Considering Accent Recognition Technology for Forensic Applications

Georgina Alice Brown

PhD

University of York

Language and Linguistic Science

October 2017

Abstract

Speaker recognition technology is becoming more available to forensic speech analysts to help to arrive at conclusions around how likely the speech in multiple recordings was produced by the same speaker. However, there is not currently a suitable technological tool that could assist with speaker profiling tasks (i.e. tasks where we wish to deduce information about an unknown speaker). Accent recognition technology could play a role in speaker profiling tasks. This thesis therefore presents numerous automatic accent recognition experiments that have been motivated by forensic applications.

This thesis conducts a detailed examination of one automatic accent recognition system in particular, the York ACCDIST-based automatic accent recognition system (the Y-ACCDIST system). It is trained to assign an accent label to a speaker's speech sample. Unlike other accent recognition system architectures, Y-ACCDIST takes a segmental approach by forming models of speakers' accents using representations of individual phonemes. Implementing a segmentation phase comes at a practical cost, but it is expected that Y-ACCDIST's segmental approach captures a more detailed reflection of a speaker's accent than other accent recognition systems. When classifying speech samples into one of four categories, Y-ACCDIST achieved a recognition rate of 86.7% correct, while the best-performing text-independent system obtained 47.5%.

This thesis also shows Y-ACCDIST's performance on spontaneous speech data. On a three-way classification task on Northern English accents, we witness a recognition rate of 86.7% correct. Additionally, we achieved 63.1% correct when classifying recordings into one of seven non-native English categories. The latter task is also a demonstration of Y-ACCDIST's capabilities on telephone data.

Contents

List of Tables

List of Figures

Acknowledgements

First of all, I would like to thank my supervisor, Dom Watt, for agreeing to support me and my project in the first place. I have particularly appreciated his willingness to promote my work to the research community and the great introduction to academia he has given me. Throughout my time at York, I have thoroughly enjoyed our meetings and really hope we can continue to develop ideas together in the future.

I would also like to thank Peter French for being on my Thesis Advisory Panel. I am grateful for the interest he has shown in my research and the advice and suggestions that have followed.

I have also appreciated the company of Vince Hughes during the times we have muddled through the world of speech technology together. I have enjoyed our discussions and conference trips, which have definitely added great value to this thesis.

I am very grateful to members of the *Audias* and *BiDA LAB* research groups at Universidad de Autónoma de Madrid, who made me feel extremely welcome during my research visit there. In particular, I would like to thank Joaquín González-Rodríguez, Javier Franco-Pedroso and Daniel Ramos-Castro. Their generosity and advice have helped to shape this thesis. I certainly hope this collaboration will continue.

Another collaborator I would like to acknowledge is Jess Wormald. It was a joy to write a paper with her and she has helped me to view my research in a different light.

My undergraduate supervisor, Simon King, has also played a key role in getting me to this point. Simon was extremely patient with me as a linguistics

undergraduate trying to get to grips with speech technology. The skills he passed on to me have been vital to the work presented here.

I also want to take this opportunity to thank the people who kick-started my interest in language, Jill Lavender and Kasia Davies. As my A-Level teachers, Jill and Kasia managed to trigger a momentum that has led to this thesis. I hope that I can similarly inspire and enthuse others in my future ventures.

Additionally, Jill has been a major support throughout my journey so far, and I am so pleased that we have continued our discussions about language (and more) for so long. Thanks for being my pal.

Finally, I would like to thank the Economic and Social Research Council for the Advanced Quantitative Methods scholarship that has enabled me to produce this thesis.

Declaration

I declare that the contents of this PhD thesis are my own original work. This work has not been previously presented for an award at this, or any other, University. All sources are acknowledged as references. Parts of this work, however, have been published. The details of these publications are listed below with details of author involvement:

- Most of the contents of Chapter 2 and a small part of Chapter 4 have been published in the paper below, where I was the sole author:

  Brown G. (2016). Automatic Accent Recognition Systems and the Effects of Data on Performance. In Proceedings of Odyssey: The Speaker and Language Recognition Workshop. Bilbao, Spain. Paper number 29.

