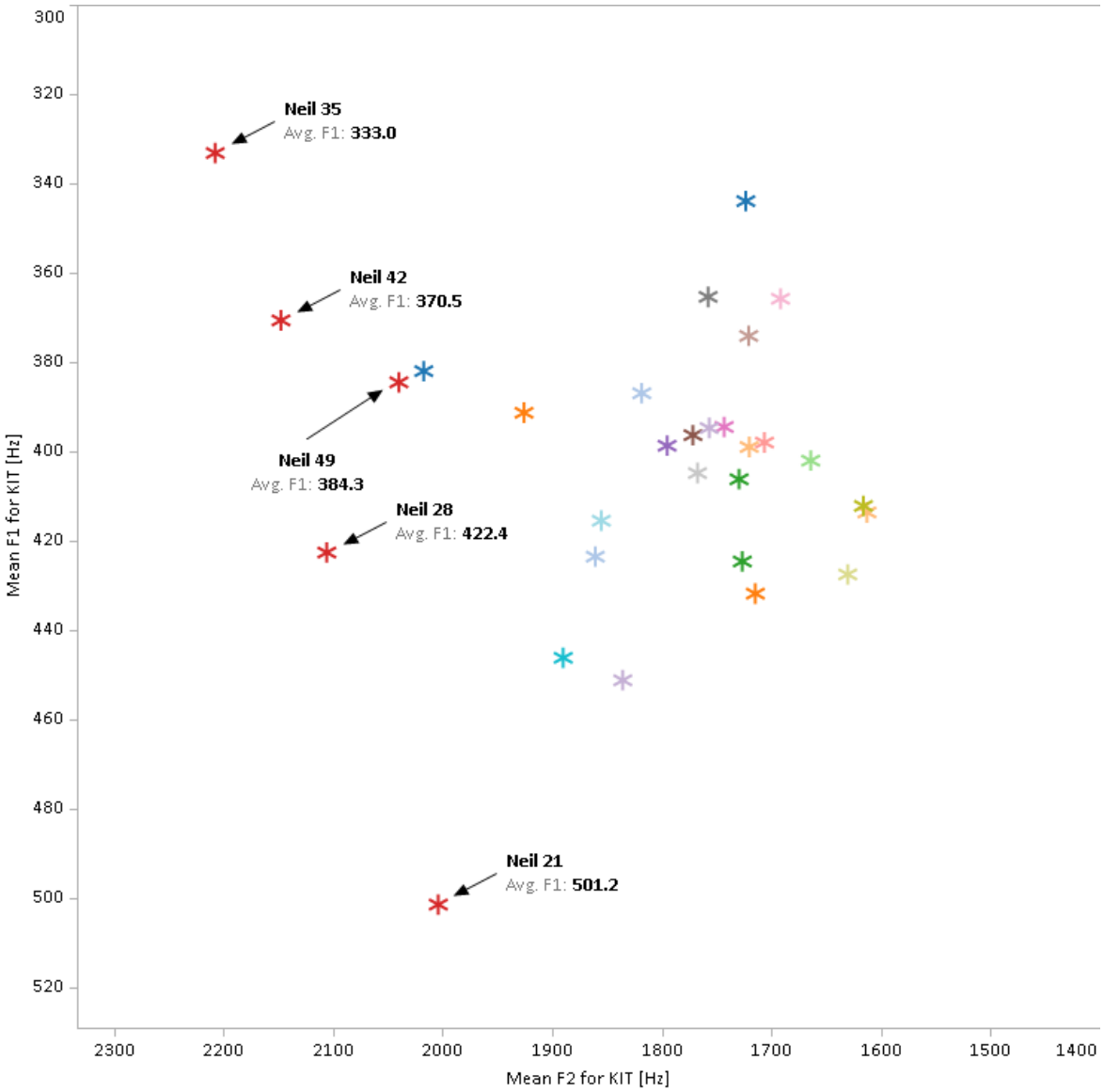


Figure 77 - Scatterplot showing F1 and F2 means for KIT for Neil at each stage and the DyViS reference speakers



Not only are the evidential samples far removed from the F1 distributions in these tests, but the changes across the delays are substantial. Both of these factors are exactly what the LR is designed to test, but there are problems with the modelling behind the LR when samples are at the extreme of the reference distribution, such as in this case.

Lindh et al. (2012) demonstrate that where (in this case disputed utterance) test data is on the extremity or 'tail' of the reference distribution, small perturbations in the data can cause dramatic changes to the LR score. Furthermore, they used Monte Carlo simulation to indicate that in fact their massive LR (of 10^{77}) was way out of the distribution of 95% of the results in simulated tests, and that a better estimate should be around 10^{9}. Although LRs in the present study are not as large as their disputed utterance LR, they are based on single formant vectors, whereas Lindh et al. (2012) tested LRs based on a (combined)

F1,F2,VOT vector. Furthermore, with the relatively small number of reference speakers in this test, the effect could be further amplified; Hughes (2012) shows that LRs are not reliably estimated in tests with reference populations as small as 25. It is also possible that this is partly due to a mismatch of the reference population and the test sample, and that the influence of Neil's original Liverpool accent, or articulatory settings apparent at 21, mean the RP reference population is unsuitable. It is unclear how to proceed in cases such as this, where it would be very difficult to recruit a large '1970s-Liverpool-origin adopted-SSBE speaker' reference population, other than to exclude certain data or abstain from using a numerical LR. There is clearly a tension between how we define a 'well-suited' reference population and assess divergence from their distribution for forensic purposes.

It is also worrying that, according to Lindh et al. (2012), where test data are further from reference data, yielding higher LRs (in the same vein as the typicality or high between-speaker variability criterion) the modelling behind the LR seems more unstable. That is to say stronger evidence, which is more desirable and of higher value to the forensic 'client', seems to be less reliably estimated. They suggest that Monte Carlo simulation may be one way of offsetting this risk, by estimating the likely distribution of test scores in thousands of simulated cases. However, this simulation of testing might be considered unacceptable and would probably not be understood well by a jury. This is certainly a point which requires further investigation.

Returning to the less extreme data between -5 and 5 for Neil in Figure 75 below, the summary graph (top left, red) displays a similar pattern to that for Bruce's data, where LLR scores for age 21-21 tests are (in the majority) showing positive values. There are two tests at this stage which show minimal negative LLRs somewhere between 0 and -1. All others show a positive score between 0 and 1, which again accords with Bruce's data. For tests with suspect data at age 28 and 35 (7 and 14 year delays) the majority of scores remain between 0 and 1, with an increasing minority showing stronger negative LLRs at 14 years delay than at 7.

Results for a 21 year delay indicate that LLR scores are spread around 0, with just over half of all LLRs showing false support for the defence hypothesis. For Bruce, this kind of pattern was only found after 28 years. These negative LLRs are fairly well spread over the formants, though fewer F2 scores are negative (at almost every stage). The spread of

scores is similar for Neil at the 28 year delay, although there are slight differences between the specific vowels and formants which result in negative LLRs.

Figure 78 - Plots showing $\log_{10}$ LR scores (within -5 to +5) for Neil's monophthong data with increasing delay (years). Greyed bars represent stages of a verbal scale for presenting LRs. Reference population – DyViS, N = 25

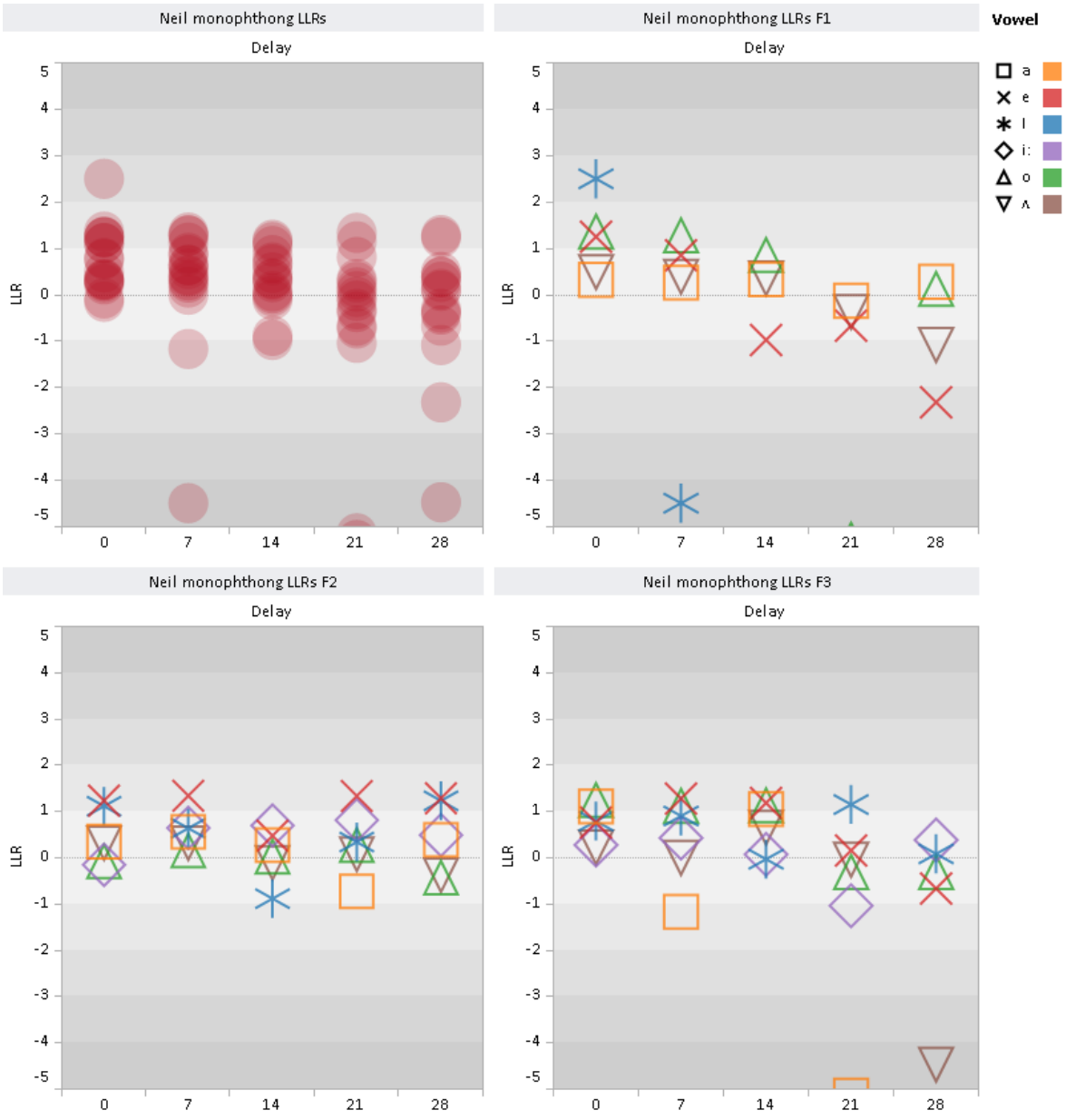


Figure 78 above illustrates results between -5 and +5, and is useful for referring to changes in specific formants. The pattern across formants is similar to that in Bruce's data, although in Neil's case F2 seems to display fewer negative LLRs than F3. For F1, the LLR results concord with what is predicted based on vowel category (and height in particular), where open vowels are least likely to be affected (certainly for TRAP, START and to a lesser extent STRUT).

In the case of F2, none of the LLR scores fall outside of the -5 to 5 range displayed above. In fact, although there is variation around 0 at each stage there are only very minimal changes, and not predictably in vowel categories. While we might expect more open vowels to change, based on data observed from all subjects, Neil's F2 is relatively stable compared with the group and that would explain why the scores are remaining within the 'limited' evidential value boundary of -1 to 1.

For F3, (ignoring the extreme TRAP scores) again we would expect LLRs for close front vowels to be most resistant to aging, given overall frequency changes and those for Neil in particular. This is demonstrable from the 21 and 28 year delays, where KIT and FLEECE vowels are somewhat more likely to show positive LLRs.

6.3.1.4 Summary

Table 34 below summarises the LLR scores for all monophthong tests. In general almost all contemporaneous tests showed positive LLRs. Increasing delays leads to increasing numbers of results in false support of the defence hypothesis. This is predicted by the age-related vocal changes described in previous chapters. Changes in LLR scores are somewhat predictable from the vowel-category-specific changes in monophthong formants. This pattern was not complete, as the distance from reference population distribution also affects scores. In some cases, tests from single formant frequency vectors show large magnitude LLR estimates. Reasons for this are presented and discussed above.

Table 34 - Showing log_{10} LR scores for all monophthong tests across increasing delays, colour formatting is dependent on magnitude of LR (green = positive, red = negative)

		Bruce 0	Neil 0	Neil 7	Bruce 14	Neil 14	Bruce 21	Neil 21	Bruce 28	Neil 28
F1	o	2.1	1.4	1.3	0.7	0.9	-0.2	-5.3	-2.3	0.1
	ʌ	0.6	0.5	0.4	0.7	0.4	0.6	-0.4	-0.1	-1.1
	a	0.3	0.3	0.3	0.1	0.3	0.2	-0.1	-0.2	0.3
	e	0.5	1.3	0.9	0.3	-1.0	-0.3	-0.7	-1.3	-2.3
	ɪ	0.4	2.5	-4.5	-1.3	-24.6	-1.1	-13.0	-1.1	-10.5
	i:	0.1	9.6	-7.5	-2.4	-13.8	-3.9	-19.9	-3.3	-13.1
F2	o	0.3	-0.1	0.2	0.1	0.0	0.0	0.3	0.3	-0.4
	ʌ	1.5	0.3	0.3	0.4	-0.1	1.2	0.1	-0.5	-0.4
	a	1.0	0.4	0.6	1.4	0.3	1.1	-0.7	-2.5	0.4
	e	0.8	1.2	1.3	1.5	0.5	0.3	1.3	1.4	1.3
	ɪ	0.7	1.1	0.6	0.9	-0.9	0.4	0.3	-1.5	1.2
	i:	0.6	-0.1	0.6	0.7	0.7	0.4	0.8	-0.7	0.5
F3	o	0.4	1.2	1.1	-0.9	1.1	0.2	-0.3	0.5	-0.3
	ʌ	0.5	0.2	0.0	-1.4	0.7	0.2	0.0	-0.5	-4.5
	a	0.5	1.1	-1.2	-0.6	1.1	-0.3	-5.1	-2.3	-14.9
	e	0.2	0.8	1.3	0.5	1.2	-0.2	0.2	0.2	-0.7
	ɪ	0.6	0.8	0.9	0.5	0.0	0.3	1.2	-0.6	0.1
	i:	0.4	0.3	0.4	0.6	0.1	0.3	-1.0	0.5	0.4

6.3.2 **Diphthongs**

LLR scores for diphthong data presented in this section are calculated based on polynomial (cubic) coefficients of formant frequency contours of PRICE tokens from Bruce, Neil, Symon and Tony. Tests in this section present an estimate of the effects of age-related changes on LLR scores for dynamic measures. To put this into context, these results are compared with LLR scores calculated using static measures of the same diphthong data. The reference population used in estimating these LLRs is much larger than for the monophthong tests, with 100 speakers from the DyViS corpus. However, limited comparisons are still possible between the two analyses. Token numbers for the reference speakers are comparable with the monophthong tests, with around 12-14 tokens per speaker.

6.3.2.1 Predictions

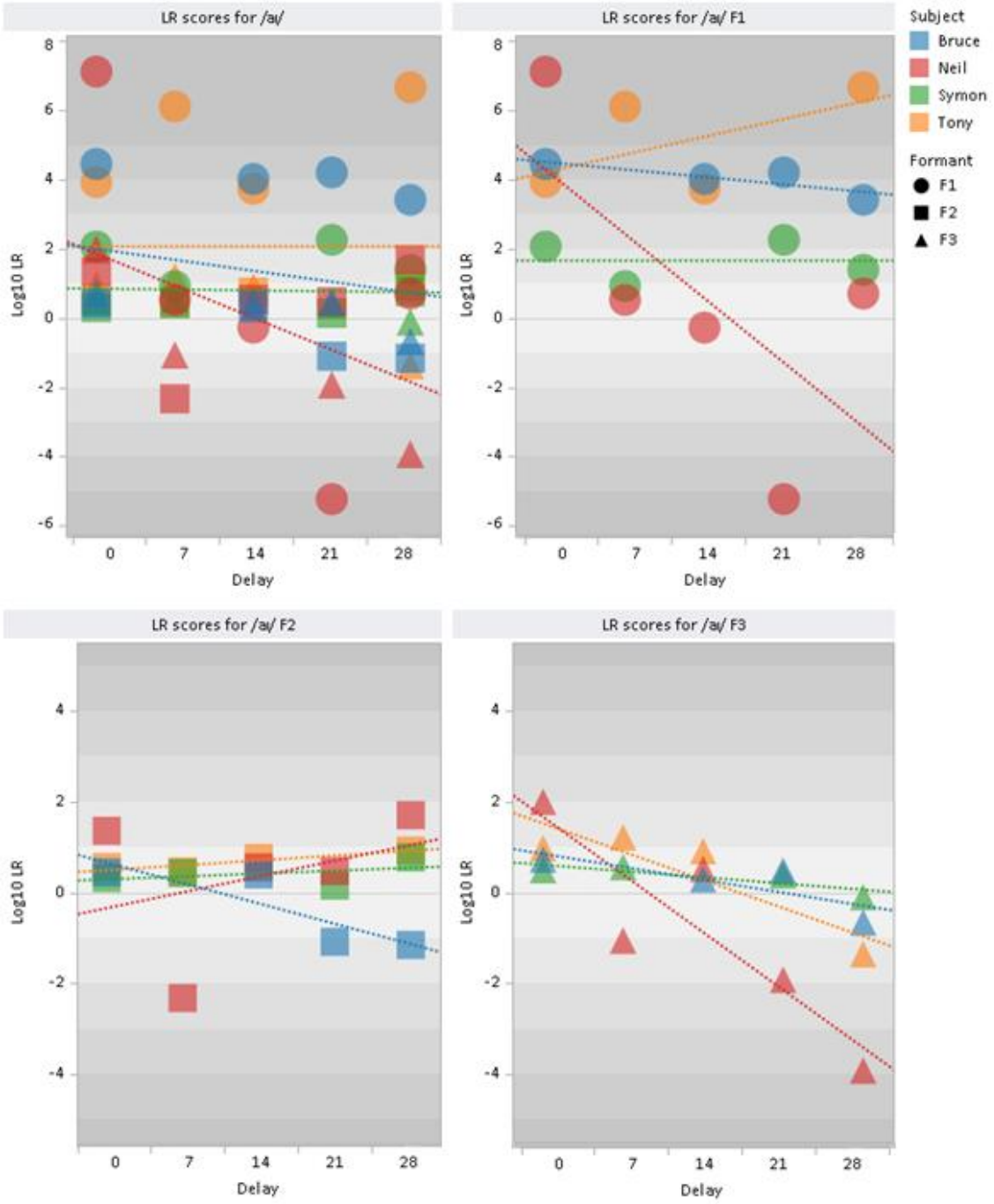
As this analysis of aging effects on FD measures is unprecedented, it is not possible to make specific comparisons with existing research. However, the literature tells us that in some cases, dynamic measure outperform other formant based parameters (cf. studies summarised in section 2.2.2). Therefore, LLRs for these tests could show greater strength of evidence than static monophthong or diphthong parameters.