- One of the system descriptions in Chapter 2 also resembles the system description in the following publication where I was the sole author:

  Brown G. (2015). Automatic recognition of geographically-proximate accents using content-controlled and content-mismatched speech data. In Proceedings of the 18th International Congress of Phonetic Sciences. Glasgow, UK. Paper number 458.

- Most of the contents of the first half of Chapter 3 (up to Section 3.3) have been published in the following paper, where I was the sole author:

  Brown G. (2016). Exploring Forensic Accent Recognition using the Y-ACCDIST System. In Proceedings of the 16th Australasian International Conference on Speech Science and Technology. Sydney, Australia. pp 305-308.

- Elements of the discussion in Chapter 9 can be found in the following publication where I was the first author:

  Brown G. and J. Wormald. (2017). Automatic Sociophonetics: Exploring corpora with a forensic accent recognition system. Journal of the Acoustical Society of America. 142. pp 422-433.

- Finally, this thesis has been a continuation of work presented in my Masters (by Research) dissertation:

  Brown G. (2014). Y-ACCDIST: An automatic accent recognition system for forensic applications. MA (by research) dissertation. University of York, UK.

Introduction

Thanks to a combination of court rulings (e.g. Daubert v Merrell Dow Pharmaceuticals [1993]) and academic literature (e.g. Saks and Koehler (2005)), we no longer take a forensic expert's testimony at face value. There has been a change in attitude towards how we view evidence presented by an expert witness and, instead, we are much more likely to question the methods he or she has used. This comes under the so-called "paradigm shift" in forensic science (Saks and Koehler, 2005: 283; Kuhn, 1962). Some forensic sciences are more advanced than others in the methods that they standardly use, and which are accepted by the court. DNA typing, for example, has well-established methods and technologies that are now widely accepted by judicial systems, albeit after a lengthy course of scrutiny (Jobling and Gill, 2004).

The area of forensic speech science, on the other hand, has not advanced in the same way as DNA typing. Forensic speech science is the forensic subdiscipline that is largely concerned with speech recordings when they occur as evidence in a case. Forensic speaker comparison cases make up the majority of a forensic speech analyst's workload (French, Harrison, Kirchhübel, Rhodes and Womald, 2017). These are cases where we have an unknown speaker's speech sample (the questioned sample) and a suspect sample (such as a recording of a police interview). The objective is to compare these speech

samples to draw conclusions about whether the speech in the two recordings was produced by the same speaker or not (Nolan, 1983). Automatic speaker recognition technology is becoming more available to assist with this sort of forensic analysis, but it is not currently accepted in the same way as DNA typing methodology. A linguistic-acoustic approach (Foulkes and French, 2012), that involves a significant manual role by the forensic analyst, is very much still a large part of the forensic speech science subdiscipline. Because of the variable nature of speech evidence, this is likely to continue to be the case, but we cannot ignore the potential methodological advantages that technology could bring to forensic speech science. Technology could contribute more objective and testable qualities to forensic speech analysis, which are favourable properties within forensic science, and ones that could complement those of the linguistic-acoustic approach.

While automatic speaker recognition technology is finding its feet within the forensic domain, speaker comparison tasks are not the only type of task that forensic speech scientists are asked to conduct. Forensic science, by its unpredictable nature, presents a whole range tasks to an analyst. Another type of task is *speaker profiling*. This is where we have a recording of an unknown speaker and the aim is to extract information about that speaker, such as age, geographical origin, etc. This is an appropriate analysis when we perhaps do not have a suspect (and therefore no suspect recording) to compare against, and we are simply trying to narrow down the pool of potential suspects (Watt, 2010). A typical scenario for this might be a ransom telephone call made by an unknown speaker, for example. The task of speaker profiling is elaborated on with specific case examples in Chapter 1 below.

At the centre of this thesis is the forensic speaker profiling application. Currently, there has been little research on possible technologies for this kind of

task, unlike automatic speaker recognition technology for speaker comparison cases. As a result, there is not a technological option for this kind of case, and so speaker profiling is solely conducted through a linguistic-acoustic approach. The key motivation behind this thesis is to work towards an additional tool for forensic speech analysts to use in some of their casework involving speaker profiling tasks. This thesis turns to automatic accent recognition technology to explore this prospect.

The bulk of automatic accent recognition research has targeted more general speech technology applications, not necessarily considering forensic casework. The most common target application is automatic speech recognition (Humphries and Woodland, 1997; Zheng *et al*, 2005; Vergyri, Lamel and Gauvain, 2010). We are more likely to achieve a lower error rate if we have correctly estimated the speaker's accent category before the system attempts to recognise the spoken content (Najafian, DeMarco, Cox and Russell, 2014). Automatic accent recognition has therefore often been considered as a step before automatic speech recognition takes place. The focus on the speech recognition application has meant that past automatic accent recognition research design has catered for this cause. Research has not necessarily addressed the kinds of challenges encountered in forensic casework. Nevertheless, automatic accent recognition technology has the testable, repeatable and data-driven properties that methodologies in the forensic sciences should aim for. These are ideal methodological traits that all forensic sciences should move towards, in parallel with the change in attitude by criminal justice systems (as promoted by the Home Office Forensic Science Regulator's Code of Practice (Tully, 2016)). This thesis is therefore devoted to considering automatic accent recognition technology for forensic casework.

One particular automatic accent recognition system will be in focus through-

out this thesis, the York ACCDIST-based automatic accent recognition system (the Y-ACCDIST system). The Y-ACCDIST system is based on the ACCDIST metric (Huckvale, 2004). Y-ACCDIST is a development on past ACCDIST-based systems, in that Y-ACCDIST is designed to be able to process spontaneous speech, whereas ACCDIST-based systems in past research (Huckvale, 2004, 2007; Hanani, Russell and Carey, 2013) have only been able to process controlled speech data in the form of reading passages or read prompts. Crucially, past ACCDIST-based systems have required the spoken content of unknown speakers to match exactly that of the training speakers. Obviously, this kind of data constraint is not compatible with forensic applications. Y-ACCDIST was therefore originally developed in Brown (2014) to test whether we can overcome this limitation without losing the elegant modelling process an ACCDIST-based approach offers. It is a segmental approach that takes into account the specific realisations of speech segments that we expect embodies a detailed representation of a speaker's pronunciation system, more than other system architectures. The details of Y-ACCDIST's inner workings and how it compares with other types of system are presented in Chapter 2. Because of the seemingly fine-grained model of accent Y-ACCDIST forms, it is expected that Y-ACCDIST shows promise for forensic applications. Y-ACCDIST is therefore carried through the entirety of this thesis (after a comparison with other types of accent recognition system), and all research questions are asked of this specific system.

Research Objectives

As indicated above, the overarching goal of this thesis is to further test accent recognition technology with a view to using it for forensic applications. This

thesis breaks this overarching aim down into the following objectives:

1. Test and evaluate a number of different automatic accent recognition system architectures on a dataset of fairly similar accents.
2. Compare automatic accent recognition performance across datasets which present different challenges that are of relevance to the forensic domain.
3. Consider what unknown speech samples should contain to be accurately analysed by an automatic accent recognition system.
4. Apply conclusion frameworks that are widespread across the forensic sciences.
5. Explore whether we can transfer a novel accent recognition system architecture to speaker recognition tasks.

Thesis Outline

This thesis will take the following steps to investigate whether automatic accent recognition technology can be implemented in a forensic context:

Chapter 1 is a literature review that gives further details about the forensic speaker profiling application, the current position of technology in forensic speech science, and a brief introduction to research on accent variation.

Chapter 2 compares six different automatic accent recognition systems on a dataset of accents that are expected to be fairly similar to one another, rather than accents that are very different from one other. This angle separates this

thesis from much of the existing automatic accent recognition research. Automatic accent recognition research tends to try to combat the great variation within a language, because of the focus on the automatic speech recognition application, which suffers due to great degrees of pronunciation differences within the same language. This has meant that corpora of accents that are very different from one another are usually used for this kind of research. When considering forensic applications, we are much more interested in discovering how sensitive accent recognition systems can be, and so Chapter 2 evaluates their performance by testing them on a set of similar accents. The system that achieves the highest recognition rate in these experiments is then taken further through the thesis for a much more detailed examination of its performance and potential.