It is also possible to make predictions from the underlying theory of the formant dynamic method. Speakers' vocal physiology is shown to change with age, however, a dynamic method is designed to characterise idiosyncratic speaker behaviour. Although physiology is changing, within this conception the organic properties of the vocal tract are the outer constraints of movements, not the main determinant (Nolan, 1983). It could be predicted then, that speaker behaviours persist in spite of biological changes and therefore dynamic measures will be more resistant to aging than static parameters. This may be further complicated by the idea that this idiosyncratic behaviour manifests in transitions between targets which are set by a common communicative system (Nolan & Grigoras, 2005; McDougall, 2005). It is unknown how speakers would react if these targets changed, as is apparent for some monophthongs, where speakers conform to widespread changes (as in Harrington et al. (2000a; 2000b; 2005)). Are speaker-specific transition behaviours stable if linguistic goalposts are moved over time?

6.3.2.2 Results

Figure 79 below shows LLR results for all subjects in this analysis, overall and by different formants. The superimposed linear trend lines are purely for illustrating the overall pattern for each speaker, and are not designed to reflect a statistical relationship (they are significant for Neil in the overall, F1 and F3 tests, but nevertheless the data in the figure are not designed for this kind of test). The following sections explain the results in relation to different speakers and different formants.

Figure 79 - Plots showing LR scores for /aɪ/ with increasing delay. Greyed bars represent stages of a verbal scale for presenting LRs, dotted lines show linear trend lines. Reference population – DyViS, N = 100



6.3.2.3 Speakers

This section examines LLR estimates for diphthongs with respect to different speakers. Overall, Symon and Tony show more stable scores than Bruce and Neil. For Symon (green) the stability is remarkable, with all F1 tests falling between 1 and 2.5 and all but the one F3 test falling between 0 and 1. This reflects frequency patterns in the data for Symon, which are the most stable of the diphthong data, both in terms of change of

frequency and extent of gestural movement. There is larger change in Symon's F3, represented by the negative LLR for the longest delay.

For Tony (yellow, insufficient data for a 21 year delay test), all F1 scores are positive, variably between 3.5 and 7, which is higher than the other subjects. This might reflect the fact that Tony (a speaker of a Cockney variety) is less well matched in terms of accent to the reference population than the other subjects, or the fact that his formant frequencies are relatively high compared with most male speakers (as he is physically small with a high F0). The fact that the LLR is especially high for F1 might suggest the latter, as the cockney PRICE nucleus differs from SSBE more in backness than height. Moreover, there is not a similar difference in result for F2, suggesting support for the accent mismatch explanation. These F2 scores for Tony remain stable between 0 and 1. In F3 tests, scores are stable around 1 for the first three tests, but in the 28 year delay test a -1.5 score shows that acoustic changes (in this extraordinary case, frequency increases) have resulted in a negative LLR.

LLR estimates for Bruce (blue) follow a pattern of slight decline with time for each formant. Although for F1 all scores are above 3, and therefore maintain 'moderately strong' support for the prosecution hypothesis, F2 and F3 show negative LLRs after 21 and 28 years delay respectively.

Neil's data (red) show the most significant changes in LLRs, apparent both in the number of data points in the negative and the steepness of the trend lines. For F1, between 0 and 21 years delay there is a steep decline from around LLR 7 to -7, although the 28 year delay shows limited support for the prosecution. For F2 the picture is much more mixed, with no clear increase of decrease and most tests showing LLR 0 to 2. F3 is perhaps the most consistent trend, with a fairly steady transition from LLR 2 at the 0 year delay to -2 and -4 after 21 and 28 years.

6.3.2.4 Formants

This section elucidates the difference between formants' LLR scores for individual formants. F1 results show different patterns to the other two formants, and it is apparent in Figure 79 above that all but Neil's F1 scores show positive LLRs throughout. This is perhaps surprising given that F1 changes, in both frequency and extent of movement are, in general, proportionally larger than for other formants. It may be that

these relatively good results (in terms of stability and strength of evidence) are due to the incorporation of both frequency and slope or extremity information in the diphthong transition curves. It could be that using regression coefficients rather than raw frequency has made the LR estimation more reliable in the face of age-related change by capturing more information about individual's articulatory movements.

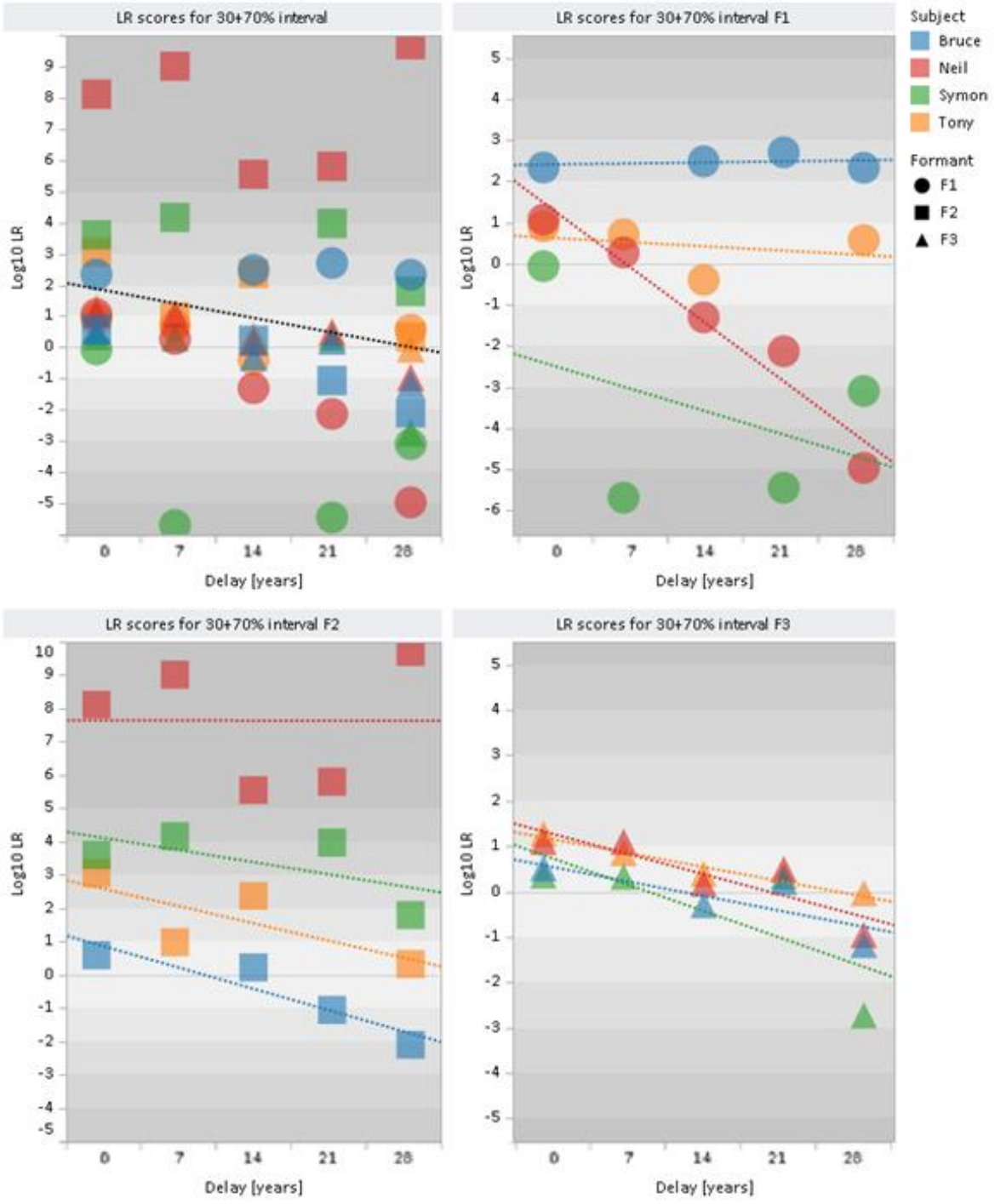
F2 findings are less conclusive. While there are the fewest changes to LLR scores for this formant (only one subject, Bruce, showed a steady decrease across the period), meaning the parameter is fairly reliable, the magnitude of the LLR, and therefore the strength of evidence, is lower than for F1.

F3 follows the most expected pattern, following both raw frequency data and from LLR results for monophthongs. All subjects show a pattern of steady reduction in LLR scores, with varying steepness. At 0 years delay, all tests show positive scores between 0 and 2 and after a 28 year delay all show a negative LLR, between 0 and -2 for all but Neil, at -4. As with F2, the strength of evidence across most F3 tests is lower than for F1. The fact that positive LLRs for both F2 and F3 are lower than F1 is surprising given the postulate that higher formants perform better as speaker discriminants.

6.3.2.5 Comparison with static measurements

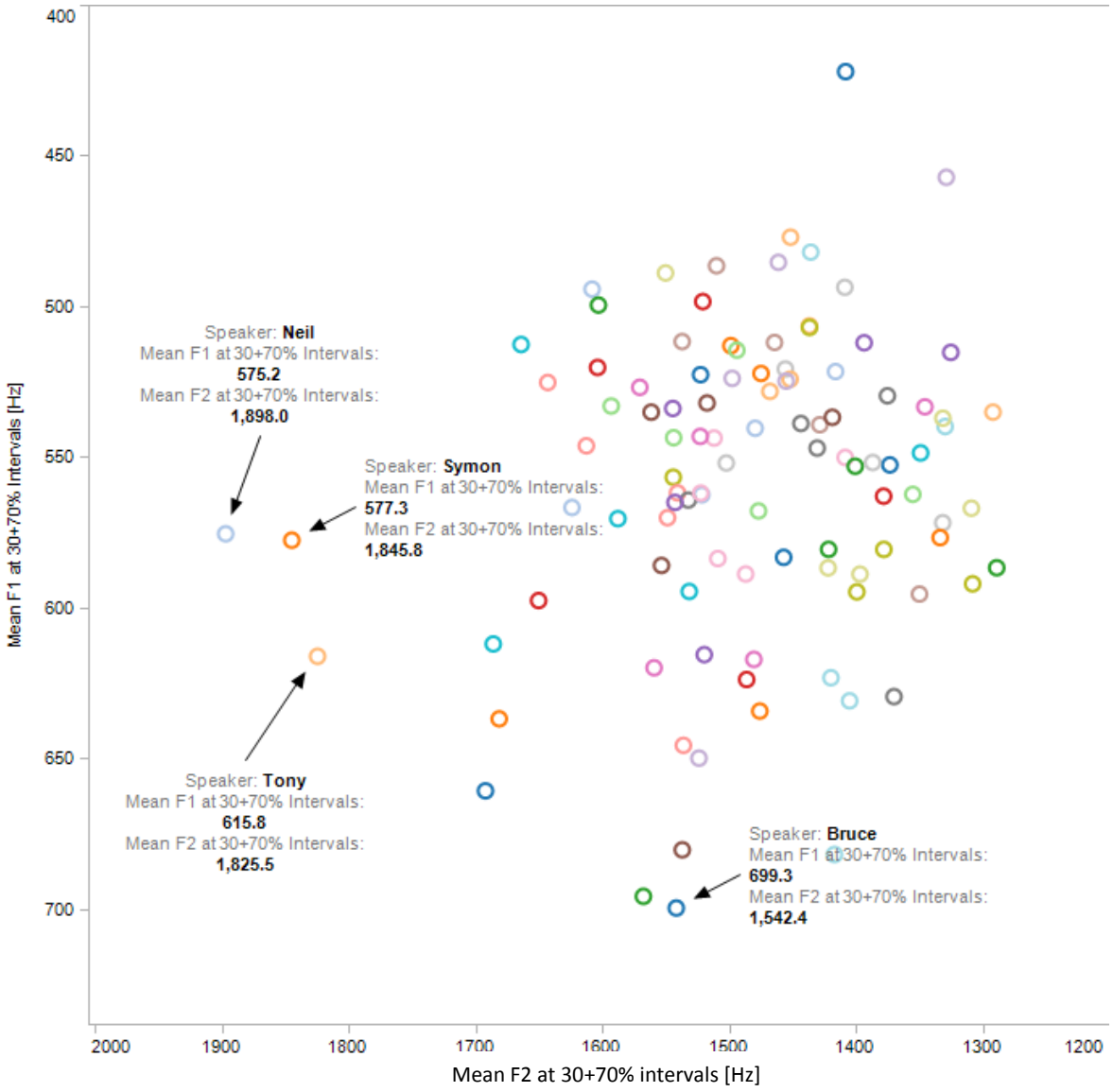
In order to examine the benefit of a dynamic approach, it is interesting to compare the resilience of cubic coefficient parameters with static measure taken at 'target' portions of the vowel. As well as capturing more data, a dynamic approach reportedly captures more individual behaviour. If this is the case, and given the changes to diphthong frequencies in chapter 0, it might be expected that dynamic measures will show greater robustness to time. Figure 80 shows LLR scores for tests using a vector of two measures, at 30 and 70% intervals across the duration of the formant, capturing (roughly) a static measure of target sections.

Figure 80 - Plots showing LR scores for 30 + 70% intervals of /aɪ/ with increasing delay. Greyed bars represent stages of a verbal scale for presenting LRs, dotted lines show linear trend lines. Reference population N = 100



Results are predictable, with declines in performance across increasing delay. The one exception is perhaps in the case of F2 for Neil, which shows variable and strong LLRs throughout. Moreover, F2 for all speakers shows relatively strong support for the same-speaker hypothesis, in spite of what has been observed from F2 changes in the PRICE vowel. This large magnitude is probably due to a mismatch with the reference sample, which can be seen in Figure 81 below:

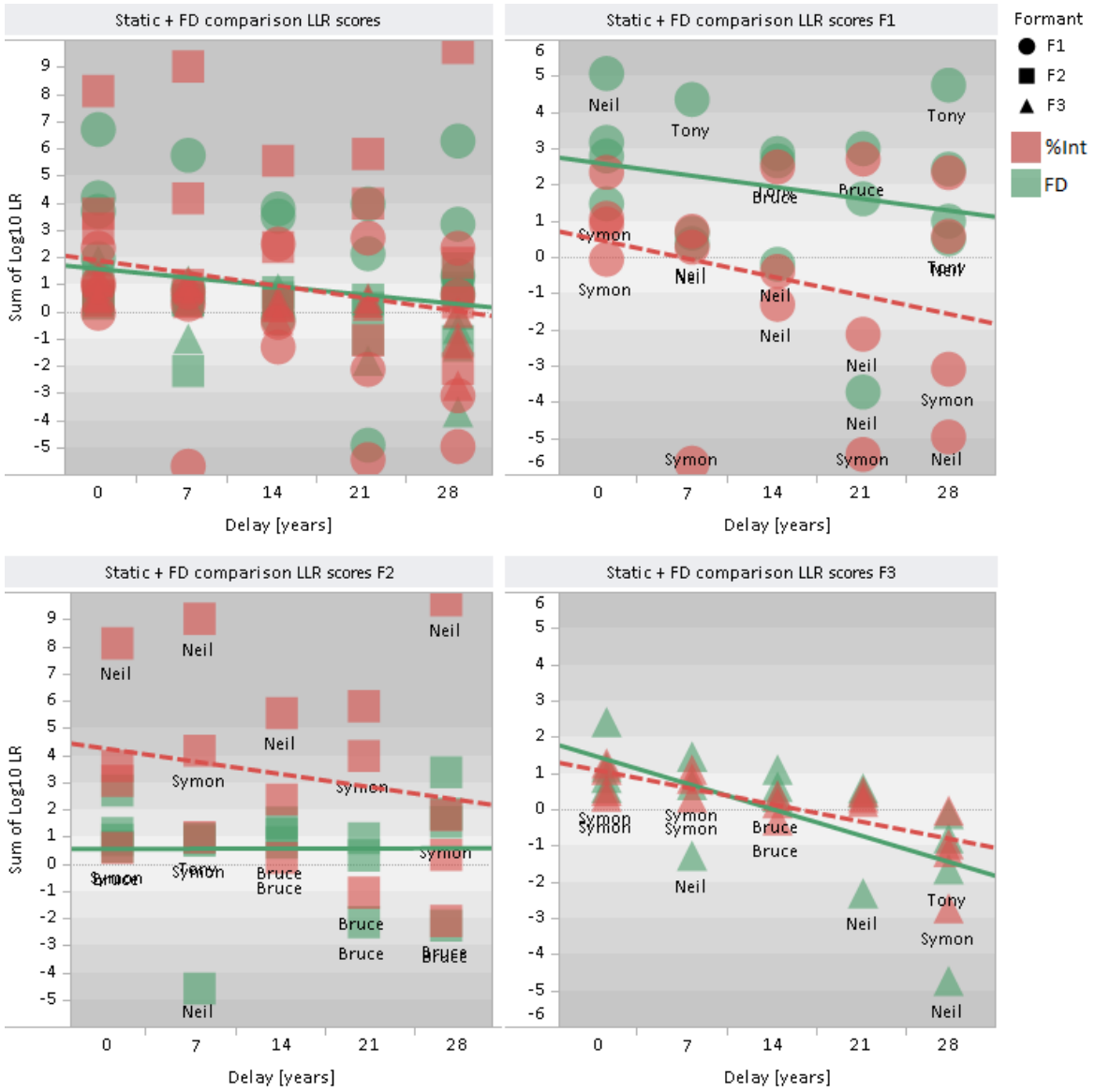
Figure 81 - Scatterplot showing F1 by F2 distributions for DyViS and *21 Up* speakers (aged 21), mean frequency from 30+70% intervals from PRICE diphthong



Although this leads to large strength of evidence, it is the resilience to aging that is of interest, and F2 scores, although high, are still declining.

The purpose of this exercise, however, is for comparison with dynamic measures, and these are illustrated below:

Figure 82 - Plots showing LR scores for dynamic measures (green) and 30 + 70% intervals (red) of /aɪ/ with increasing delay. Greyed bars represent stages of a verbal scale for presenting LRs, lines show linear trend. Reference population N = 100



In Figure 82, tests with dynamic data are in green and those with 30 and 70% interval data are in red. Linear trends are superimposed to show the general pattern, but again are not for direct analysis. There are a few key differences between the two parameters. Firstly, in F1, dynamic tests show much greater resilience to aging, with all but one test correctly showing positive LRs, even after 28 years. Not only are more static tests showing incorrect negative LLRs, but performance is degrading at a faster rate for most speakers.

For F2, however, the picture is slightly different. Although static measures seem to present stronger evidence, this is largely influenced by data for Neil. Without those large scores (due to mismatch with the reference sample), static measures show similar

performance. However, the real concern is the effect of time, and it is fairly clear that dynamic tests show greater stability and resilience to aging.

In F3, both measures are reducing steadily but not greatly, with dynamic measures seeming to reduce slightly more (although again influenced heavily by a few strong scores for Neil). Although this is a limited set of subjects and only one vowel, it does seem that a dynamic approach suffers less degradation with age. This is particularly true in F1, where the greatest impact might be predicted from changes observed in raw mean diphthong frequencies.