Chapter 3 focusses on another aspect of data that is likely to be useful to forensic applications. While Chapter 2 presents controlled experiments in which all the speakers are recorded producing the same reading passage, Chapter 3 shifts the focus to recordings of spontaneous speech by testing it on a different corpus of accents. This obviously moves us further towards more forensically-realistic data. Chapter 3 then extends this by artificially degrading the data to a quality resembling telephony. Again, this is with a view to explore accent recognition technology using forensically-relevant recordings. Not only does Chapter 3 observe accent recognition performance under these data conditions, but it also conducts a deeper investigation into how the system models these different data qualities.

Chapter 4 integrates an additional step in the engineering of the accent recognition system: *feature selection*. The primary aim of feature selection here is

to improve system performance. We observe the effects of feature selection on the two corpora used in Chapters 2 and 3 to make a cross-corpus comparison, discovering whether or not it is appropriate to implement feature selection for any dataset.

Chapter 5 turns our attention to the test samples, and asks whether the segmental content of the test sample (i.e. the specific vowels and consonants it contains) affects its likelihood of being correctly classified by an automatic accent recognition system.

Chapter 6 looks at automatic accent recognition system performance on samples of speech produced by non-native speakers of English, testing whether a system can classify speakers according to their first language. This makes use of the large National Institute of Standards and Technology Speaker Recognition Evaluation (NIST SRE) datasets. These datasets are much larger than the datasets used in the thesis up until this point. The size of the dataset opens up the opportunity to make some small changes to the engineering of the system. This allows us to more comprehensively understand the effects on recognition performance. Some engineering modifications are therefore also tested and presented in this chapter.

Chapter 7 integrates the likelihood ratio framework into the accent recognition system. This framework is used widely across the forensic sciences, as it moves us away from making "hard decisions" (i.e. outputting a specific accent label) from an analysis, and instead allows us to make "soft decisions" (i.e. outputting a likelihood of a speaker belonging to an accent category). Considering the sensitivities involved in forensic applications, this approach to

providing conclusions is more favourable.

Chapter 8 explores whether we can repurpose the accent recognition technology tested in this thesis to speaker recognition tasks, given its success at distinguishing between very similar accents.

Chapter 9 gives an overall evaluation of the experiments presented in this thesis, while offering a number of possible avenues for further research.

CHAPTER 1

Literature Review

1.1 Introduction

This thesis sits at the intersection of speech technology and sociophonetics, while targeting forensic applications. This chapter will therefore draw on background literature that covers topics on these three areas, split into three main parts. The first part (Section 1.2) elaborates on *speaker profiling*, the specific application that this thesis aims to provide further assistance with. The second part (Section 1.3) will provide a picture of what role technology currently plays in forensic speech science. This will expose the gap this thesis aims to occupy in exploring accent recognition technology for forensic applications. The third and final part (Section 1.4) draws on literature in sociophonetics, which outlines the challenge that accent variation poses to recognition technology.

1.2 Speaker Profiling

Speaker profiling is the task of deducing information about an unknown speaker, usually from a sound recording of him or her. This kind of information might be a prediction of the speaker's age, sex or geographical origin. This section mainly talks about speaker profiling by trained analysts in the context of forensic and LADO casework, but there is also some discussion about the relevance of speaker profiling by non-expert listeners.

1.2.1 Forensic Speaker Profiling

An early and high-profile case that speaker profiling played a part in was the case of *The Yorkshire Ripper*, where a serial killer committed murders over a number of years during the late 1970s and early 1980s. While the culprit was still at large, a tape recording was posted to the senior investigator of the case. The recording was a spoken message from an anonymous individual claiming to be the Yorkshire Ripper. As a dialectologist, Stanley Ellis was consulted to analyse the recording to determine any identifying properties of the speaker to help reduce the pool of potential suspects for the investigation team. Ellis' account of his analysis and involvement in the case is reported in Ellis (1994). Ellis describes parts of the analytical process he conducted to determine the likely geographical origin of the speaker. He talks about a number of spoken features he identified in the recording, such as /h/-dropping, gradually narrowing down the area the speaker was likely to be from as somewhere in Sunderland. He justifies, through spoken features, why he pinpointed the area he did over other areas of the northeast of England. Ellis' report provides a good example of the kinds of analysis that can go on in a speaker profiling

task.