6.3.2.6 Summary

Table 35 - Showing log_{10} LR scores for 4 subjects' PRICE vowels across increasing delays, colour formatting is dependent on magnitude of LR (green = positive, red = negative)

Speaker	Delay	0 years	7 years	14 years	21 years	28 years
Bruce	F1	4.5		4.0	4.2	3.4
	F2	0.5		0.4	-1.1	-1.1
	F3	0.7		0.3	0.5	-0.6
Neil	F1	7.1	0.6	-0.3	-5.2	0.7
	F2	1.4	-2.3	0.6	0.5	1.7
	F3	2.0	-1.0	0.5	-1.9	-3.9
Symon	F1	2.1	0.9		2.3	1.4
	F2	0.3	0.4		0.2	0.8
	F3	0.5	0.6		0.4	-0.1
Tony	F1	3.9	6.1	3.7		6.7
	F2	0.6	0.5	0.8		0.9
	F3	1.0	1.2	0.9		-1.3

In general, LLR results for diphthong coefficient data are similar to monophthongs, where LLR scores decrease with increasing delay. There are differences between the diphthong curve and monophthong frequency tests in the amount of reference speakers. The LR calculations for diphthongs exhibit fewer negative LLRs and show stronger and more consistent support for the (correct) same-speaker hypothesis. This is especially true for F1, which might be surprising given the extent of change in frequency and in LLRs for monophthongs. There are two likely explanations for this. Principally, the reference data in the diphthong curve testing are much more comprehensive and likely to produce more reliable results. Furthermore, using regression to model frequency contours, in simple terms, captures more information (about both frequency and slope) and is more likely to produce better discrimination within each vector of the LR test because of this. These ideas are supported by comparisons with static measures, which show dynamic measures

to be more resilient to age-related changes in frequency overall. Although it was shown that both overall frequency and the range of gestural movements were affected by aging, the dynamic approach (which takes both characteristics into account) produced more valid estimates of strength of evidence.

There are, however, limitations to the conclusions we can draw on this basis. Testing in this chapter is intended more as an explanatory case study to investigate these issues, rather than to provide definitive, comprehensive answers about discriminatory power. Further testing is required, incorporating more vowels and with more appropriate comparisons with similar reference data. Within-testing of large databases would also allow for more reliability and validity metrics, providing a more complex picture of both the number and magnitude of errors and what kinds of errors are being made. Resources such as DyViS are ideal for theoretical testing, but in practice there are remaining questions over how well matched these databases should and can be to the test data. This is especially true given the limitations of sampling methods currently used in linguistics, and the social or lifestyle factors which may be likely in general and criminal populations (see §2.1.2). In the case of some of the subjects of the current study, it would prove very difficult to find appropriate reference data which match the social factors of participants' lives, particularly where this would affect speech.

This is (to the author's knowledge) the first study to investigate the effects of aging on the strength of evidence calculated using static and dynamic formant parameters. This limits the comparisons that can be drawn with other studies. Nevertheless, the analyses in this section present findings which build on the previous investigations of age-related changes in the literature and in the present study. It also raises interesting questions about the practical issues of application of overtly numerical LR approaches with regards to mismatch between evidential, suspect and reference samples. These need to be addressed before the approach can be adopted in practice, and before reference databases for practical use can be designed.

6.4 **Discussion**

This section addresses research questions in light of the likelihood ratio estimates presented above. It expands on question 3c, where the data from chapter 0 could not directly address the question. It also uses this data to illustrate what effect age mismatch

has on LR estimation, and extrapolates from that to consider the logical and practical implications of any kind of mismatch on this approach towards estimating the strength of evidence.

6.4.1 **Research question 3c**

3 What is the discriminatory performance of FD measures of different polyphthongs in different varieties of English in spontaneous speech?

c Is making dynamic measures of vowel formants worthwhile?

Further evidence from LR estimation testing shows that making dynamic measures of diphthongs is worthwhile in cases of long-term non-contemporaneity. Despite changes to diphthong frequencies and ranges of gestural movement with age, LR estimations with cubic coefficients from PRICE vowels show, for the most part, correct positive LLRs, particularly for F1 and to a slightly lesser extent for F2. Put simply, despite there being consistent age-related changes to acoustic properties of diphthongs measured separately, formant dynamic parameters seem to show greater resistance to the effects of age in estimating the strength of evidence.

Further testing with more vowel categories, languages and larger test databases (including different-speaker testing) is required. Furthermore, comparisons between different parameters would give a more holistic view of their forensic usefulness in the face of different factors (such as aging). Despite this, it seems that, for these speakers, there may be an element of gestural behaviour that a dynamic method captures that is recoverable even in the face of predictable physiological modulation with age. Nolan (1983) postulates that physiology constrains the range of articulator movements, and despite changes to those constraints, speakers' individual articulatory strategies (Nolan, 1997; Nolan & Grigoras, 2005; McDougall, 2006) seem to remain somewhat preserved.

6.4.2 **Research question 4**

4 What effect does age mismatch between evidential and suspect recordings have on LR estimations?

It is clear from the findings of this study that speakers' acoustic outputs are changing over time. Logically, it is also clear that if samples are more different then LRs necessarily show less strong support for the prosecution hypothesis. The purpose of the cases

presented in this section was to assess the extent of these changes in LRs and elucidate how these effects would be manifested in a practical method for estimating strength of evidence.

Clearly, aging affects LRs significantly. LLR estimates based on contemporaneous data correctly supported the same-speaker hypothesis in almost all cases. Tests across increasing delays generally lead to a decrease in both the strength of evidence and the number supporting the correct hypothesis. This effect was not universal, but after 21 or 28 years the majority of tests showed erroneous support for the different-speaker or defence hypothesis. This was also variable for different formants and vowel categories, which (generally) mirrored those differences found in previous acoustic results. Again, although this was a general pattern, it was not complete, as the LR score is determined not only by the relationship between the two samples but also the relationship with the reference population.

This raises an interesting and important question regarding this approach towards formulating strength of evidence estimates: how do we match a reference population to test samples where there are (sometimes expected or predictable) mismatches between suspect and offender samples? In the UK position framework (French & Harrison, 2007) the analyst would take expected social or other variability into account, based on their knowledge and experience. How is this possible within an overtly numerical LR approach, without necessarily biasing a test towards either hypothesis? It might be possible to perform normalisation for age, but these effects are not always present or consistent (cf. consistent acoustic data from the Yorkshire Ripper Hoaxer case (French, Harrison, & Windsor-Lewis, 2006)). Moreover, there are other social factors which might lead to changes in different directions to predictable physiological differences, as the present study has demonstrated.

6.4.3 **Mismatched samples in LR estimation**

This age-related question raises a general issue with the likelihood ratio approach with regards to both testing data and the selection of reference population data.

6.4.3.1 Differences between testing data

Although non-contemporaneity of the extent apparent in this study may not be fairly commonplace in forensic practice, there are a number of factors which are highly likely to

vary between suspect and evidential samples. Some of these can be dealt with at the level of test data; for example if one sample was transmitted by telephone, samples can be filtered to try to match transmission conditions. However, there are multiplex issues at hand with potential for numerous speaker, channel and other factors (see §1.1.3) which can influence speech or interfere with the recording and/or transmission of acoustic data. Furthermore there are many factors which might be at play where the extent, effects or even existence of that factor are unknown to the analyst. Although this might not be a problem for aging specifically, as it is normally possible to identify the age of a suspect and therefore their potential age at the time the evidential recording was made, it is an issue for other factors.

As with the aging example discussed above, if there is an expected mismatch between samples, how does an LR approach deal with expected variability between samples, other than excluding those features or parameters from strength of evidence estimation?

6.4.3.2 Reference databases

The second key question in relation to between-sample mismatch is how to select an appropriate reference database. The purpose of this is to provide statistics on the distribution of data, the typical inter-speaker variability, in a relevant group of speakers. In theory the reference population is delimited by the concept of a potential perpetrator population, that is to say, the group that has the same characteristics as the person in the evidential sample which is in some cases further defined by the defence hypothesis.

For instance, for a DNA test, it might be that an eye-witness could identify ethnicity, narrowing down the potential perpetrator population. The defence hypothesis might be 'it was not the suspect who left the trace, but someone else of that ethnicity'. In this case, this factor would delimit a population to be sampled for or selected from a reference database. The same is true with speech data, where discoverable facts about the speaker (such as sex) might delimit the formation of a hypothesis. In speech, however, this might be more problematic as there are much more complex interactions of factors which could impact on speech than for DNA (as discussed earlier in §2.3.5). In cases where there are differences in the factors which delimit the reference population, or significantly affect the data in the LR estimation, the picture becomes even murkier. If there are factors in a case (such as aging) which mean we would expect there to be variation between testing samples *for the same speaker*, how is this taken into account

when selecting a reference population which matches the properties of the evidential sample?

Using the case studies above and considering a 28 year delay as an example, a reference population such as the piloted Deterding (1997) data might be useful for assessing the typical variability of an acoustic parameter in a reference sample that matches speakers of SSBE in late 1970s. Although it matches the potential perpetrator limitations defined by the offender recordings, if this is being used as a measure of variation in SSBE speakers in 2006, as in the suspect data, this cannot be reliable. The same problem is true for age in the DyViS database, where 18-25 year old speakers roughly match the age of the offender recording, but not the 49 year old suspect speakers. Time is not the only factor in the present study for which this issue might arise. How should an analyst sample the relevant population for a speaker who is geographically or socially mobile where a linguistic parameter is known to vary between two locations or social groups/classes? These are just two examples of factors that are present in the current data, there are many more which play a part in linguistic and acoustic variation.

6.4.4 **Specificity and composition of reference databases**

These examples, and the findings of this study, give rise to a further unanswered question with serious practical implications for the implementation of an LR approach. In order for analysts to be able to perform LR based assessments of speech evidence, large scale databases of reference recordings need to be available. It was discussed in §2.3.1.1 that funding for forensic examination in the UK is limited and decreasing, and it is impractical to expect analysts to be able to collect bespoke databases in each case. For DNA, the national database houses large amounts of reference data, indexed by the few factors which influence DNA composition, such as gender and ethnicity. This is successful and feasible as DNA is largely inflexible within an individual. Moreover, put into perspective, strength of evidence from DNA is generally much stronger than for speech evidence, given this inflexibility, and therefore can attract greater funding. Given that speech is demonstrated to vary within an individual according to many factors, what amounts and specificity of different data need to be collected to represent a suitable reference database, indexed by how many factors?

Taking age as an example, findings from the present study show that aging effects significantly affect acoustic parameters commonly used in forensic practice. Logically we

should assume that, on average, populations at different ages are likely to have different distributions of acoustic parameters. Alongside the current findings, acoustic studies of formant frequencies have shown differences between younger and older speakers, within the same speakers (Endres, Bambach, & Flösser, 1971; Linville & Rens, 2001; Harrington, Palethorpe, & Watson, 2007; Reubold, Harrington, & Kleber, 2010) and in different groups (Linville, 2001; Watson & Munson, 2007). This idea is also supported by vocal tract measures (Xue & Hao, 2003) which indicate that elderly groups have greater volume and oral tract length than younger speakers; these differences will impact on measures of resonant frequencies and other acoustic parameters, such as MFCCs. Therefore, a reference database composed of samples taken from 21 year olds may not be suitable for comparisons of recordings from 50 year olds, for example. Furthermore, there seem to be differences in how different parameters and testing methods are affected. While, in general, results from this chapter show evidence supporting the incorrect hypothesis after a delay of around 14-21 years, results for a GMM-UBM system parameters (Kelly & Harte, 2011; Kelly, Drygajlo, & Harte, 2012) show significant degradation in performance over 5-10 years. Should databases then collate data from different parameters at different age intervals (in this case age-matching for formant data need not be as specific than for GMM-UBM), or simply keep the recordings so that analysts can select appropriate recordings and analyse them case-by-case? Aging issues are further obfuscated by the fact that aging is not a linear, nor always predictable, process, and there are developmental stages which affect physiology more significantly, such as menopause or puberty. These issues are not prevalent for DNA, which does not vary over the lifespan.

Of course aging is not the only factor which affects speech, and the number of social, biological, environmental and other factors which can affect forensic recordings is extensive. Similarly to age, how closely do these factors need to be replicated in a reference database to present sufficiently accurate strength of evidence estimates? Is a forensic speech database which allows for an accurate match according to all the factors that are recorded in forensic cases something of a pipe dream, given that each new factor would multiply the number of recordings required? This is much less problematic for DNA, where the database is indexed against ethnic background and gender. Is a speech database where recordings could potentially be indexed by factors such as gender, class, age, height, time of day, intoxication/drug intake, accent, stress, recording quality (to

name a few) prohibitively large and detailed, given the possible permutations of factors? Add to this the propensity of accent types to change and adapt, and would this database have to be reviewed and re-recorded over decades? In theory, such a database could be out-of-date by the time it was recorded.

In practice, this could be streamlined by the types of factors most likely in a forensic case, or by broadening some of the categories, but this presents problems where this might exclude factors present in a case. Furthermore, for a (presumably) publically funded resource, there are potential political problems in targeting specific groups for a forensic speech databases. There are clearly many questions which remain unanswered about reference database composition and specificity. Research into these areas is required, especially before funding is pushed towards large scale, expensive and inadequate speech databases for actual forensic casework.

7 ASR analysis using *BATVOX*

This chapter discusses findings regarding aging effects on performance of the *BATVOX* ASR system. The analyses investigate age-related changes in speakers' vocal apparatus, reflected by changes to MFCCs, which may be forensically relevant. Age mismatch between suspect models and reference data for calculating LRs also comes under scrutiny. This chapter, therefore, expands on the responses to research questions 2 and 4 (see §2.4).

7.1 ASR testing

Following the methodology of the previous chapter, ASR tests are made with different magnitude delays and illustrate the effects of age differences between evidential and suspect-type recordings for the same subjects. While LR tests are configured as direct suspect-to-evidential sample comparisons, the *BATVOX* system generates a model of the suspect distribution which is then tested against evidential material. This means that tests between different age groups are different depending on whether the later or earlier age has been used as the model or the test data.

To simplify the two types of test, this chapter adopts the distinction used in Kelly and Harte (2011), where the authors differentiate between a verification model, where the earlier known model is tested against data from a later time, and a comparison model, where the evidential or test material precedes the suspect model. This would be relevant if you were to train a verification system, and then make subsequent identification requests (this is also referred to as a 'forwards delay', as the delay comes after the model). It is also possible that a speaker comparison could follow this framework; in a case where evidential materials arises and is tested against historic police interview recordings, or those recorded for other reasons (perhaps in the media). In the more likely comparison model, however, evidential samples are made first and reference recordings come from materials from a later time. The Yorkshire Ripper Hoaxer example described in the first chapter is an extreme example of this, where the known suspect recording (in this case the basis for the ASR model) was made 26 years after the test data. This is also referred to as a 'backwards delay'.

The methodology for these tests can be found in §3.4.2.7. 100 speakers of SSBE from the DyViS corpus (Nolan, McDougall, de Jong, & Hudson, 2009) were a reference population

from which *BATVOX* selected an optimised 35 speakers for each test, mirroring procedures used in a practical setting. Unlike in the previous chapter, comparisons at the same age were not made, as the channel similarity combined with the fact that some speakers did not have an even spread of recordings from different times and environments meant that a 21 to 21 year comparison (for example) would present greatly inflated LR scores (Rose, 2002; Enzinger & Morrison, 2012) and would not be a fair control test. Moreover, in any case, most 7 year delay tests reached the maximum performance of the system. Much like in the previous chapter, the limited number of subjects prevents this study from making error-rate and -type assessments, and therefore wide-ranging generalisations about aging and the validity of the systems in light of these processes. What this chapter does is investigate aging effects as a case study and uses the results to frame further questions, corroborate previous findings and identify directions for more comprehensive research. This chapter continues to present Log_{10} LRs (LLRs) unless otherwise stated.

7.2 ASR results

This section presents results from the ASR system for each of the 6 male speakers that were subject to analysis. The figures in this section are in the following matrix format:

Table 36 - Example results table for ASR results

Age for test data (evidential sample)

Age for model (suspect sample)	Speaker	21	28	35	42	49
	21	No test	7 year forwards delay	14 year forwards delay	21 year forwards delay	**28 year forwards delay****
	28	7 year backwards delay	No test	7 year forwards delay	14 year forwards delay	21 year forwards delay
	35	14 year backwards delay	7 year backwards delay	No test	7 year forwards delay	14 year forwards delay
	42	21 year backwards delay	14 year backwards delay	7 year backwards delay	No test	7 year forwards delay
	49	**28 year backwards delay***	21 year backwards delay	14 year backwards delay	7 year backwards delay	No test

* Extreme example for comparison type delay (i.e. Yorkshire Ripper Hoaxer)

** Extreme example for verification type delay

Horizontal rows represent each training age and the vertical columns correspond to each test age. Therefore those cells below and to the left of the 'no test' cells represent the comparison type tests or backwards delay, and those above and to the right represent the verification type. For this study and for FSC in general the lower rows and the 49 row in particular are probably the most interesting; for example the bottom left cell is the most extreme example of a non-contemporaneous FSC case, similar to that of the Yorkshire Ripper Hoaxer (R v John Samuel Humble, 2005; French, Harrison, & Windsor-Lewis, 2006). In the following matrices, blue represents positive LLRs while red indicates negative LLRs, the level of shading is dependent on the magnitude of the LLR. Values of 10 are the maximum that are given by this version of *BATVOX* (Log_{10} value of 10 equates to LR = 10 billion). Although a number of the LLR scores presented here are extremely large, it is worth remembering that both the model and test data are of excellent technical quality, matched in terms of channel and do not feature many of the other limitations normally found in the forensic condition.

7.2.1 **Predictions**

Given the results relating to other parameters presented in this study, and the GMM-UBM system testing that has already been performed with other longitudinal data (Kelly & Harte, 2011; Kelly, Drygajlo, & Harte, 2012), it would be sensible to predict a fairly steady decrease in performance over the period. Findings also expand on the results of Künzel (2007), which concluded that (an earlier version of) the same system was almost completely robust to an 11 year delay. Furthermore, tentative results in Suzuki et al. (1996) suggest that cepstral coefficients are likely to change with age.

If this degradation is apparent, the matrices below will show fewer dark blue cells and even negative LLR red cells after the longer delays (i.e. in those cells furthest from the centre diagonal). Results of further interest highlight the magnitude of changes (more than 10 years was generally significant in some of the studies above), any different patterns for differences between forwards and backwards delays, effects of training models at different ages (the difference might evince effects of age mismatch between model and reference population) and, perhaps most importantly, individual speaker effects.

7.2.2 **Andrew**

Table 37 - Matrix showing LLR scores for Andrew for *BATVOX* tests between each model and test age

Age for test data (evidential sample)

Age for model (suspect)	Andrew	21	28	35	42	49
	21		-0.48	-0.91	-0.95	0.16
	28	1.66		9.07	7.19	6.77
	35	0.42	3.28		6.71	10.00
	42	1.37	7.75	10.00		5.97
	49	0.70	1.93	7.39	1.89	

Table 37 shows that, as predicted, data for Andrew show degradation in magnitude of LLR with increasing years of delay. Those cells closer to the centre diagonal, with the smallest delays, show moderate to very strong support for the correct hypothesis, up to the inherent *BATVOX* maximum in two cases. In contrast those tests further from the centre, especially with the age 49 training model, show lower LLRs. These would fall around the 'limited support' boundary in a verbal alternative scale. Three of four results with age 21 training models show false support for the different-speaker hypothesis, and those tests with 21 test data generally show weaker support for the correct hypothesis. This is probably because the 21 sample for Andrew was the shortest (of his data, and across all speakers), but this was still above the minimum threshold for *BATVOX* and well within a normal range for forensic material at just over a minute net speech (particularly for test or evidential data). For this speaker, there are a number of errors, and an overall pattern of decreasing performance with longer delays.