Unfortunately, in the case of the Yorkshire Ripper, the tape recording was created by a hoaxer, which distracted investigative efforts away from the area where the real Yorkshire Ripper was at large. At the time, the hoaxer was not identified, but decades later, in 2005, he was identified through DNA evidence. At this point, the case became a speaker comparison case, rather than a speaker profiling case (the details of this later analysis can be found in French, Harrison and Lewis (2006)). The individual admitted to the hoaxing offence, and this confirmed that Ellis' earlier speaker profiling analysis was indeed accurate, because this individual was from a part of Sunderland.

A speaker profiling analysis can also play a role in speaker comparison cases. Nolan and Grigoras (2005) make a point of this in their case report of a speaker comparison task concerning an unknown caller making obscene phonecalls to staff working within a bank in London. Despite having recordings of the unknown caller and recordings of two suspects, there was some value in determining whether the unknown caller was a speaker of Australian English or New Zealand English. It was suspected that the unknown caller was a member of staff within the London bank, which employed a number of Australians and New Zealanders. Having knowledge that the primary suspect was a New Zealander (with a typical New Zealand accent), it was important to establish whether the unknown caller was a speaker of Australian or New Zealand English. The authors acknowledge that this distinction is not always clear. After some acoustic analysis, the authors established that it was likely that the unknown caller was also from New Zealand, and so could not eliminate the suspect at this point in the analysis.

Like forensic speaker comparison, speaker profiling involves a combination of aural-perceptual analysis and acoustic-phonetic analysis (Schilling and

Marsters, 2015). Ideally, when conducting a speaker profiling case, an analyst would have access to a representative sample of recorded data obtained from speakers of the 'suspect' varieties, and any sociophonetic research literature or dialectological accounts that are relevant to the suspect varieties. Such resources can lay down the features that we could expect from speakers of the relevant varieties. Observations or measurements (e.g. vowel formant measuerments or observations on intonation patterns) that we gather from the speech recording in question can then be put against these expectations.

Foulkes, French and Wilson (2018) provide an overview of forensic speaker profiling, including some discussion of the Yorkshire Ripper case, as well as other individual speaker profiling cases. Foulkes, French and Wilson move on to talk about a specific type of speaker profiling, known as *Language Analysis for the Determination of Origin* (LADO). In the context of asylum seeker applications, LADO aims to establish where claimants were "socialized" (Cambier-Langeveld, 2010: 68) as just one part of the overall application assessment process, in which there may be some doubt over a claimant's stated origin. Foulkes, French and Wilson highlight the importance of linguists collaborating with native speakers of the spoken variety in a given case, and describe this collaboration as "essential". In their discussion, they describe how this collaborative approach is not necessarily taken by organisations that regularly undertake this kind of work (because there is a lack of enforced guidelines in this area). Foulkes and Wilson (2011), in some experimental work, showed that native speakers of the relevant linguistic variety actually have a lot to offer these kinds of cases. While the area of LADO is very much up for debate about the specific approach that should be taken to offer valid and reliable analyses, there have been recent reports that the German authorities are trialling some technology to assist with these sorts of tasks (e.g. Toor, 2017).

This could provide an additional option for LADO.

Realistically speaking, the work presented here is more likely to benefit the LADO application than forensic speaker profiling for criminal cases. French (personal communication, 2018) points out that a forensic speech and acoustic laboratory in the UK only conducts 3-5 speaker profiling cases per year. On the other hand, there is an overwhelming number of LADO cases that could make use of technology. Like Foulkes, French and Wilson (2018), this thesis considers LADO as a type of speaker profiling and suggests that accent recognition technology might be more appropriate for this cause.

Little research or literature on conducting speaker profiling exists and while there might be some recommendations on how it should be done, practice has not been sufficiently standardised. There may well be room to consider speech technology for these kinds of purposes. One argument for trialling technology for speaker profiling is that it can be effectively tested. This issue will be discussed further in Section 1.3 below, but testing human analysts who conduct this kind of work is very difficult, due to the length of time it can take. We can, however, potentially run thousands of trials using technology, which could shed light on the strengths and weaknesses of the methodology. This is not to say that this is a reason to replace human analysts for speaker profiling, but it is certainly an advantage of using technology in these sorts of cases, where care should be taken.

1.2.2 Speaker Profiling by non-expert listeners

The case examples and scenarios above all involve linguistically trained listeners and analysts. Some experiments have been run to evaluate lay listeners' ability to assign accent labels to recordings of unknown speakers. This could

have some relevance to the forensic domain in the context of witness statements. When a crime is committed, it is not unusual for witnesses to comment on the culprit's accent. This was one of the key foci of Atkinson (2015).