7.2.3 **Bruce**

Table 38 - Matrix showing LLR scores for Bruce for *BATVOX* tests between each model and test age

Age for test data (evidential sample)

Age for model (suspect)	Bruce	21	28	35	42	49
	21			4.78	4.00	0.88
	28					
	35	4.47			8.74	8.83
	42	2.63		7.53		10.00
	49	1.86		9.52	10.00	

Data for Bruce (although limited by the absence of reliable 28 data) show a consistent pattern following the prediction of reduction in LLR, with no miss errors and no tangible difference between forwards and backwards tests. LLR scores reduce from the maximum in 7 year delay tests to 'limited support' values around 1, and the effect of age seems to be steadily increasing (within the logarithmic scale). The effect of age on ASR performance for Bruce's data is very strong and steady; this is in line with the changes in Bruce's previous acoustic and LR test data.

7.2.4 **Neil**

Table 39 - Matrix showing LLR scores for Neil for *BATVOX* tests between each model and test age

Age for test data (evidential sample)

Age for model (suspect)	Neil	21	28	35	42	49
	21		4.05	7.11	0.96	1.36
	28	9.68		10.00	9.81	4.33
	35	8.76	10.00		10.00	10.00
	42	3.73	6.00	10.00		10.00
	49	6.09	2.89	10.00	10.00	

Neil's recordings contain the largest quantities of net speech, so we might expect stronger LLRs and most consistent changes to those scores. There are maximum LLR scores around the smallest delay tests. Generally those tests with a 7 year delay (in either direction) result in very strong support for the same-speaker hypothesis. Those tests with

14 years show almost the same level of strength of evidence. Tests with 21 or 28 year delays show a drastic decline in strength of evidence, with scores ranging between around LLR=3-6 for comparison tests and LLR=1-4 for the verification tests. The overall picture matches the prediction, with reducing performance in general, particularly strongly after delays greater than 14 years.

7.2.5 **Nick**

Table 40 - Matrix showing LLR scores for Nick for *BATVOX* tests between each model and test age

Age for test data (evidential sample)

Age for model (suspect)	Nick	21	28	35	42	49
	21		9.78	3.70	9.98	0.14
	28	8.44		10.00	10.00	9.99
	35	7.50	10.00		10.00	10.00
	42	10.00	10.00	10.00		10.00
	49	1.20	10.00	10.00	10.00	

Tests for Nick are almost all at the maximum LLR with the exception of 21-35 comparisons (in both directions) and the longest delay (in both directions). It seems Nick is relatively stable over short term delays, and is atypical with respect to the reference data. VTLe bears this latter observation out, estimating Nick's vocal tract length at between 16.5 and 17 cm, generally the longest of the subjects (apart from Symon at two stages). It could also be that the strength of evidence throughout tests is due to a difference in habitual vocal setting that Nick adopts after moving to America, putting him apart from the SSBE speakers of the reference database, leading to inflated LLR scores.

This issue needs testing formally using voice quality analysis tools, and with larger and more varied datasets, especially considering the findings of Harrison and French (2010) with regards to the effects of accent variation on ASR performance. It is known that speakers and learners of different languages use different vocal settings (Gick, Wilson, Koch, & Cook, 2004; Wilson, 2006) and that this should be directly measurable (Schaeffler, Scobbie, & Mennen, 2008; Mennen, Scobbie, de Leeuw, Schaeffler, & Schaeffler, 2010). However, there is no research which has investigated the ability and

disposition of (specifically) mobile speakers to adopt host vocal settings, which might inform this explanation.

Given the lower LLR scores for the 21-35 samples, it is likely that those tests are the furthest apart in terms of speaking style or channel. They are not different enough from the remaining samples to reduce performance in other comparisons, however. For the 28 year delay tests, performance is reduced by a considerable degree, with scores of 1.2 (comparison) and 0.14 (verification). Clearly for Nick there is something of a cliff-edge effect for the longest delays, where there is a rapid decline in the system's performance.

7.2.6 **Symon**

Table 41 - Matrix showing LLR scores for Symon for *BATVOX* tests between each model and test age

Age for test data (evidential sample)

Age for model (suspect)	Symon	21	28	35	42	49
	21		1.58		0.71	-0.89
	28	3.78			10.00	10.00
	35					
	42	0.97	7.22			9.91
	49	0.45	9.96		10.00	

Scores for Symon's tests follow the general pattern, but vary slightly from the other speakers. Initially it is clear that increasing delay is degrading LLR scores, especially in those extreme cases (with the verification example showing a negative LLR). Nevertheless, results for all age 21 model and test comparisons shows by far the weakest strength of evidence. Almost all other tests show near-maximum LLRs. Unlike in the case of Andrew, the amount of net speech for Symon's 21 recording is comparable with the rest of the tests and other speakers' recordings. Clearly then, Symon's MFCC distribution is very different in the 21 recording from the other samples. Even within the 21 tests though, increased delay causes reduced performance, with non-evidential values after 21 and 28 year delays.

7.2.7 **Tony**

Table 42 - Matrix showing LLR scores for Tony for *BATVOX* tests between each model and test age

Age for test data (evidential sample)

Age for model (suspect)	Tony	21	28	35	42	49
	21		10.00	10.00	10.00	6.29
	28	10.00		10.00	10.00	0.73
	35	10.00	9.99		10.00	10.00
	42	10.00	7.15	10.00		10.00
	49	10.00	4.61	10.00	10.00	

Tony's data present consistently large LLRs in almost all tests, with only a few of the longer delay tests exhibiting less than the maximum LLR. This is perhaps to be expected given that the MFCC data reflect vocal tract dimensions of the speakers; VTLe for Tony suggest that he has a relatively short vocal tract (the shortest for males in this study for three of the age stages). It is apparent from the documentary that Tony is also relatively diminutive in terms of physical height, and this has been shown to correlate somewhat with vocal tract and articulator proportions and therefore acoustic outputs (Künzel, 1989; Huber, Stathopoulus, Curione, Ash, & Johnson, 1999; González, 2004). Therefore it is probably not surprising that data designed to map the vocal tract as a biometric have identified Tony as atypical of a population of 100 18-25 year old males. Furthermore, Tony is the furthest from the population in terms of accent; similarly to Nick's data, it is difficult to interpret the effects of this, as ASR systems function, in principle, independent of accent. This is even harder to assess especially given the difference in both their vocal tract length estimates.

Tony's stature does raise an interesting question in terms of matching those characteristics which are known to affect vocal tract proportions within the reference population. Even if height and VTL correlations are weak, they still have an effect and perhaps should be taken into account even before *BATVOX* selects an optimised reference sample. Overall almost all of the tests for Tony present maximum LLR scores, probably as a result of mismatch with the DyViS speakers in terms of VTL (and perhaps accent).

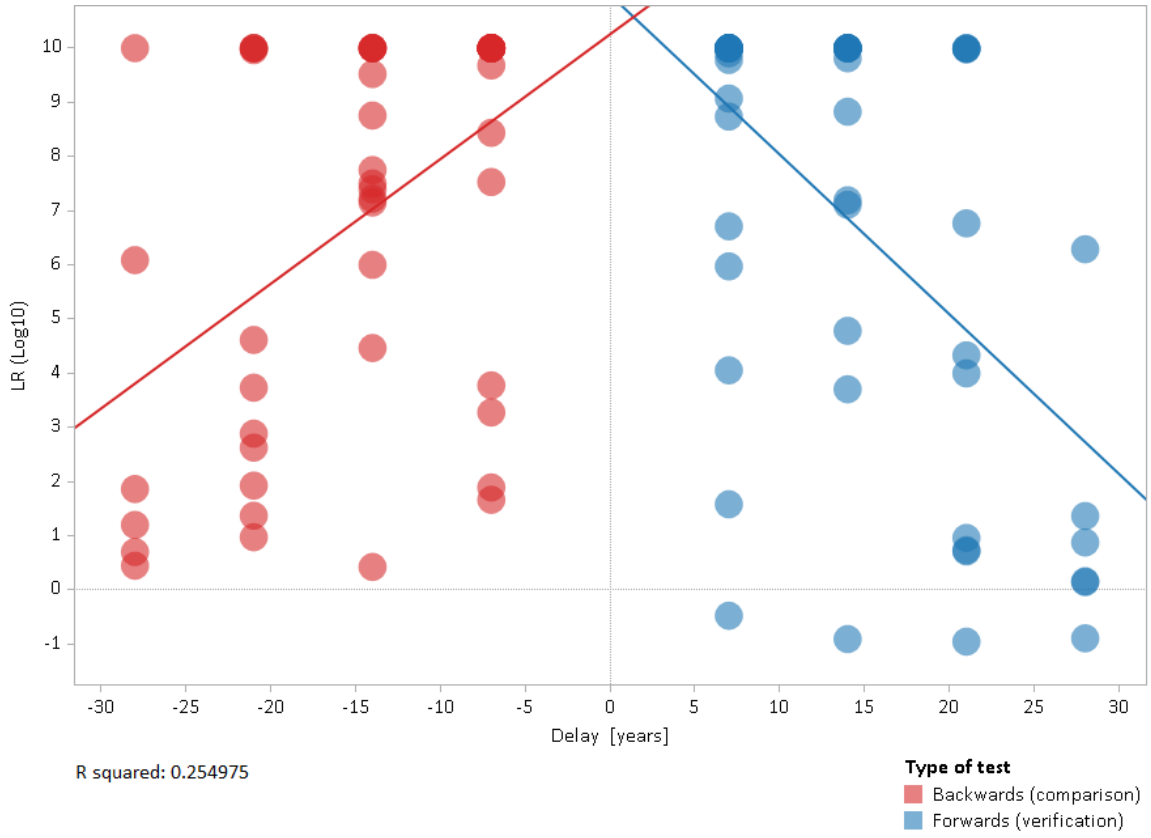
7.2.8 **Overall summary**

The results for speakers above demonstrate that age has a significant effect on strength of evidence estimates generated from MFCC data. Moreover, this effect seems to vary between speakers, being fairly steady for some speakers, while for others (generally those with higher LLRs overall) there is something of a cliff-edge effect after 21 or 28 years.

7.2.8.1 Age and performance correlations

This section presents an overview of all tests and averages across different lengths and types of delay to illustrate the general pattern. Figure 83 below shows the average result for each speaker, across each type of test. For example, all 7 year delays in each test type for each speaker are averaged to a single value, (whereas each speaker has only one 28 year delay in each test type). Figure 83 shows the pattern that was observed in the previous section, that increasing age affects LLR scores negatively. What it also reveals is that while there is degradation in performance between 7 and 14 years delay, the largest shift comes between 14 and 21 years delay. In general tests with a 7 or 24 year delay have LLRs falling between 4 and 10, while the majority of those in the 21 and 28 year delay conditions are between 0 and 5 (with a number of scores at 10 in the 21 delay condition). The quick drop off for some speakers after 28 years can also be seen, with the high frequency (intense circles) of tests at maximum LLR=10 not present at this longest delay stage.

Figure 83 - Log_{10} LR average value for each speaker in each type of test across different delays, with linear trend-line (N tests = 112)



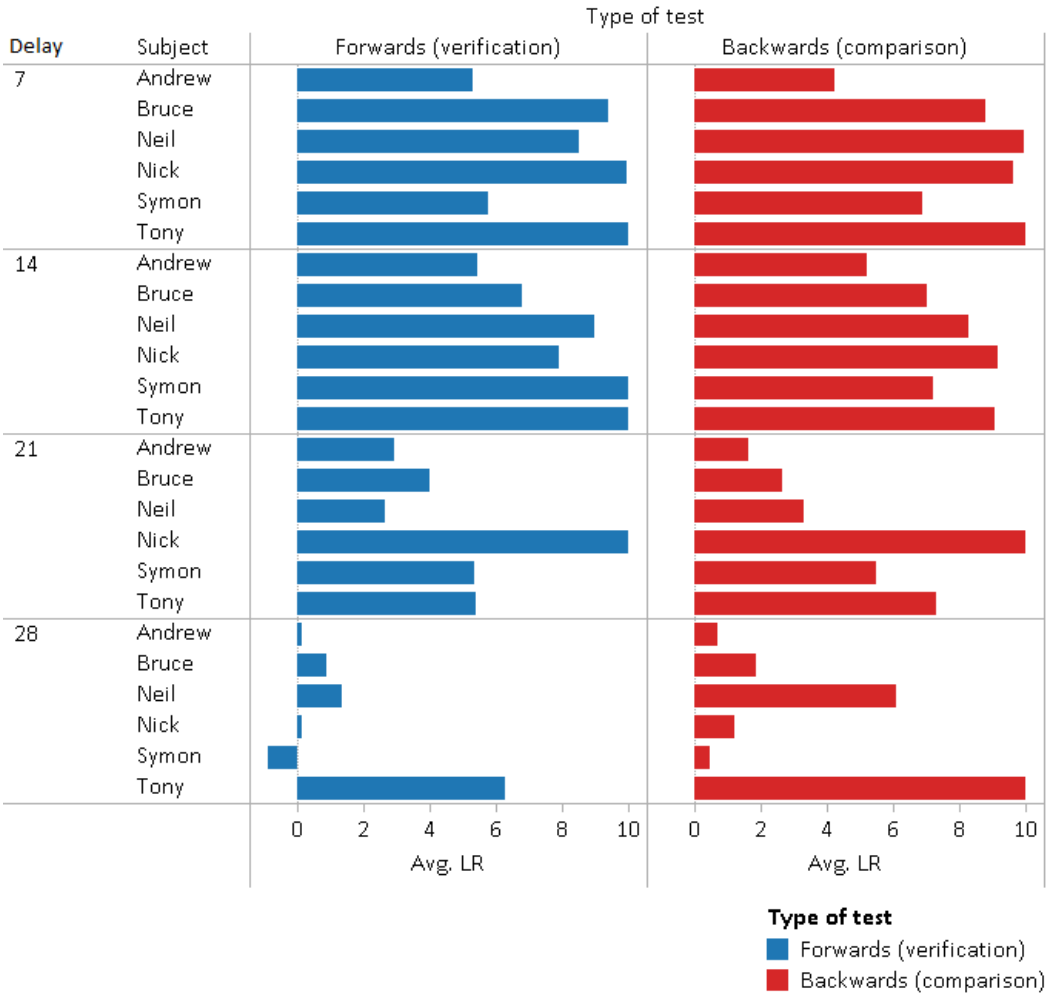
What Figure 83 also demonstrates is that the significant effects of aging on LLR scores for MFCC data can be described (fairly weakly with overall R squared = 0.25, $p < 0.0001$) in a linear relationship. This means that about a quarter of the variation is explained by a statistical model of aging effects. Bearing in mind that these strength of evidence estimates are on a base-10 logarithmic scale, this relationship represents huge shifts in the strength of evidence estimate. What this also demonstrates is that, in general, performance in the two types of tests is fairly similar (R squared is 0.22 for comparison and 0.28 for verification). Although the errors fall only in verification type tests, and R squared is slightly stronger for this type of test, there are other reasons for these results, mainly based on the limited quantity of net speech in Andrew's age 21 data (discussed in §7.2.2).

7.2.8.2 Differences between test types

Any significant difference between the two formats of tests might reveal an effect of mismatch between test and reference data used by the ASR. While the forwards model matches the verification data for age (21 against 18-25 year old reference speakers), the longer comparison models have a mismatch. With changes in vocal anatomy expected,

MFCCs which reflect this might be predicted to vary between mismatched reference and test data. However, the figure below shows more specific results across each type of test with average LLR for each speaker, and shows that in general the types of test are performing similarly across delay conditions and also within each speaker at each condition. This similar level of performance is probably to be expected given the relatively long and well-matched samples, and would probably not be apparent in testing with actual forensic material.

Figure 84 - Log10 LR averages for each speaker across different delays and types of tests (N tests = 112)



7.2.8.3 Comparison with previous studies

There are two groups or authors who have investigated the effects of age on ASR systems, this section will compare results from this study with those previous studies.

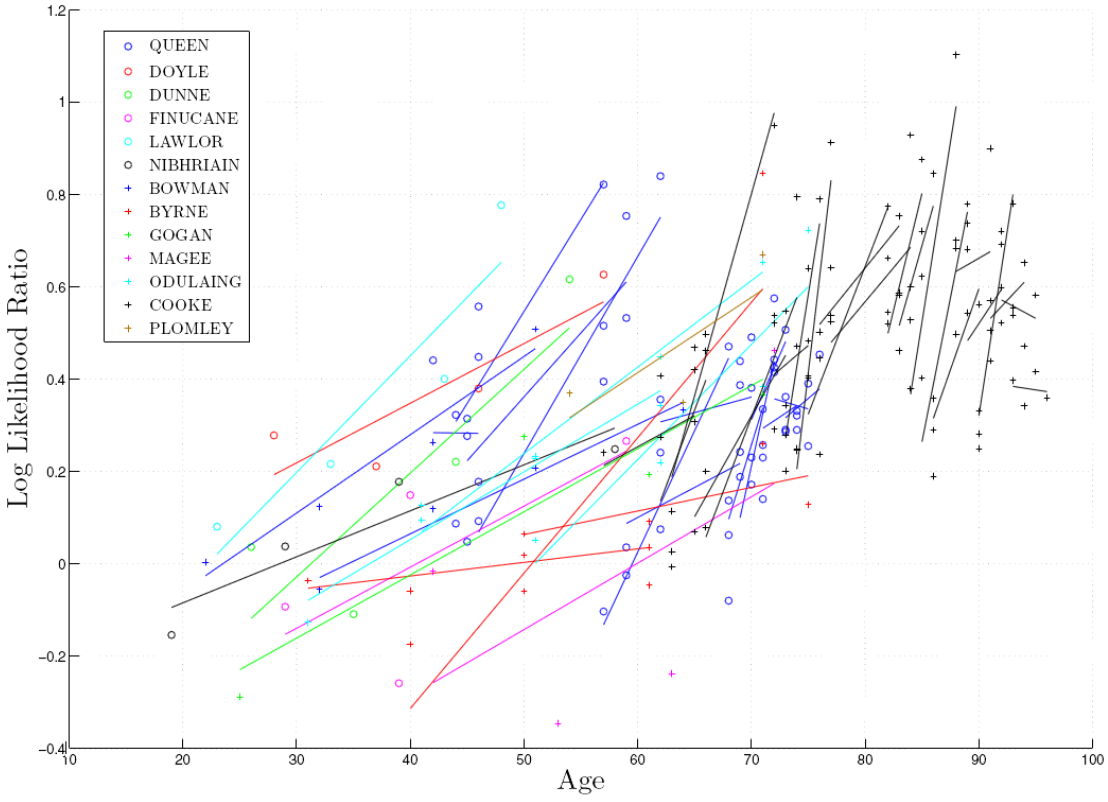
Künzel (2007) investigated the performance of an earlier version of the same *BATVOX* ASR system across 11 years in ten male German speakers aged (in the first recording) between 21 and 51. He found that LRs degraded somewhat for most speakers and very sharply for

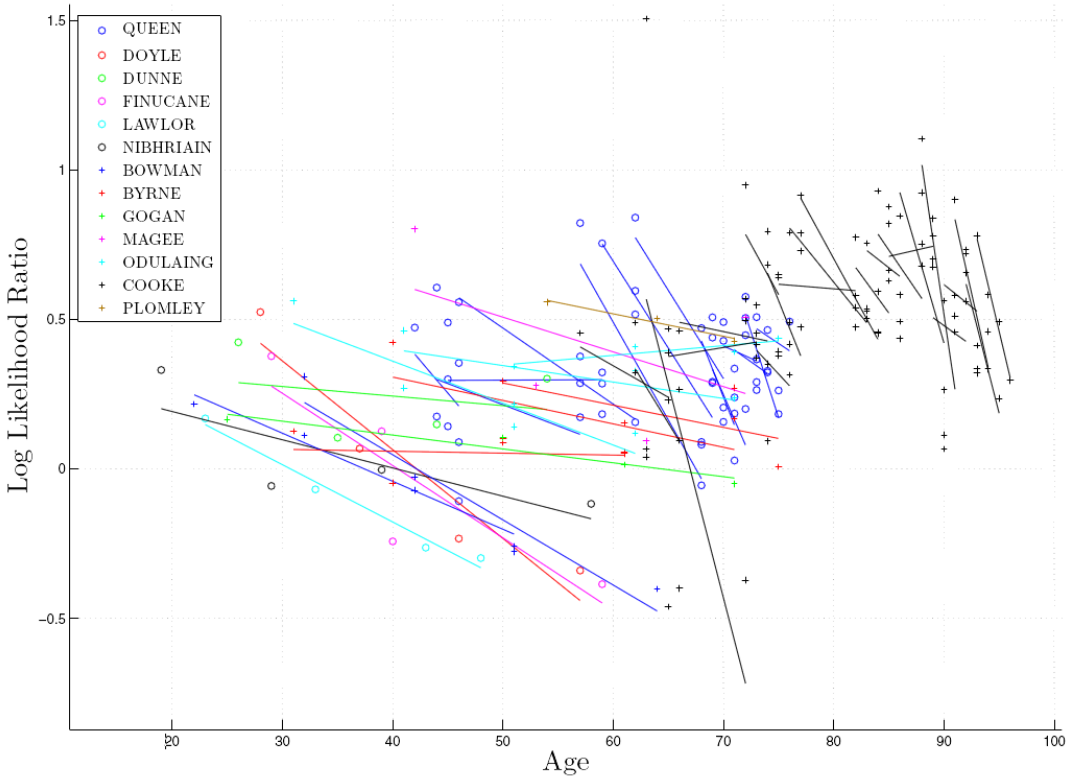
one speaker (who gave up smoking). However, the majority of reductions of performance were minimal and the paper concluded that:

> Automatic speaker identification with the same set of non-contemporary voices shows that an 11 year delay has almost no influence on the identification performance. (Künzel, 2007, p. 131)

Research by Kelly et al. (Kelly & Harte, 2011; Kelly, Drygajlo, & Harte, 2012) presents different conclusions for 18 speakers over a 30-60 year delay. Different speakers are represented by different colours and tests between ages are represented by lines. As with the Künzel study, age differs across subjects. ASR LLRs show linear decline with age over the period, in both backwards and forwards type tests:

Figure 85 - Backwards and Forwards delay ASR testing with 13 speakers





Source: Kelly and Harte (2011)

Clearly performance is degrading, much like in the present study, and seems to continue after the 27 year delay investigated by the present study. What this also illustrates is the difference across different age groupings, with seemingly steeper reductions in performance after 50-60. Reductions are present throughout the 21-49 period and differ between speakers.

There are differences between the current research and these studies. Although Künzel (2007) uses the same system, the ASR tested in Kelly et al is different (although broadly a similar structure). The ages of speakers in the earliest sample differ, and lengths of testing samples are different (30s for Kelly et al., unknown for Künzel). The available reference sample in Künzel is fairly similar (although in German), with 90 speakers from a large corpus. For Kelly et al. the sample is universal, even across genders, with speakers from a number of corpora. However, in all cases, the system selected an optimised reference population for each test, of around 35 speakers in the current study. These differences might temper comparisons, but they are not very dramatic and might represent differences between implementation of systems in practice.

In general, findings from this study corroborate those findings from Kelly et al. (Kelly & Harte, 2011; Kelly, Drygajlo, & Harte, 2012) that performance degrades significantly over time, and that age-variation after 10 years represents larger effects than intra-speaker variation for most speakers (for some this was greater, up to 20 or 30 years, depending on speaker and age range). The range of 10-20 years is similar to the current study, which found significant changes after 21 and 28 years for most parameters and limited changes for some speakers after 7-14 years. Findings from Kelly et al. and the present study do not support the conclusion in Künzel (2007) that a delay of 11 years has almost no influence. For some speakers, even a 7 year delay was shown to have some influence and although 10 (for Kelly et al.) or 14 years (in the present study) were generally the upper limit for resilience to aging, effects are still apparent.

7.2.8.4 Summary

In summary, age has a significant effect on the performance of an ASR system. Whilst some speakers show a steady decrease in the magnitude of LLRs, for others there is a sharper drop at 21-28 years delay. In general LLRs were extremely high, especially compared with other acoustic measures in the present study, but the test conditions were idealistic: the data are of good technical quality, present large durations of net speech and do not feature the channel and transmission mismatch that would be expected in forensic materials. It would be interesting to see further ASR research into aging (and any other limiting factors) to include variable testing of forensic conditions, with differing lengths, quality and mismatch of model and test samples. Much of this research into the impact of individual factors on ASR performance is already ongoing, and researchers are beginning to assess the combined effect of these forensically realistic factors.

Between the 6 subjects and across the 112 tests there were 4 ‘misses’ (i.e. same speaker tests with negative LLRs). However, this does not take into account the more damaging ‘false hits’ error type (i.e. different speaker tests falsely yielding positive LLRs). In this study, model data for each speaker was piloted against all other test data, with false hits throughout. This requires testing over more comprehensive datasets. Results from the current study corroborate those in Kelly et al. (Kelly & Harte, 2011; Kelly, Drygajlo, & Harte, 2012) but, to some extent, are at odds with the results in Künzel (2007).

7.3 Discussion

This section considers the ASR results with respect to research questions 2 and 4, set out below. While these findings are from a limited dataset, they do corroborate findings from other studies on the effects of aging on ASRs (Kelly & Harte, 2011; Kelly, Drygajlo, & Harte, 2012) and concur with the acoustic and LR changes observed in previous chapters.

7.3.1 Research question 2

2 What is the magnitude of change in individuals' vocal output during adulthood?

LR scores generated from the *BATVOX* ASR demonstrate that there are significant age-related changes to vocal outputs (measured by MFCC) during the testing over almost three decades. The magnitude, rate and timing of changes to LR estimates were different for different speakers. For all speakers, a 28 year delay lead to severe degradation of the system's performance. In general, this difference between the smallest and largest delays was about 5 orders of magnitude, and in many cases was up to 9 or 10.

Using LR estimations allows us to put results into a practical framework and consider the difference aging causes to the 'output' of a forensic analysis. Using data from Bruce's non-contemporaneous comparison test (49 model), presented in LR-compatible conclusions, the difference between a 7 year and 28 year delay becomes patently clear:

> 7 year delay: "Based on my evaluation of the evidence, I have calculated that one would be **ten billion times more likely** to obtain the acoustic differences between the voice samples if the question-voice sample had been produced by the accused than if it had been produced by someone other than the accused. What this means is that whatever you believed before this evidence was presented, you should now be **ten billion times more likely than before** to believe that the voice on the questioned-voice recording is that of the accused."
>
> *or*
>
> "This evidence provides **very strong support** for the prosecution hypothesis"
>
> *compared with*
>
> 28 year delay: "Based on my evaluation of the evidence, I have calculated that one would be **seventy three times more likely** to obtain the acoustic differences between the voice samples if the question-voice sample had been produced by the accused than if it had been produced by someone other than the accused. What this means is that whatever you believed before this evidence was presented, you should now be **seventy three times**

> **more likely than before** to believe that the voice on the questioned-voice recording is that of the accused."
>
> *or*
>
> "This evidence provides **limited (to moderate) support** for the prosecution hypothesis"
>
> Conclusion text sources: numerical conclusion from Morrison (2010a), verbal alternative based on Champod and Evett (2000)

Given that these results are from idealistic materials, the assumption should be that forensic conditions would reduce system performance even further. The difference between these two conclusions is extensive, especially in the format that the evidence would eventually be presented in to triers of fact.

The changes in magnitude of LLR from the *BATVOX* system are perhaps of the largest magnitude of any of the changes observed in the present study, and should corroborate the hypothesis that there may be significant changes to the vocal tract, even in early adulthood. The following sections dissect the changes further in three parts.

7.3.1.1 Research question 2a

a Which features remain more stable than others?

Comparing changes in this chapter with other acoustic results is difficult given that the LLR scores reflect the performance of the system in comparing samples, which gives only indirect access to the changes in actual MFCC data. What is possible, however, is to compare changes in strength of evidence estimates for different features in the study. These comparisons should be tempered by the limited number of subjects and the issue that the reference population for LRs from acoustic monophthong data were not consistent with that of diphthong or ASR LR calculations. Furthermore, testing with the ASR was done with far more material (in relation to what would be available in a canonical forensic setting) with very high technical quality, whereas some acoustic tests had fairly low token Ns.

The first thing to note is that ASR tests produced much larger magnitude (correct) LLRs in general and especially for two of six speakers. Therefore the system was performing at a higher level than the acoustic LR calculations at smaller delays. It also produced fewer (miss) errors than the acoustic data, especially when compared with the monophthong tests (although false hit errors are probably more damaging). However, the ASR system

showed a greater degree of degradation than both the monophthong and the PRICE FD LR estimations. After the longest 28 year delay, formant transitions for F1 and F2 from a single diphthong actually present greater strength of evidence than the MFCC-based ASR system. *BATVOX* outperforms the monophthong tests, which showed a large proportion of negative LRs after 21 or 28 years (although based on the limited Deterding (1997) data for reference population). Out of all parameters, age had the greatest effect on the ASR, despite its efficacy in comparing speakers after short delays.

7.3.1.2 Research question 2b

b To what extent are changes predictable from a model of sociolinguistics or gero-physiology?

One of the main differences between acoustic data and MFCC data is that it is possible to relate formant information to predictions based on sociolinguistic principles. MFCCs are not directly related to linguistic features but relate to physical features of the vocal tract, meaning sociolinguistic interpretation is difficult, if not impossible. There is the suggestion that habitual vocal setting might correlate with or affect MFCC data (Harrison & French, 2010; French & Foulkes, 2012), but without further analysis of voice quality this study cannot address the ASR data in sociolinguistic terms.

It is possible, however, to make predictions about the likely consequences of senescence on vocal physiology: generally that aging has an effect on the vocal tract and that this in turn affects MFCC-based systems. The results from this chapter support the wider hypothesis that age manifests significant changes to the vocal apparatus, particularly in the length and proportions of the vocal tract.

To an extent it is hard to tease apart the performance of system as a whole (in relation to reference population optimisation and subsequent LR calculations) from actual MFCC changes as a component of the ASR processing. It would be interesting to look at changes in raw MFCCs alone, and as suggested in previous chapters, compare this parameter with other acoustic parameters and physical measures of the vocal tract in the same individuals to observe connections and correlations between physical and acoustic data. Further investigation of the MFCCs behind systems would also help to break down the 'black box' view of ASRs, where the automated parts of the system (isolating speech,

profiling noise, selecting optimised reference populations) are not the features of interest from a phonetic perspective.

7.3.1.3 Research question 2c

c What effect should this have on how we evaluate forensic speech evidence?

The response to this practical question is similar to that in previous chapters: analysts should act with caution and awareness of the likely influence of aging on the performance of the system. With ASR systems, this caution should probably be further enforced, as the degradation of performance is much more striking than for acoustic measures. In specific terms, a delay of 28 years affected all subjects' tests significantly, with some other speakers' data showing reduced performance after 7 or 14 years delay. This effect is present in high quality recordings, matched for channel and transmission which presented a relatively large amount of speech.

ASR systems also present a further issue, as the parameters at the heart of the system are linguistically opaque. Therefore it is much harder to make predictions about the likely changes with factors that may present in forensic casework. In this case that factor is age, but there are probably a number of other factors which influence vocal tract physiology and speakers' behavioural modification of their vocal apparatus. While analysts may be able to predict the direction of formant differences with advancing age, it is not currently possible to do this with an ASR system (although cf. Kelly, Drygajlo and Harte (2012) who have developed a verification system which incorporates aging information to improve accuracy).

7.3.2 **Research question 4**

4 What effect does age mismatch between evidential and suspect recordings have on LR estimation?

As mentioned above, it is hard to tease apart the effects of age on MFCCs from assessing the impact on LRs of mismatch in age between suspect and evidential recordings. As with LR estimates from acoustic data, physical changes result in acoustic changes which affect the strength of evidence with increasing age. A more interesting question raised in chapter 6 relates to mismatch not between suspect and evidential samples, but between

those samples and the reference population. ASR tests in the present study address this point with regards to the difference between older and younger training models.

Although theoretically the reference population should represent a potential perpetrator population based on the evidential sample; *BATVOX* attaches the reference population to the speaker model and tests the distribution distance between suspect and reference data. The difference in analysis technique is beyond the scope of this study, but it does mean that it is possible to observe potential effects of age mismatch between model and reference data. For example, model data at 21 (the verification test type) is matched in terms of age with the DyViS reference database, whereas age 49 (comparison tests) model data are mismatched by around the same 28 year time period. If there were significant effects of mismatch with the model data, a difference between the two types of test would be expected. However, §7.2.8 demonstrates that there are few differences between the two test types (beyond negative LLR results for Andrew) and that therefore there does not seem to be a tangible effect of mismatch with the reference population and the model, at least within the present study. These results are based, however, on very ideal materials and a single reference sample and this point requires testing with realistic materials and differently aged reference populations.

8 Conclusions

This final chapter summarises the findings of each data chapter and discusses their implications for each of the three focus areas of the study: aging, dynamic measurements and LR estimation. It also presents directions for further research emerging from these findings and their limitations, and frames results within practical recommendations for forensic speech science and linguistic research.

8.1 Summary of findings

Findings are summarised briefly in this section. Although references to aging are made generally, it should be remembered that this study is concerned with aging over a specific 28 year period between 21 and 49 years. It is sensible to assume that individuals experience aging in a unique way on individual timescales and that different physiologically-induced processes at different stages of life produce varying results. It is also important to remember that male and female biology is different, particularly in relation to hormonal and 'purposeful' genetic programming. Although the different parameters and systems in this study exhibit different levels of performance in discriminating between speakers, the tests are not designed equally and in relation to realistic conditions so as to fairly assess performance across parameters. It is worth keeping in mind that the effects of age are the focus of this study and that it is concerned with assessing stability or the magnitude of age and time-related changes, and not comparisons between the discriminant power of each parameter *per se*.

8.1.1 Fundamental frequency

It is reported in chapter 4 that there are few firm findings from the fundamental frequency data for male subjects, apart from a slight tendency for some speakers' F0 to decrease, especially after 35. For the two female subjects, F0 is shown to reduce. For one female subject this is very marked, and probably related to the fact that she appears to be a consistent smoker. F0 shows variability between age stages, probably due to the uncontrolled conditions and natural speaking styles in the recordings, and the propensity for F0 to vary within a speaker in response to a number of factors (other than aging).

8.1.2 **Formants of monophthongs**

Chapter 4 reports on the most consistent finding of the project: reductions in the formant frequencies taken at central portions of nine monophthongs: /iː ɪ e a ɑː ʌ ɒ ʊ & uː/. While these are predicted from most of the literature on acoustic aging effects, there are fairly comprehensive patterns in the data that were not expected. Reductions in formant frequencies vary interdependently on the formant and vowel quality under examination. These vowel-type changes in each formant are consistent and expand on the existing literature on acoustic modulation with aging.

F1 shows the largest magnitude and most consistent reductions, with average 8.5% changes in frequency, and all speakers showing an overall reduction. This result supports the notion that reduced flexibility in the temporomandibular joint with age impacts on the extent of jaw opening (Reubold, Harrington, & Kleber, 2010). Moreover, close front vowels show much greater reductions in F1 over the period, somewhere between 13 and 21%. In F2 there are also reductions in frequency, but less consistently and with proportionally smaller changes (mean 3.7% reduction). Overall, F2 for open vowels reduced more than other vowels, between 6-9%, although reductions in the close front vowels could be offset by more widespread linguistic changes (Hawkins & Midgley, 2005). F3 shows the smallest proportional reductions, which may not be surprising given higher F3 frequencies, but is significantly reduced in the same number of tests as for F2. As with F2, there is a tendency for F3 changes to be of greater magnitude in open vowels.

Although these general patterns are consistent, there are other examples where changes are apparent in non-conventional directions. In a number of cases these are predictable from social information regarding widespread changes to accents or social and geographical mobility, as in the case of monophthongs for Suzy and Neil (see their profiles in §4.4).

8.1.2.1 Using estimation formulae to explore physiological causes

Data from chapter 4 are used to postulate on the feasibility of possible explanations for frequency reductions in section 4.2. These estimation formulae provide exploratory guidance that changes in formant frequencies are not likely to be due to speakers making adjustments to preserve the distance between F0 and F1. They do seem, however, to

support the notion that two asynchronous processes may affect changes in the acoustic parameters measured.

The first of these processes, vocal tract lengthening, is widely put forward as a potential aging explanation in the literature. Estimations based on the formant data in this study elucidate a fairly consistent pattern of vocal tract lengthening between age 21 and 49. The second of these predicted processes concerns a change to the habitual vowel space used by speakers, potentially due to changes in the flexibility of articulators (affecting general size) and the mechanisms controlling articulators, such as the temporomandibular joint (affecting height of habitual vowel space). Estimations using this data seem to suggest that this process is at work from the age of 35 with an estimated 47% mean reduction in vowel space area. This is calculated using nine monophthongs to map the vowel space geometrically.

8.1.3 **Formants of diphthongs**

Chapter 0 presents results for analysis of average frequency and the range of gestural movement of the diphthongs /aɪ & eɪ/. Average formants across diphthongs are shown to reduce in line with monophthongs, although there are more marked reductions in F2 and F3 than for monophthong data. Slope of transitions seem to increase with age, despite the opposing literature and concurrent reduction in habitual vowel space estimated from monophthongs. Overall diphthong frequency measures show similar aging patterns to monophthongs. In order to assess the dynamic approach and investigate diphthong data practically using an LR approach, polynomial (cubic) regression coefficients are calculated based on formant transitions. Aging effects are calculated in an LR approach in chapter 6. LR estimations calculated from dynamic formant measures show largely strong support for the correct hypothesis and exhibit fewer negative LLRs than monophthongs and static measures of the same diphthong data. This is especially true for F1. It is postulated that the incorporation of both frequency and extent of movement in one parameter gives more reliable results over time. In articulatory terms, there is tentative evidence to show that even though there are changes to vocal tracts, articulatory behaviours persist over time.

8.1.4 **LR estimation**

In chapter 6, formant data from monophthongs and diphthongs are used to calculate numerical strength of evidence estimates at different lengths of delay. Results show a clear pattern across all acoustic measures for increasing age to negatively affect the correct attribution and magnitude of LLRs. In general, those features which are shown to change the most in terms of frequency show larger changes in LR. However, there are exceptions to this, most likely due to changes in relative distance to distributions in the reference population. Generally FD measures are more robust to aging and present better strength of evidence than monophthongs, most likely as a result of containing more information about the speaker behaviour than a single central measure.

The analyses in this chapter raise a number of questions about applying the LR approach in cases where there is a mismatch of age, either between evidential and suspect samples, or between either of these and a reference population. Most forensic cases exhibit one or more kinds of mismatch between evidential and suspect samples and it is currently unclear how to proceed within a numerical LR approach when faced with this problem. This also has consequences for the design and collection of reference databases and how specific they need to be in order to address samples in a case.

8.1.5 **ASR system (*BATVOX*)**

Chapter 7 presents results across different length delays from an ASR system (*BATVOX*). The analysis demonstrates that although the system produces very large (correct) strength of evidence estimates (to the system maximum of $LR=10^{10}$), performance degrades significantly with aging. Like all other parameters, this reduction is variable between speakers, with steady reductions and some stark drop-offs in LLR. Overall, there is a linear relationship between length of delay and LR (on a base-10 logarithmic scale). This represents a striking degradation. After the maximum 28 year delay, all tests across all six male speakers exhibit extremely large reductions in the strength of evidence estimate, in many cases this reduction in performance corresponds to reduction of LLR≥5. These results corroborate the hypothesis that physiological modulations are at the heart of age-related acoustic changes and that changes to the proportions of the vocal tract are present in early adulthood.