Some research has managed to gauge how lay listeners perform in accent classification tasks, and in some cases report surprisingly low recognition rates. Clopper and Pisoni (2004), for example, conducted human accent classification experiments. They asked participants to assign speakers to one of six North American English categories, using North American English listeners. They report an overall classification rate of 30% correct. The chance level we would expect here is 16.7%. Another similar experiment was by Vieru, de Mareüil and Adda-Decker (2011), who ran some human perceptual experiments on their non-native French accent data. They looked into whether lay listeners could distinguish between native speakers of Arabic, English, German, Italian, Portuguese and Spanish, all speaking French. The listeners were given the chance to familiarise themselves with sample data from these varieties. The overall classification rate to come out of this experiment was 52% correct on this six-way classification task, clearly sitting well above the chance expectation of 16.7% correct. It might be the case that Vieru, de Mareüil and Adda-Decker's non-native accent recognition task was in fact easier than Clopper and Pisoni's native North American English task. A number of factors come into play, however, such as the listener's degree of previous exposure to the different accents involved (which could be linked to the listeners' mobility). These sorts of factors are very difficult to control.

What we should take away from these lay listener experiments is that we should be aware of these kinds of results when considering witness statements which may include an accent recognition claim.

1.3 Technology in Forensic Speech Science

The above section refers to the task of *speaker profiling*, but the majority of cases that forensic speech scientists come across are *speaker comparison* cases (approximately 70% of the overall caseload (French, Harrison, Kirchhübel, Rhodes and Wormald, 2017)). As already stated in the *Introduction* of this thesis, these are cases in which we have two or more speech recordings, and we aim to determine how likely they are to have been produced by the same speaker, against how likely it is that they were not. The linguistic-acoustic approach is an established method of conducting a forensic speaker comparison task. Foulkes and French (2012) provide an overview of the kinds of analyses that can take place using this approach. Techniques might include a *vocal profile analysis* from an auditory analysis of a recording. This is based on the work of Laver (1980), which essentially provides a descriptive framework to gather an overall picture of a speaker's voice quality. For example, the analyst would rate the creakiness and breathiness of the voice, among other qualities. A more acoustic analysis is also typically conducted, wherein, for example, vowel formant measurements are taken from the signal. In sum, the linguistic-acoustic approach measures the individual linguistic feature components that make up the overall speech signal. Together, these separate measurements form a comprehensive picture of a speech recording. This analytical process sets individual recordings up for comparison, ultimately to deliver a conclusion that indicates how likely it is to find the evidence if the speech in the two recordings were produced by the same speaker, relative to how likely it is to find the evidence if the speech in the two samples were produced by different speakers.

Now, automatic speaker recognition is becoming more of an option in this kind of analysis. Automatic speaker recognition research aims to serve applications beyond forensic ones, some of which we might consider comparatively "low-risk" (Broeders, 2001: 54). Commercial telephone services, for instance, could benefit from speaker recognition technology. This might be in the context of resetting an account password down the phone, for example (Kinnunen and Li, 2010). Building security technology could also benefit from speaker recognition technology, where speech might be used as an access medium. Related to this, we are beginning to see some banks integrating speaker recognition technology, whereby customers can use their speech to access their account information (HSBC is one example of this)[1].

1.3.1 For and Against

While the advantages of using speaker recognition technology may include convenience and the prospect of additional biometric support to security access systems, the advantages and disadvantages of using speaker recognition technology for the kinds of forensic applications we have discussed are more complex. Despite trained automatic systems naturally producing outputs faster than a human analyst, convenience and speed should not be the main reasons for using these technologies in the criminal justice system. We of course want to aim to achieve the most accurate and reliable analysis for a given case. In her annual report for 2016, the UK Forensic Science Regulator stressed the importance of using transparent methodologies across the forensic sciences (Tully, 2017). With similar aims to the UK Forensic Science Regulator report,

[1]HSBC website information on Voice ID: https://www.hsbc.co.uk/1/2/contact-and-support/banking-made-easy/voice-id [accessed 13/06/17].