8.2 **Aging**

This study has demonstrated that age is manifested in consistent and largely predictable changes to acoustic measures of speech. Physiological modulation of the vocal apparatus with age is a generally accepted hypothesis which is supported by acoustic data in this thesis. Exploratory estimation techniques seem to suggest that two of the popular theories of senescence, vocal tract lengthening and reduction of articulator movement, affect speakers asynchronously. Mel-frequency cepstral coefficients, which reflect the physiognomy of the vocal tract, are also affected by aging and corroborate this explanation. Results also suggest that physiological changes affect various acoustic parameters differently. Different resonances (particularly the lowest formant) experience greater shifts, moreover different vowel properties are affected systematically and have an interdependent relationship with different formants.

That is not to say that aging is a totally linear, systematic, or entirely predictable process. The general trends are not borne out in every case or realised steadily across time. There are other factors which affect acoustic output, such as the interaction between language and the individual's lifestyle and social environment, which must be taken into account. It is important to remember that speakers are individuals and not just uniformly expanding vocal tracts; they also smoke and move to America and want to be politicians.

All these factors are relevant to forensic applications of speech science. It is important to know that certain frequencies are reduced by aging but also that they may increase due to changes in a widespread language variety over time. The job of a forensic analyst does not simply entail measuring these frequencies. It requires interpretation of the data by the application of knowledge and experience of the likely effects of these factors to ascertain whether a certain distribution is likely or not in a certain set of circumstances, usually differing between forensic samples.

8.3 **Formant dynamics**

Although it encompasses a range of acoustic measures, this thesis is designed to investigate the efficacy of dynamic measures of formant transitions in the face of aging processes. Tentative results show that although aging affects all frequency outputs, dynamic measures of formant transitions are more resilient to the process of aging than static formant- and MFCC-based parameters. This thesis judges that if available, making

these kinds of measures is worthwhile given that, put simply, they capture more information than static measures and characterise speaker behaviours which remain relatively stable over time.

8.4 Likelihood ratio estimation

The likelihood ratio approach is implemented in the current study to assess changes in strength of evidence in age-mismatched tests. Where acoustic measures were affected by age, LRs were expectedly shown to reduce. While the analysis is useful for considering the practical consequences of age-related changes, it also raised questions about the application of the LR approach, particularly in relation to (those regular) cases where analysts expect there to be variation in one or more factors between evidential and suspect samples. Factors present in this thesis which are highlighted include mismatch in age or geographical location between test and reference samples. Issues for concern also include the specificity of composition of reference databases for forensic casework, given the likely distributions of acoustic data across different age groups and widespread changes in accents which speakers may conform to. Further issues are also raised in section 2.3.5. Despite the benefits of an overtly empirical approach for new regulations regarding assessment criteria for forensic practices, there are many questions still to be answered before this approach can be universally applied to speech evidence in practice.

8.5 Further research

This section provides areas for further work emerging from the present thesis. Although the *Up* recordings are a rich resource for longitudinal language study, it is recognised that the number of subjects is limited, and a number of tests in this study would benefit from larger, more comprehensive testing (this is noted throughout the text). Other than this, there are features not analysed here that are routinely used in forensic analysis which should receive further attention, in terms of individual changes and those patterns likely to be found in different age populations. Generally these features have been analysed with reference to elderly speakers, but the adulthood period in the current study is perhaps more relevant to the forensic domain. Voice quality assessment requires further examination with respect to age, as it was shown to be one of the most universally utilised parameters across many practitioners (Gold & French, 2011). Other features which might be of relevance would include a comparative assessment of LTAF measures

to investigate possible correspondence with the manual formant measures in this study. Temporal features may also be affected, given that they are reported to be heavily modulated in older age.

Throughout this thesis it is also put forward that future aging studies should incorporate measurements of both acoustic and physical parameters, in order that the relationship between physiological cause and acoustic effect be more closely investigated. The estimation formulae used in the present study are somewhat blunt and can only offer suggestion or corroboration. Direct physiological evidence combined with acoustic measurements would give a much better picture of which acoustic differences are the result of physiology and which are modified by speakers. Modern imaging hardware such as MRI or the use of techniques such as acoustic reflectometry would provide a basis from which to assess the relationship between changes to the vocal apparatus and vocal output. While there is a wealth of research on development of the vocal tract during childhood and adolescence, more focus on forensically-relevant periods, such as adulthood, would be welcome.

The third focus area of this thesis concerns the likelihood ratio approach. Numerous questions are raised throughout chapters 2 and 6 relating to taking this approach from theory to practice. Aging-specific questions were related above in §8.4 but general questions remain about accent variation, sound changes, composition of large reference databases and the required composition of samples in specific analyses. While there is much debate and research around the theoretical implications of this approach, it is important that these are framed within sensible boundaries with regards to the financial limitations of any forensic exercise, especially in the currently contracting forensic 'marketplace'.

8.6 **Practical recommendations**

As well as contributing to the understanding of aging effects on the vocal apparatus and speech behaviour, this thesis can contribute practical recommendations. The principal application of these findings is to the forensic domain. In this respect, aging results presented can have a direct influence in long-term non-contemporaneity between samples, or less directly, in the composition of reference databases.

With regard to the former, those analysts working within the auditory-acoustic method should be aware of and expect changes of the type observed in the chapters above. An awareness of the direct impact of age, much like knowledge of the likely effects of telephone transmission or Lombard speech, give the analyst a better understanding of resulting variation. Furthermore, as aging logically entails the passage of time, analysts should be aware that time offers the opportunity for speakers to adapt to wider accent changes, to move geographically and to become socially mobile. All of these factors, indeed factors not limited to these, can have an influence on the measurable features of speech and practitioners should use sociolinguistic and physiological literature to guide. A further recommendation for practitioners within an auditory-acoustic method is that formant dynamic methods seem to be resilient in the face of the aging process, particularly in F1 (which might be surprising given overall F1 modulation). That is, of course, taking into account the feasibility of collecting sufficient diphthong data with shorter samples.

The direct influence of aging is also shown to have a significant effect on the performance of an ASR system, even with ideal materials. The recommendation for practitioners using an ASR system (especially for those using it in isolation) is to view results with a 7-14 year non-contemporaneous delay with extreme caution. For delays of a magnitude greater than 14 years, the data above corroborate findings from other studies that an ASR system suffers severe degradation in performance and should not be used for forensic analysis (at least until satisfactory accuracy with aging samples is developed). This also applies to verification applications of GMM-UBM systems, where model samples should be re-trained at intervals no longer than around 5-7 years.

Aging also has an indirect impact on forensic analysis, where reference databases are used to frame the strength of evidence given to a forensic assessment. Given that there are physiological processes that affect speech with age, different ages in a population are likely to exhibit different distributions of this data. It is important that reference populations reflect the test samples accurately so as not to inflate the magnitude of the LR, even in cases where the two test samples are fairly contemporaneous. The specificity of this matching should emerge from further research into distribution of speech data in different age groups, given that age is a complex, individual and non-linear process.

Outside of the forensic domain, the findings of this study address the apparent time construct in sociolinguistics. Analysts should be aware that there are likely physiologically-induced changes in formants (which affect sounds and measurements). These may be in a direction that points to sound change or age-grading and should not be confused with either process. Apart from age-related changes, most speakers were remarkably stable, while others showed a small number of changes in contrast to the widely held belief that language fossilises in adolescence. Sociolinguistic studies should seek to incorporate these 'non-traditional, non-stable' speakers in future research for two reasons: firstly our model of speech variation should not be based on a proportion of the population who never move or shift place in the social ladder, especially given the frequency of these factors in the UK population. Secondly, and perhaps more importantly for forensic speech science, the model of speech variation that is relied on in casework should be built on a representative sample including these 'non-traditional' speakers. It is not possible to control for these factors when presented with criminal data.

Reference List

Agha, A. (2003). The social life of cultural value. *Language & communication 23*, 231-273.

Agnitio Corporation. (2012). *Batvox (web pages)*. Retrieved 11 19, 2012, from Agnitio corp: http://www.agnitio-corp.com/producto.php?id_producto=2

Aitken, C. G. G. (1995). *Statistics and the evaluation of evidence for forensic scientists.* Chichester: Wiley.

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., & Taroni, F. (2004). *Statistics and the evaluation of evidence for forensic scientists.* London: John Wiley & Sons.

Alderman, T. (2004). The Bernard data set as a reference distribution for Bayesian likelihood ratio-based forensice speaker identification using formants. *Proceedings of the 10th Australian conference on speech science and technology, Dec 8-10*, (pp. 510-515). Sydney.

Alexander, A. (2007). Forensic automatic speaker recognition using Bayesian interpretation and statistical compensation for mismatched conditions. *International journal of speech, language and the law 14 (1)*, 145-155.

Alexander, A., & Forth, O. (2011). 'No, thank you for the music': an application of audio fingerprinting and automatic music signal cancellation for forensic audio enhancement. *Presentation delivered to the 20th annual conference of the international association for forensic phonetics and acoustics, 24-28 July.* Vienna.

Apted, M. (Director). (1964). *'Up' series* [Motion Picture].

Apted, M. (Director). (1998). *42 Up* [Motion Picture].

Atkinson, N. (2009). *Formant dynamics for SSBE monophthongs in unscripted speech.* MSc Dissertation: University of York.

Bailey, G. (2004). Real and apparent time. In J. K. Chambers, P. Trudgill, & N. Schilling-Estes (eds.), *The Handbook of language variation and change* (pp. 312-333). London: Blackwell.

Balogh, K., & Lelkes, K. (1961). The tongue in old age. *Gerontologica clinica 3 (supplementary edition)*, 38-54.

Barfüsser, S., & Schiel, F. (2010). Disfluencies in alcholised speech. *Presentation delivered to the 19th annual conference of the international association for forensic phonetics and acoustics, 18-21 July.* Trier.

Baumeister, B., & Schiel, F. (2010). On the effect of alcholisation on fundamental frequency. *Presentation delivered to the 19th annual conference of the international association for forensic phonetics and acoustics, 18-21 July.* Trier.

Bautista-Tapias, R. (2005). *Sistemas forenses de reconocimiento automático de locutor. Determinación y análisis de sus variables más críticas.* Unpublished dissertation, Universidad Politecnica de Madrid.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical transactions of the royal society of London*, 370-418.

Beal, J. (2004). English dialects in the North of England. In E. W. Schneider (ed.), *A handbook of varieties of English: Phonology, vol. 1* (pp. 113-133). Berlin: Mouton de Gruyter.

Beck, J. (1997). Organic variation of the vocal apparatus. In W. J. Hardcastle & J. Laver (eds.), *The handbook of phonetic sciences* (pp. 256-299). Oxford: Blackwell.

Benjamin, B. (1997). Speech production of normally aging adults. *Seminars in speech and language 18*, 135-141.

Bernard, J. R. (1970). Towards the acoustic specification of Australian English. *Zeitschrift fur phonetik, sprachwissenschaft und kommunikations-forschung 23*, 113-128.

Beyerlein, P., Cassidy, A., Kholhatkar, V., Lasarcyk, E., Noth, E., Potard, B., et al. (2008). *Vocal aging explained by vocal tract modelling.* JHU Summer workshop final report.

Biber, D., & Finegan, E. (1989). Styles of stance in English: lexical and grammatical marking of evidentiality and affect. *Text 9*, 93-124.

Birkholz, P., & Kröger, B. J. (2007). Simulation of vocal tract growth for articulatory speech synthesis. In Saarbrücken: J. Trouvain and W. Barry (eds.), *Proceedings of the 16th international congress of phonetic sciences, 6-10 August* (pp. 377-381).

Birren, J., & Schroots, J. (1996). History, concepts, and theory in the psychology of aging. In J. Birren, & K. Schaie (eds.), *Handbook of the psychology of aging (4th ed.)* (pp. 3-23). San Diego: Academic Press.

Blake, R., & Josey, M. (2003). The /ay/diphthong in Martha's Vineyard community: what can we say 40 years after Labov? *Language in Society 32*, 451-485.

Blanden, J., Gregg, P., & Machin, S. (2005). *Intergenerational mobility in Europe and North America.* London: Centre for Economic performance.

Bless, D., Biever, D., & Shaik, A. (1986). Comparisons of vibratory characteristics of young adult males and females. *In Hibi, J., Hirano, M. & Bless, D (eds.) Proceedings of the international conference on voice.*, (pp. 46-54). Kurume.

Boreham, R., Fuller, E., Hills, A., & Pudney, S. (2006). *The arrestee survey annual report: October 2003 - September 2004, England and Wales.* London: Home Officer: Home Office Statistical Bulletin 04/06.

Bourdieu, P. (1977). The economics of linguistic exchanges. *Social science information 16*, 645-688.

Bourdieu, P. (1991). *Language and symbolic power.* Cambridge, Mass.: HUP.

Bourdieu, P., & Passeron, J. C. (1990). *Reproduction in education, society and culture.* New York: Sage Publications Inc.

Bowie, D. (2000). *The effect of geographical mobility on the retention of a local dialect.* PhD Thesis, University of Pennsylvania.

Bowie, D. (2005). Language change over the lifespan: a test of the apprent time construct. *University of Pennsylvania working papers in linguistics*, 45-58.

Bowie, D. (2006). Adult linguistic stability and the gathering of linguistic evidence. *Presentation to the international conference on linguistic evidence, 3rd Feb.*

Bowie, D. (2010). The ageing voice: changing identity over time. In C. L. (eds), *Language and Identities* (pp. 55-66). Edinburgh: EUP.

Braun, A. (1995). Fundamental frequency - how speaker-specific is it? *Beiträge zur Phonetik und Linguistik 64*, 9-23.

Braun, A., & Diehl, N. (2010). Age estimation in whispered speech. *Presentation delivered to the 19th annual conference of the international association for forensic phonetics and acoustics, 18-21 July.* Trier.

Broeders, A. P., Cambier-Langevald, T., & Vermuelen, J. (2002). Arranging a voice lin-up in a foreign language. *Forensic Linguistics 9 (1)*, 104-112.

Browman, C. P., & Goldstein, L. (1990). Gestural specification using dynamically-defined articulatory structures. *Journal of Phonetics 18 (3)*, 299-320.

Brown, S., & Willis, S. (2009). Complexity in forensic science. *Forensic science and policy management 1*, 192-198.

Brown, W., Morris, R., Hollien, H., & Howell, E. (1991). Speaking fundamental frequency as a function of age and professional singing. *Journal of voice 5*, 310-315.

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

Buckleton, J. S., & Walsh, K. (1991). Who is "random man"? *Science and Justice 31 (4)*, 463-468.

Byrne, C., & Foulkes, P. (2004). The 'mobile phone effect' on vowel formants. *Forensic Linguistics 11*, 83-102.

Carré, R., Pellegrino, F., & Divenyi, P. (2007). Speech dynamics: epistemological aspects. In J. Trouvain & W. Barry (eds.), *Proceedings of the 16th International congress of phonetic sciences, 6-10 August* (pp. 569-572). Saarbrücken.

Cellmark Forensic Services (FSS73). (2010). *Forensic Science Service (written evidence submitted to the Science and Technology Commons select committee).*

Chambers, J. K. (1988). Acquisition of phonological variants. In A. R. Thomas (ed.), *Methods in dialectology.* Multilingual matters.

Chambers, J. K. (1992). Dialect acquisition. *Language 68*, 673-705.

Chambers, J. K. (1995). *Sociolinguistic theory.* London: Blackwell.

Champod, C., & Evett, I. W. (2000). Commentary on: Broeders, A.P.A. (1999) some observations on the use of probabaility scales in forensic identification. *Forensic Linguistics 6 (2)*, 228-241.

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

Chan, T. W., & Boliver, V. (2011). Social mobility over three generations in Britain. *Paper presented at the 2011 spring meeting of the ISA RC28 in Essex.*

Chao, M., Colombo, S., & Bowie, D. (2007). Linguistic stability and variation across the lifespan. *Presentation to the American dialect society annual meeting. 5 Jan.*

Chin, S., & Pisoni, D. (1997). *Alcohol and speech.* San Diego: Academic Press.

Clark, J., & Foulkes, P. (2007). Identfication of voice in electronically disguised speech. *International journal of speech, languague and the law, 14 (2)*, 195-221.

Clermont, F. (2002). Systemic comparison of spoken and sung vowels in formant-frequency space. *Proceedings of the 9th International conference on speech science and technology, 2-5 December* (pp. 124-129). Melbourne: ASSTA.

Clermont, F. (2007). A linear-scaling approach to speaker variability in poly-segmental formant ensembles. In C. Mueller (ed.), *Speaker Classification II* (pp. 116-129). Berlin: Springer-Verlag.

Clermont, F. (2009). Linear scaling effects of co-articulation in vowel space. *Presentation at the 18th annual conference of the international association for forensic phonetics and acoustics, 2-5 August.* Cambridge.

Clermont, F. (2011). Speaker-variance ratios in forensically-realisatic vowel formant data: normalising for consonantal context. *Presentation delivered to the 20th annual*

conference of the international association for forensic phonetics and acoustics, 24-28 July. Vienna.

Clermont, F., & Mokhtari, P. (1998). Acoustic-articulatory evaluation of the upper vowel-formant region and its presumed speaker-specific potential. *Proceedings of the International conference on spoken language processing, 30 November-4 December* (pp. 527-530). Sydney: ICSLP.

Cohen, J., & Gitman, L. (1959). Oral complaints and taste perception in the aged. *Journal of gerontology 14*, 294-298.

Corak, M. (2004). *Generational income mobility in North America and Europe.* Cambridge: CUP.

Crawford, C., Johnson, P., Machlin, S., & Vignoles, A. (2011). *Social Mobility: a literature review.* London: Department for business, innovation and skills.

Criminal Expert (Experts) Bill. (2011). London.

Cudmore, A. (2011). *Juror interpretations of forensic speaker comparison evidence.* Unpublished MSc Dissertation, University of York.

Cukor-Avila, P. (2002). She say, she go, she be like: verbs of quotation over time in African American vernacular English. *American Speech 77(1)*, 3-31.

Daubert v Merrell Dow Pharmaceuticals Inc (1993).

DeCoster, W., & Debruyne, F. (2000). Longitudinal voice changes: facts and interpretation. *Journal of voice 14*, 184-193.

Denmett, A., & Stillwell, J. (2011). A new area classification for understanding internal migration in Britain. *Population Trends 145, Office for National Statistics*, 146-172.

DePinto, O., & Hollien, H. (1982). Speaking fundamental frequency characteristics of Australian women: then and now. *Journal of phonetics 10*, 367-375.

Deterding, D. (1997). The formants of monophthong vowels in standard southern British English. *Journal of the international phonetic association, 27*, 47-55.

Drager, K. (2006). Social categories, grammatical categories and the likelihood of "like" monophthongisation. *Proceedings of the 11th Australian International Conference on Speech Science & Technology, Paul Warren & Catherine I. Watson (eds.)* (pp. 384-387). Auckland, New Zealand: Australian Speech Science and Technology.

Drygajlo, A. (2010). Deterministic and statistical methods for quantitative interpretation of recorded speech as biometric evidence. *Presentation delivered to th 19th annual conference of the international association for forensic phonetics and acoustics, 18-21 July.* Trier.

Eckert, P. (2000). *Linguistic variation as social practice.* Oxford: Blackwell.

Endres, W., Bambach, W., & Flösser, G. (1971). Voice spectrograms as a function of age, voice disguise, and voice imitation. *Journal of the acoustical society of America 49 (4)*, 1842-1848.

Eriksson, E. J., & Sullivan, K. P. (2008). An invesitgation into the effectiveness of a Swedish glide + vowel segment for speaker discrimination. *International journal of speech, language and the law 15 (1)*, 51-66.

Evans, B. G., & Iversen, P. (2007). Plasticity in vowel perception and production: a study of accent change in young adults. *Journal of the acoustical society of america 121*, 3814-3826.

Evett, I. W. (1995). Avoiding the transposed conditional. *Science and Justice 35 (2)*, 127-131.

Enzinger, E. (2010). Parametric representations of diphthongal formant trajectories of Viennese German /ae/. *Presentation delivered to the 19th annual conference of the international association for forensic phonetics and acoustics, 18-21 July.* Trier.

Enzinger, E., & Morrison, G. S. (2012). The effect of session variability on the validity of forensic-voice-comparison systems. *Poster presented at the 27th annual conference of the international association for forensic phonetics and acoustics.* Santander, Spain.

Fabricius, A., Watt, D., & Johnson, D. E. (2009). A comparison of three speaker-intrinsic vowel formant frequnecy normalisation algorithms for sociophonetics. *Language variation and change, 21*, 413-435.

Fant, G. (1960). *Acoustic theory of speech production.* The Hague: Mouton.

Fecher, N. (2011). Speaking under cover: the impact of face concealing garments on the acoustics of fricatives. *Presentation delivered to the 20th annual conference of the international association for forensic phonetics and acoustics, 24-28 July.* Vienna.

Ferragne, E., & Pellegrino, F. (2010). Formant frequencies of vowels in 13 accents of the British Isles. *Journal of the international phonetic association 40 (1)*, 1-34.

Ferreri, G. (1959). Senscence of the larynx. *Italian general rev oto-rhino-laryngol 1*, 640-709.

Field, A. (2005). *Discovering statistics using SPSS (3rd ed.).* London: Sage.

Fienberg, S. E. (1989). *The evolving role of statistical assessments as evidence in the courts (vol. 1).* New York: Springer-Verlag.

Flügel, C., & Rohen, J. W. (1991). The craniofacial proportions and laryngeal positions in monkeys and man of different ages (a morphometric study based on CT-scans and radiographs). *Mechanisms of aging and development 61*, 65-83.

Foulkes, P., & Docherty, G. (2006). The social life of phonetics and phonology. *Journal of phonetics 34*, 409-438.

Foulkes, P., Scobbie, J. M., & Watt, D. (2010). Sociophonetics. In W. L. Hardcastle (ed.), *Handbook of phonetic sciences (2nd ed.)* (pp. 703-754). Oxford: Blackwell.

Frankel, M. (2011). Doomed Forensic Science Service is world-leading lab. *Research Fortnight-Today*, [URL:http://www.researchresearch.com/index.php?option=com_news&template=rr_2col&view=article&articleId=1085593].

French, J. P. (1998). Mr Akbar's nearest ear versus the Lombard reflex: a case study in forensic phonetics. *Forensic Linguistics 5 (1)*, 58-68.

French, J. P. (2011). Expert evidence in court: developments and changes in the UK position. *Presentation delivered to the 20th annual conference of the international association for forensic phonetics and acoustics, July 24-28.* Vienna.

French, J. P., Harrison, P., & Windsor-Lewis, J. (2006). R v John Samuel Humble: The Yorkshire Ripper Hoaxer trial. *International journal of speech, language and the law 13 (2)*, 255-273.

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.

French, J. P., Nolan, F., Foulkes, P., Harrison, P., & McDougall, K. (2010). The UK position statement on forensic speaker comparison: a rejoinder to Rose and Morrison. *International journal of speech, language and the law 17 (2)*, 143-152.

French, J. P., & Foulkes, P. (2012). The quest for a biometric identifier: Why the voice is different. *Frontiers in forensic science, public lecture series.* University of York, UK.

French, J. P., Foulkes, P., Harrison, P., & Stevens, L. (2012). Vocal tract output measures: relative efficacy, interrelationships and limitations. *Presentation delivered to the 27th annual conference of the international association for forensic phonetics and acoustics, August.* Santander, Spain.

Frye v United States (1923).

General Medical Council v Meadow, EWCA Civ 1390 (2006).

Gick, B., Wilson, I., Koch, K., & Cook, C. (2004). Language-specific articulatory settings: evidence from inter-utterance rest position. *Phonetica 61*, 220-233.

Giddens, A. (1991). *Modernity and self-identity.* Cambridge: Polity.

Giles, H. (1971). *A study of speech patterns in social interaction: accent evaluation and accent change.* (Unpublished PhD thesis, University of Bristol).

Gold, E. (2012). Considerations for the analysis of interlocutors' speech in forensic speech science casework. *Presentation delivered to the 27th annual conference of the International Association for Forensic Phonetics and Acoustics.* Santander, Spain.

Gold, E., & French, J. P. (2011). International practices in forensic speaker comparison. *International journal of speech, language and the law, 18 (2)*, 293-307.

Gold, E., & Hughes, V. (2012). Defining interdependencies between speech parameters. *Poster delivered at BBfor2 Short Summer School in Forensic Evidence Evaluation and Validation, 18 June.* Madrid, Spain.

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

Gordon, E., Maclagan, M., & Hay, J. (2007). The ONZE corpus. In J. Beal, & K. &. Corrigan, *Models and methods in the handling of unconventional digital coropora. Vol. 2: diachronic corpora.* Palgrave.

Griesbach, R., Esser, O., & Weinstock, C. (1995). Speaker identification by formant contours. In A. Braun, & O. Köster (eds.), *Studies in forensic phonetics* (pp. 49-55). Trier: Wissenschatlicher Verlag.

Halliwell, J., Keppens, J., & Shen, Q. (2003). Linguistic Bayesian networks for reasoning with subjective probabilities in forensic statistics. *ICAIL, 24-28 June .* Edinburgh.

Harrington, J. (2006). An acoustic analysis of 'happy-tensing' in the Queen's annual Christmas broadcasts. *Journal of Phonetics 34*, 439-457.

Harrington, J., Palethorpe, S., & Watson, C. J. (2000a). Monphthongal changes in received pronunciation: an acoustic analysis of the Queen's Christmas broadcasts. *Journal of the International Phonetic Association 30*, 63-78.

Harrington, J., Palethorpe, S., & Watson, C. J. (2000b). Does the Queen speak the Queen's English? *Nature 408*, 927.

Harrington, J., Palethorpe, S., & Watson, C. J. (2005). Deepening or lessening the divide between diphthongs: an analysis of the Queen's annual Christmas broadcast. In J. Laver, W. J. Hardcastle, & J. M. Beck (eds.), *A figure of speech: a festschrift for John Laver* (pp. 227-235). London: Routledge.

Harrington, J., Palethorpe, S., & Watson, C. J. (2007). Age-related changes in fundamental frequency and formants: a longitudinal study of four speakers. *INTERSPEECH-2007*, (pp. 2753-2756). Antwerp.

Harrison, P. (2001). GSM interference cancellation for forensic audio: a report on work in progress. *International journal of speech, language and the law 8 (2)*, 9-23.

Harrison, P. (2010). Formant measurement errors for multiple synthetic speakers. *Presentation delivered to the 19th annual conference of the international association of forensic phonetics and acoustics, 18-21 July.* Trier.

Harrison, P. (2011). Formant measurement errors from real speech. *Presentation delivered to the 20th annual conference of the international association for forensic phonetics and acoustics, 24-28 July.* Vienna.

Harrison, P., & French, J. P. (2010). Assessing the suitability of BATVOX for UK casework. *Presentation delivered to the 19th annual conference of the international association of forensic phonetics and acoustics, 18-21 July.* Trier.

Hartman, D., & Danhauer, J. (1976). Perceptual features of speech for males in four perceived age decades. *Journal of the acoustical society of America 59*, 713-715.

Haskins Laboratories. (2007). *Haskins Laboratories.* Retrieved November 20, 2007, from http://www.haskins.yale.edu/research/gestural.html

Hawkins, S., & Midgley, J. (2005). Formant frequencies of RP monophthongs in four age groups of speakers. *Journal of the international phonetic association 35*, 183-199.

Hay, G., Gannon, M., Casey, J., & Millar, T. (2009). *Estimtes of the prevalence of opiate use and.or crack cocaine use, 2008/9: Sweep 5 report.* Glasgow: The Centre for Drug Misuse Research, University of Glasgow.

Herman, S. (1961). Explorations in the social psychology of language choice. *Human relations 14* , 149-64.

Higgins, M., & Saxman, J. (1991). A comparison of selected phonatory behaviours of healthy aged and young adults. *Journal of speech and hearing research 34*, 1000-1010.

Hillman, R., Holmberg, E., Perkell, J., Walsch, M., & Vaugh, C. (1989). Objective assessment of vocal hyperfunction: an experimental framework and initial results. *Journal of speech and hearing research 32*, 373-392.

Hirano, M., Kakita, Y., Ohmaru, K., & Kurita, S. (1982). Structure and mechanical properties of the vocal fold. In N. Lass (ed.), *Speech and language: advances in basic research and practice* (pp. 211-97). New York: Academic Press.

HM Treasury. (2010). *Spending Review.* URL:http://www.direct.gov.uk/prod_consum_dg/groups/dg_digitalassets/@dg/@en/documents/digitalasset/dg_191696.pdf.

Hockett, C. F. (1950). Age-grading and linguistic continuity. *Language 26*, 449-57.

Hollien, H., & Jackson, B. (1973). Normative data on the speaking fundamental frequency characteristics of young adult males. *Journal of Phonetics 1*, 117-120.

Hollien, H., & Shipp, T. (1972). Speaking fundamental frequency and chronologic age in males. *Journal of speech and hearing research 15*, 155-159.

Honjo, I., & Isshiki, N. (1980). Laryngoscopic and voice characteristics of aged persons. *Arch Otolaryngol 106*, 149-50.

Huber, J. E., Stathopoulus, E. T., Curione, G. M., Ash, T. A., & Johnson, K. (1999). Formants of children, women, and men: The effects of vocal intensity variation. *Journal of the Acoustical Society of America, 106*, 1532-1542.

Hudson, R. (1996). *Sociolinguistics.* Cambridge: CUP.

Hudson, T., de Jong, G., McDougall, K., Harrison, P., & Nolan, F. (2007). F0 statistics for 100 male speakers of SSBE. In J. Trouvain and W. Barry (eds.).*Proceedings of the 16th international congress of phonetic sciences, 6-10 August* (pp. 1809-18182). Saarbrücken.

Hughes, A., Trudgill, P., & Watt, D. (1996). *English accents and dialects: an introduction to social and regional varieties of English in the British Isles.* London: Arnold.

Hughes, V. (2009). *Diphthong dynamics in unscripted speech.* MSc dissertation: University of York.

Hughes, V. (2012). The effect of variability on the outcome of likelihood ratios. *Presentation delivered to the 21st international conference of the international association for forensic phonetics and acoustics, 5-8 August.* Santander, Spain.

Hughes, V., McDougall, K., & Foulkes, P. (2009). Diphthong dynamics in unscripted speech. *Presentation delivered to the 18th annual conference of the international association for forensic phonetics and acoustics, 2-5 August.* Cambridge.

Ingram, J. C., Prandolini, R., & Ong, S. (1996). Formant trajectories as indices of phonetic variation for speaker identification. *Forensic linguistics 3 (1)*, 129-145.

Israel, H. (1968). Continuing growth in the human facial skeleton. *Arch oral biology 13*, 133-137.

Israel, H. (1973). Age factor and the pattern of change in craniofacial strcutres. *American journal of physiological anthropology*, 111-128.

Jessen, M. (2008). Forensic Phonetics. *Language and linguistics compass 2 (4)*, 671-711.

Jessen, M. (2009). Forensic phonetics and the infuence of speaking style on global measures of fundamental frequency. In G. Grewendorf, & M. Rathert (eds.), *Formal linguistics and law* (pp. 115-141). Berlin: Walter de Gruyer.

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., Ladefoged, P., & Lindau, M. (1993). Individual differences in vowel production. *Journal of the acoustical society of America 94*, 701-721.

Johnstone, B. (2000). The individual voice in language. *Annual review of anthropology 29*, 405-424.

Johnstone, B. (2009). Stance, style and the linguistic individual. In A. Jaffre (ed.), *Sociolinguistic perspectives on stance.* New York: OUP.

Johnstone, B., Andrus, J., & Danielson, A. E. (2006). Mobility, indexicality, and the enregisterment of 'Pittsburghese'. *Journal of English linguistics 34*, 77-104.

José, B. (2010). The apparent-time construct and stable variation: final /z/ devoicing in North-western Indiana. *Journal of sociolinguistics 14(1)*, 34-59.

Joseph Crosfield & Sons v Techno-Chemical laboratories Ltd., 29 TLR 378 (English High Court 1913).

Kahane, J. (1980). Anatomic and physiologic changes in the aging peripheral speech mechanism. In D. Beasley, & G. Davis (eds.), *Aging: communication processes and disorders.* New York: Grune and Stratton.

Kahane, J. (1983). Postnatal development and aging of the human larynx. *Seminars in speech and language 4*, 189-203.

Kaltieider, N., Fray, W., & Hyde, H. (1938). The effects of age on the total pulmonary capacity and its subdivisions. *American review of tuberculosis 37*, 662-689.

Kavanagh, C. M. (2012). *New consonantal acoustic parameters for forensic speaker comparison.* PhD Thesis, University of York.

Kelly, F., & Harte, N. (2011). Effects of long-term ageing on speaker verification. *Biometrics and ID Management, vol 6583 of Lecture notes on Computer Science*, 113-124.

Kelly, F., Drygajlo, A., & Harte, N. (2012). Speaker verification with long-term ageing data. *Proceedings of the 5th IAPR International conference on Biometrics.* New Delhi March 29-April 1.

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., Ishihara, S., & Rose, P. (2009). Exploring the discriminatory potential of F0 distribution parameter in traditional forensic speaker recognition. *International journal of speech, language and the law 16 (1)*, 59-90.

Kirchübel, C. (2010). The effect of Lombard speech on vowel formant measures. *Presentation delivered to the 19th annual conference of the international association of forensic phonetics and acoustics, 18-21 July.* Trier.

Kirchhübel, C., Howard, D. M., & Stedmon, A. W. (2011). Acoustic correlates of speech when unders stress: Research, methods and future directions. *International journal of speech, language and the law 18 (1)*, 75-98.

Kirk, P. L., & Kingston, C. R. (1964). Evidence evaluation and problems in general criminalistics. *Journal of forensic science 9 (4)*, 434-444.

Krook, M. (1988). Speaking fundamental frequency characteristics of normal Swedish subjects obtained by glottal frequency analysis. *Folia Phoniat (Basel) 40*, 82-90.

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

Künzel, H. J. (1995). Field procedures in forensic speaker recognition. In J. Winsor-Lewis (ed.), *Studies in general and English phonetics in honour of Professor J.D. O'Connor* (pp. 68-84).

Künzel, H. J. (2000). Effects of voice disguise on fundamental frequency. *Forensic Linguistics 7*, 50-62.

Künzel, H. J. (2001). Beware of the 'telephone effect': the influence of telephone transmission on the measurement of formant frequencies. *Forensic Linguistics, 8 (1)*, 80-99.

Künzel, H. J. (2007). Non-contemporary speech samples: auditory detectability of an 11-year delay and its effects on automatic speaker identification. *International journal of speech, language and the law 14 (1)*, 109-136.

Künzel, H. J. (2009). Automatic speaker recognition of identical twins. *Presentation delivered to the 18th annual conference of the international association for forensic phonetics and acoustics, 2-5 August.* Cambridge: UK.

Labov, W. (1972). On mechanisms of linguistic change. *Sociolinguistic patterns*, 160-82.

Labov, W. (1994). *Principles of linguistic change, vol. 1: Internal factors.* London: Wiley-Blackwell.

Labov, W. (1962). *The social history of sound change on the island of Martha's Vineyard, Massachusetts.* Columbia University: Master's Essay.

Labov, W. (1963). The social motivation of a sound change. *Word 19*, 273-309.

Labov, W. (1966). *The social stratification of English in New York City.* Washington DC: Center for Applied Linguistics.

Lasker, G. (1953). The aging factor in bodily measurements of adult male and female Mexicans. *Human Biology 25*, 50-63.

Laver, J. (1980). *The phonetic description of voice quality.* Cambridge: CUP.

Laver, J., & Trudgill, P. (1979). Phonetic and linguistic markers in speech. In K. Scherer, & H. Giles (eds.), *Social markers in speech* (pp. 1-26). Cambridge: CUP.

Law Commission. (2011). *Expert evidence in criminal proceedings in England and Wales (pursuant to section 3(2) of the Law Commissions Act 1965).* London: The stationery office.

Lewontin, R. C. (1991). *The doctrine of DNA: biology and ideology.* London: Penguin.

Lindblom, B., & Sundberg, J. (1971). Acoustical consequences of the lip, tongue, jaw and larynx movement. *Journal of the acoustical society of America 50*, 1166-1179.

Lindh, J., Eriksson, A., & Nelhans, G. (2010). Methodological issues in the presentation and evaluation of speech evidence in Sweden. *Presentation delivered to the 19th annual conference of the international association for forensic phonetics and acoustics, 18-21 July.* Trier.

Lindh, J., Ochoa, F., & Morrison, G. S. (2012). Calculating the reliability of a likelihood ratio from a disputed utterance. *Presentation delivered to the 21st annual conference of the international association for forensic phonetics and acoustics, 5-8 August.* Santander.

Linville, S. E. (1992). Glottal gap configuration in two age groups of women. *Journal of speech and hearing research 35*, 1209-1215.

Linville, S. E. (1996). The sound of senescence. *Journal of voice 10 (2)*, 190-200.

Linville, S. E. (2001). *Vocal aging.* Canada: Singular.

Linville, S. E., & Fisher, H. (1985). Acoustic characteristics of perceived versus actual vocal age in controlled phonation by adult males. *Journal of the acoustical society of America 78*, 40-48.

Linville, S. E., Skarin, B., & Fornatto, E. (1989). The interrelationship of measures related to vocal function, speech rate and laryngeal appearance in elderly women. *Journal of speech and hearing research 32*, 323-330.

Linville, S. E., & Rens, J. (2001). Vocal tract resonance analysis of aging voice using long-term average spectra. *Journal of voice 15 (3)*, 323-330.

Liss, J., Weismer, G., & Rosenbeck, J. (1990). Selected acoustic characteristics of speech production in very old males. *Journal of gerontology psychological sciences*, 35-45.

Llamas, C., & Watt, D. (eds.) (2010). *Language and Identities.* Edinburgh: EUP.

Loakes, D. (2004). Front vowels as speaker-specific: some evidence from Australian English. *Proceedings of the 10th Australian international conference on speech science and technology, 8-10 December*, (pp. 289-294). Sydney.

Loakes, D. (2006). *A forensic phonetic investigation into the speech patterns of identical and non-identical twins.* PhD Thesis: University of Melbourne.

Lynch, M., & McNally, R. (2003). "Science", "common sense" and DNA evidence: a legal controversy about the public understanding of science. *Public understanding of science*, 83-103.

Maeda, S. (1990). “Compensatory articulation during speech: Evidence from the analysis. In W. H. (eds.), *Speech production and speech modelling* (pp. 131-149). Amsterdam: Kluwer academic publisher.

Maeda, S. (1979). Un modèle articulatoire de la langue avec des composantes linéaires. *Actes 10èmes Journées d’Etude sur la Parole*, 152-162.

McDougall, K. (2005). *The role of formant dynamics in determining speaker identity.* PhD Thesis: 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 /u:/ in British English. In J. Trouvain & W. Barry (eds.).*Proceedings of the 16th International congress of phonetic sciences, 6-10 August* (pp. 1825-1828). Saarbrücken.

McQuiston-Surret, D., & Saks, M. J. (2009). The testimony of forensic identification science: what expert witnesses say and what factfinders hear. *Law and human behaviour 33*, 436-453.

Mennen, I., Schaeffler, F., & Doherty, G. (2012) Cross-language differences in fundamental frequency range: a comparison of English and German. *Journal of the Acoustical Society of America, 131 (3)*. pp. 2249-2260

Mennen, I., Scobbie, J. M., de Leeuw, E., Schaeffler, S., & Schaeffler, F. (2010). Measuring language-specific phonetic settings. *Second language research (26:1)*, 13-41.

Metropolitan Police Service (FOI request). (2011). *MPS Freedom of information request (ref. 2011010004810).* London: MPS.

Meyerson, M. (1976). The effects of aging on communication. *Journal of gerontology 31*, 29-38.

Milroy, L., & Milroy, J. (1992). Social network and social class: toward an integrated sociolinguistic model. *Language in society 21 (1)*, 1-26.

Mittman, C., Edelman, N., Norris, A., & Shock, N. (1965). Relationship between chest wall and pulmonary compliance and age. *Journal of applied physiology and age 20*, 1211-1216.

Morrison, G. S. (2008). Forensic voice comparison using likelihood ratios based on polynomial curves fitted to the formant trajectories of Australian English /aI/. *International journal of speech, language and the law*, 249-266.

Morrison, G. S. (2009a). Likelihood-ratio-based forensic speaker comparison using parametric representations of vowel formant trajectories. *Journal of the acoustical society of America*, 2387-2397.

Morrison, G. S. (2009b). The place of forensice voice comparison in the ongoing paradigm shift. *The 2nd international conference on evidence, law and forensice science, conference thesis (1)* (pp. 20-34). Beijing: The key laboratory of evidence science of the ministry of education (the institute of evidence law and forensic science, China University of political science and law).

Morrison, G. S. (2009c). Response to the law commission of England and Wales consultation paper no. 190 "the admissability of expert evidence in criminal proceedings in England and Wales: a new approach to the determination of evidentiary reliability. Retrieved 08 16, 2009, from http://geoff-morrison.net

Morrison, G. S. (2009d). Comments on Coulthard and Johnson's (2007) portrayal of the likelihood-ratio framework. *Australian journal of forensic sciences 41 (2)*, 155-161.

Morrison, G. S. (2010a). Likelihood Ratio framework for forensic voice comparison. *Presentation at the University of York.* York.

Morrison, G. S. (2010b). Empirically assessing the validity and reliability of forensic-comparison systems. *Presentation delivered to the 19th annual conference of the international association for forensic phonetics and acoustics, 18-21 July.* Trier.

Morrison, G. S. (2010c). *An introduction to the evaluation of forensic-voice-comparison evidence.* Retrieved November 20, 2010, from http://www.geoff-morrison.net/documents/TALK%20-%20%20An%20introduction%20to%20the%20evaluation%20of%20forensic-voice-comparison%20evidence.pdf

Morrison, G. S. (2010d). *Forensic voice comparison and the paradigm shift.* Retrieved November 20, 2010, from http://www.geoff-morrison.net/documents/TALK%20-%20Forensic%20voice%20comparison%20and%20the%20paradigm%20shift%20in%20foresnic%20science.pdf

Morrison, G. S. (2010e). Forensic voice comparison. In I. Freckleton, & H. Selby (eds.), *Expert evidence* (p. Ch. 99). Sydney: Thompson Reuters.

Morrison, G. S. (2011). The calculation of likelihhod ratios for speech data. *Workshop delivered at the University of York, 19 April.*

Morrison, G. S., Thiruvaran, T., & Epps, J. (2010). Estimating the precision of the likelihood-ratio output of a forensic voice-comparison system. *The speaker and language recognition workshop, 28 June - 1 July.* Brno.

Morrison, G. S., Ochoa, F., & Thiruvaran, T. (2012). Database selection for forensic voice comparison. *Proceedings of Odyssey 2012: The language and speaker recognition workshop.* Singapore: International Speech Communication Association.

Murphy, C. H., & Doyle, P. C. (1987). The effects of cigarette smoking on voice-fundamental frequency. *Otolaryngol head and neck surgery 97 (4)*, 376-380.

Mysak, E. (1959). Pitch and duration characterisitcs of older males. *Journal of speech and hearing research 2*, 46-54.

Mysak, E., & Hanley, T. (1959). Vocal aging. *Geriatrics 14*, 652-655.

Nahkola, K., & Saanilahti, M. (2004). Mapping language changes in real time: a panel study on Finnish. *Language Variation and Change*, 75-92.

NHS Information Centre. (2008-2010). *Statistics on drug misuse - various collated: URL: http://www.ic.nhs.uk/statistics-and-data-collections/health-and-lifestyles/drug-misuse [accessed on 09/01/2012].* Leeds: The health and social care information centre.

NHS Information Centre. (2011a). *Adult critical care in England April 2009 to March 2010: experimental statistics.* Leeds, URL: http:\\www.hesonline.nhs.uk [accessed on 09/01/2012]: The health and social care information centre.

NHS Information Centre. (2011b). *Smoking patterns in adults and children.* URL: http://www.ic.nhs.uk/statistics-and-data-collections/health-and-lifestyles/smoking [accessed on 09/01/2012]: The health and social care information centre.

NHS Information Centre. (2011c). *Provisional monthly HES data for admitted patient care.* Leeds, URL: http://www.hesonline.nhs.uk/Ease/servlet/ContentServer?siteID=1937&categoryID=1122 [accessed on 09/01/2012]: The health and social care information centre.

NHS Information Centre. (2011d). *In-patients formally detained in hospitals under the Mental Health Act 1983 - and patients subjected to supervised community treatment, Annual Figures, England, 2010/11.* Leeds, URL: http://www.ic.nhs.uk/statistics-and-data-collections/mental-health/nhs-specialist-mental-health-services [accessed on 10/01/2012]: The health and social care information centre.

NHS Information Centre. (2011e). *Mental Health Bulletin: Fourth report from mental health minimum dataset (MHMDS) annual returns, 2010.* Leeds, URL:

http://www.ic.nhs.uk/statistics-and-data-collections/mental-health/nhs-specialist-mental-health-services [accessed on 10/01/2012]: The health and social care information centre.

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

Nolan, F. (1997). Speaker recognition and forensic phonetics. In W. Hardcastle, & J. Laver (eds.), *A handbook of phonetic sciences* (pp. 744-768). Oxford: Blackwell.

Nolan, F. (2001). Speaker identification evidence: its forms, limitation and roles. *Proceedings of the conference 'Law and language: prospect and retrospect'.* Levi, Finnish Lapland.

Nolan, F. (2003). A recent voice parade. *Forensic Linguistics 10 (2)*, 277-291.

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.

Nolan, F., McDougall, K., de Jong, G., & Hudson, T. (2009). The DyViS databse: style-controlled recordings of 100 homogenous speakers for forensic phonetic research. *International journal of speech, language and the law, 16(1)*, 31-57.

Nycz, J. R. (2011). *Second dialect acquisition: implications for theories of phonological representation.* PhD Thesis, New York University.

Ochs, E. (1992). Indexing gender. In A. Duranti & C. Goodwin (eds.), *Rethinking context: language as interactive phenomenon.* (pp. 335-358). New York: CUP.

Office for National Statistics. (2003-2010). *Internal migration for the years 2002-2009 year end December - various collated: URL: www.ons.gov.uk [accessed on 01/10/2011].* London: ONS.

Office for National Statistics. (2010). *Internal Migration within England and Wales.* London: ONS.

Office for National Statistics. (2011). *Population estimates for UK, England and Wales, Scotland and Northern Ireland, Mid-2010 Population estimates.* London, URL: http://www.ons.gov.uk/ons/rel/pop-estimate/population-estimates-for-uk--england-and-wales--scotland-and-northern-ireland/mid-2010-population-estimates/index.html [accessed on 10/01/2012]: Office for National Statistics.

Orlikoff, R. (1990). The relationship of age and cardiovascular health to certain acoustic characteristics of male voices. *Journal of speech and hearing research 33*, 450-457.

Oyer, H., & Deal, L. (1985). Temporal aspects of speech and the aging process. *Folia Phoniatr (Basel) 37*, 109-112.

Paige, A., & Zue, V. W. (1970). Calculation of vocal tract length. *IEEE transactions on audio and electroacoustics 18 (3)*, 268-271.

Papp, V. (2009). The effects of heroin on the spectral quality of vowels. *Presentation delivered to the 18th annual conference of the international association for forensic phonetics and acoustics, 2-5 August.* Cambridge.

Papp, V., Schreuder, M., Theunissen, E., & Ramaekers, J. (2011). Reference corpus of Dutch drug users I: MDMA/ecstasy. *Presentation deliverd to the 20th annual conference of the international association for forensic phonetics and acoustics, 24-28 July.* Vienna.

Parliamentary Questions (#229-354). (2011). *Uncorrected transcript of evidence (Science and Technology Commons Committee) (HC 855-iii).*

Parliamentary Questions (#295-354). (2011). *Examination of Witnesses.* [URL: http://www.publications.parliament.uk/pa/cm201012/cmselect/cmsctech/855/11042703.htm accessed on 23/08/11].

Pierce, J., & Ebert, R. (1965). Fibrous network of the lung and its change with age. *Thorax 20*, 468-473.

R v Adams, 2 Cr App Rep 467 (Royal courts of London April 26, 1996).

R v Bonython, SASR 45 (Australian Supreme Court, Appeal 1984).

R v John Samuel Humble (Leeds Crown Court October 2005).

Ramig, L. (1983). Effects of physiological age on speaking and reading rates. *Journal of communication 16*, 217-226.

Ramig, L. (1986). Aging speech: physiological and social aspects. *Language and communication 6*, 25-34.

Ratstatter, M., & Jacques, R. (1990). Formant frequency structure of the aging male and female vocal tract. *Folia Phoniatr (Basel) 42*, 312-319.

Ratstatter, M., McGuire, R., Kalinowski, J., & Stuart, A. (1997). Formant frequency characteristics of elderly speakers in contextual speech. *Folia phoniatrica et lodopaedica 49*, 1-8.

Reubold, U., Harrington, J., & Kleber, F. (2010). Vocal aging effects on F0 and the first formant: a longitudinal analysis in adult speakers. *Speech Communication 52 (7-8)*, 638-651.

Reuter, P., & Stevens, A. (2007). *An analysis of UK drug policy.* UK Drug Policy Commission.

Rhodes, R. (2009). *Using /aɪ/ to discriminate between Derby speakers using formant dynamics in spontaneous speech.* MSc Dissertation: University of York.

National Research Council Governing Board Committee on the Assessment of Risk (abbr. Risk) (1981). *The handling of risk assessments in NRC reports.* Washington DC.

Roberts, L. (2011). Acoustic characteristics of distress speech in real victims and trained actors. *Presentation delivered to the 20th annual conference of the international association for forensic phonetics and acoustics, 24-28 July.* Vienna.

Robertson, G., & Vignaux, G. A. (1995). *Interpreting evidence: evaluating forensic science in the courtroom.* Chichester: Wiley.

Robertson, J. (2007). Forensic speech science from a Police perspective. *Paper presented at the Australian Research Council Network in Human Communication science workshop: FSI not CSI: Perspectives in state-of-the-art forensic speaker recognition.* Sydney.

Rose, P. (1999). Long- and short-term within-speaker differences in the formants of Australian hello. *Journal of the international phonetic association 29*, 1-31.

Rose, P. (2002). *Forensic speaker identification.* London: Taylor-Francis Ltd.

Rose, P. (2010, September 9). Personal communication.

Rose, P. (2010). The effect of correlation on strength of evidence estimates in forensic voice comparison: uni- and multivariate likelihood ratio-based discrimination with Australian English vowel acoustics. *International journal of biometrics 2 (4)*, 316-329.

Rose, P., Osanai, T., & Kinoshita, Y. (2003). Strength of forensic speaker identification evidence: multispeaker formant- and cepstrum-based segmental distribution with a Bayesian likelihood ratio as threshold. *International journal of speech, language and the law 10 (2)*, 179-203.

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.

Russel, A., Penny, L., & Pemberton, C. (1995). Speaking fundamental frequency changes over time in women: a longitudinal study. *Journal of speech and hearing research 38*, 101-109.

Ryan, W. (1972). Acoustic aspects of the aging voice. *Journal of geronotology 27*, 265-268.

Saltzman, E. (1986). Task dynamic coordination of the speech articulators: a preliminary model. In H. Heuer, & C. Fromm (eds.), *Generation and modulation of action patterns* (pp. 129-144). Berlin: Springer-Verlag.

Sankoff, G. (2004). Adolescents, young adults and the critical period: two case studies from "Seven Up". In C. Fought (ed.), *Sociolinguistic variation, critical reflections* (pp. 121-140). Oxford: OUP.

Sataloff, R. T., Rosen, D. C., Hawkshaw, M., & Spiegel, J. R. (1997). The three ages of voice: the aging adult voice. *Journal of voice 11 (2)*, 156-160.

Saxman, J., & Burk, K. (1967). Speaking fundamental frequency characteristics of middle-aged females. *Folia Phoniatr (Basel) 19*, 167-172.

Schaeffler, S., Scobbie, J. M., & Mennen, I. (2008). An evaluation of inter-speech postures for the study of language-specific articulatory settings. *Proceedings of the Eighth International Seminar on Speech Production (ISSP)*, (pp. 121-124).

Schilling-Estes, N. (2004). Introduction to chapter on time. In J. K. Chambers, P. Trudgill, & N. Schilling-Estes (eds.), *The Handbook of language variation and change* (pp. 311-312). London: Blackwell.

Silverman, S. (1972). Degeneration of dental and oral strcutures. In ASHA reports 7: *Orofacial function: clinical research in dentistry and speech pathology.* Washington: ASHA.

Silverstein, M. (2003). Indexical order and the dialectics of sociolinguistic life. *Language and Communication 23*, 193-229.

Soderston, M., & Lindestad, P. (1990). Glottal closure and perceived breathiness during phonation in normally speaking subjects. *Journal of speech and hearing research 33*, 601-611.

Stevens, L., & French, J. P. (2012). Voice quality in studio quality and telephone transmitted recordings. *Presentation delivered to the BAAP Colloquium.* Leeds.

Stoicheff, M. (1981). Speaking fundamental frequency characteristics of nonsmoking female adults. *Journal of speech and hearing research 24*, 437-441.

Sullivan, K. P., & Schlichting, F. (2000). Speaker discrimination in a foreign language: first language environment, second language learners. *Forensic linguistics 7 (1)*, 97-214.

Suzuki, T., Tanimoto, M., Osanai, T., & Kido, H. (1996). Acoustic variation of voice with aging of male speakers on vowels and nasal sounds. *Proceedings of the 1996 annual meeting of the American academy of forensic sciences*, (p. 85). Nashville.

Tagliamonte, S. A., & Molfenter, S. (2007). How'd you get that accent? Acquiring a second dialect of the same language. *Language in society 36*, 649-675.

Taroni, F., & Aitken, C. G. (1998a). Probabalistic reasoning in the law. Part 1: assessment of the probabilities and explanation of DNA evidence. *Science & Justice*, 165-177.

Taroni, F., & Aitken, C. G. (1998b). Probabalistic reasoning in the law. Part 1: assessment of the probabilities and explanation of trace evidence other than DNA. *Science and justice 38*, 179-188.

Thompson, W. C. (1989). Are jurors competent to evaluate statistical evidence? *Law and contemporary problems 9*, 52.

Thompson, W. C., & Schumann, E. L. (1987). Interpretation of statistical evidence in criminal trials: the Prosecutor's fallacy and the Defence attorney's fallacy. *Law and human behaviour 11 (3)*, 167-187.

Tillery, J., & Bailey, G. (2003). Approaches to real-time in dialectology and sociolinguistics. *World Englishes*, 351-365.

Tranmüller, H. (1981). Perceptual dimensions of openness in vowels. *Journal of the acoustical society of America 69*, 1465-1475.

Trudgill, P. (1972). Sex and covert prestige: linguistic change in the urban dialect of Norwich. *Language in society 1*, 179-95.

Trudgill, P. (1989). Contact and isolation in linguistic change. In *Language Change: Contributions to the Study of its Causes.* (pp. 227-237). Berlin: Mouton de Gruyter.

Turk, A., Nakai, S., & Sugahara, M. (2006). Acoustic segment durations in prosodic research: a practical guide. In S. Sudhoff, D. Lenertova, R. Meyer, S. Pappert, P. Augurzky, I. Mleinek, et al. (eds.), *Methods in empirical prosody research* (pp. 1-28). Berlin: Walter de Gruyer.

Van Buuren, L. (1988). Margaret Thatcher's pronunciation: an exercise in ear-training. *Journal of the international phonetic association 18*, 26-38.

Verdonck-de-Leeuw, I. M., & Mahieu, H. F. (2004). Vocal aging and the impact on daily life: a longitudinal study. *Journal of voice 18 (2)*, 193-202.

Vermeulen, J. (2009). *Beware the 'distance effect' on vowel formants.* University of York: Unpublished MSc Dissertataion.

Vipperla, R., Renals, S., & Frankel, J. (2010). Ageing voices: the effects of changes in voice parameters on ASR peformance. *EURASIP Journal on audio, speech and audio processing*.

Watson, P. J., & Munson, B. (2007). A comparison of vowel acoustics between older and younger adults. *Proceedings of the 16th International Congress of the Phonetic Sciences.* Saarbrücken, Germany.

Watt, D., Llamas, C., & Harrison, P. (2010). Differences in perceived sound quality between speech recordings filtered using transmission loss spectra of selected fabrics. *Presentation delivered to the 19th annual conference of the international association for forensic phonetics and acoustics, 18-21 July.* Trier.

Wells, J. C. (1982). *Accents of English.* Cambridge: CUP.

Wilder, C. (1978). Vocal aging. In Weinberg, B. (ed.) *Transcripts of the seventh symposium: care of the professional voice. Part II: life span changes in the human voice.* New York: Voice Foundation.

Wilson, I. (2006). *Articulatory settings of French and English monolinguial and bilingual speakers.* PhD thesis, University of British Columbia.

Wind, J. (1970). *On the phylogeny and ontogeny of the human larynx.* Groningen: Wolters-Noordhoff.

Windsor-Lewis, J. (1971). The American and British Accents of English. *ELT Journal 25 (3)*, 239-248.

Woolard, K. A. (2009). Language variation and cultural hegemony: toward an integration of sociolinguistic and social theory. *American ethnologist 12 (4)*, 738-748.

Xue, S. A., & Hao, G. J. (2003). Changes in the human vocal tract due to aging and the acoustic correlates of speech production: a pilot study. *Journal of speech and hearing research 46*, 689-701.

Zaino, C., & Benventano, T. (1977). Functional, involutional and degenerative disorders. In C. Zaino, & T. Benventano (eds.), *Radiographic examination of the aoapharynx and esophagus.* New York: Springer-Verlag.