Drygajlo *et al* (2016) developed guidelines for best practice in forensic speaker recognition on behalf of the European Network for Forensic Science Institutes (ENFSI). This was in an effort to standardise practice across analysts, ensuring that responsible approaches to casework are taken, while also reporting as clearly and as accurately as possible to the legal parties involved (e.g. solicitors, judges or juries). We must be able to confidently conduct our analyses knowing the strengths and weaknesses of our tools and techniques. In favour of using technology with regard to methodological transparency is the fact that we can run large numbers of tests to determine a system's performance under specific conditions. This point was made in the section above. We can test automatic systems to an extent that we cannot test human analysts, due to the time required to do the latter. An argument against using technology with respect to methodological transparency is that these systems may be branded as "black boxes". In other words, it could be claimed that the user of these tools has insufficient knowledge or control over the inner workings of these systems. Alexander, Forth, Atreya and Kelly (2016) address this concern by presenting their automatic speaker recognition software as an "open box", which aims to provide an analyst with as much flexibility and control as possible surrounding the configurations of a single analysis. They demonstrate that such flexibility can be possible.

Another advantage of using technology in forensic tasks is that there is a lower risk of bringing bias into an analysis. There is a body of work that looks at the effects of *cognitive bias* on the work of forensic analysts across the forensic sciences (Dror, Charlton and Péron, 2006; Dror and Hampikian, 2011; Kassin, Dror and Kuckucka, 2013; Zapf and Dror, 2017). In these studies, they demonstrate how contextual information about a specific piece of evidence can affect the conclusion a forensic analyst might put forward. In

Dror, Charlton and Péron,(2006), for example, this was done in the context of forensic fingerprint analysis. Dror (2015) offered a number of ways in which forensic analysts across the subdisciplines can alleviate some of the effects of cognitive bias. Among these were integrating thorough checking procedures within the overall process and incorporating technology into our analyses. In the specific area of forensic speech science, contextual information may well be intertwined with the speech evidence itself. While an analyst mainly needs to focus on the spoken features of a recording, regardless of what is being said, the spoken content is difficult to ignore, and so taking in context can be unavoidable. For this reason, Rhodes (2016) points out that forensic speech science is at particular risk of the effects of cognitive bias. Technology obviously is not affected by contextual bias in the same way. Speaker recognition technology cannot be biased by the meaning of the spoken content of the speech samples it analyses and so this could be seen as one argument for using technology in an analysis.

Despite the advantages of technology's relative objectivity, there are still issues around submitting automatic analyses as evidence. In a recent UK ruling, Slade & Ors v Regina [2015], an automatic speaker recognition system was used to analyse the speech evidence in a court case. The judge ruled that the evidence could not be admitted for the case because of the level of uncertainty around the performance of these systems on forensically relevant recordings. While it is in our interests to use the most effective methodologies for a forensic analysis, it is also important to have a comprehensive understanding of these methodologies under the specific conditions of the case at hand. It is also imperative to be able to communicate how a methodology works to legal professionals and juries. This is another responsibility of the forensic expert, and could be viewed as a challenge in the future as he or she will be required

to be able to sufficiently and successfully explain how these technologies work to non-expert audiences.

This subsection has largely talked about technology's role in forensic speech science, with reference to automatic speaker recognition technology. However, it is likely that other new technologies, such as accent recognition technology, if implemented, would face similar challenges.

1.3.2 Development Data

To develop the technology, we of course require suitable databases for training and testing. Similarly, relevant datasets are required for more manual approaches to forensic analysis (and this will be discussed further below). Overwhelmingly, automatic speaker recognition studies make use of datasets made available by the National Institute of Standards and Technology (NIST)[2]. Every one or two years, NIST releases a Speaker Recognition Evaluation (SRE) dataset for automatic speaker recognition researchers from across the globe to train and test their systems and approaches. They are then required to submit their results by a given deadline. This allows for all these systems to be directly compared with one another. NIST datasets are extremely large, now making available telephone recordings of thousands (3000 +) of different speakers for experiments. This great volume of data allows for certain system architectures to be sufficiently trained and tested (such as i-vector-based systems (Dehak, Kenny, Dehak, Dumouchel and Ouellet, 2011) and Deep Neural Networks (DNNs) (Lei, Scheffer, Ferrer and McLaren, 2014). The performance of these sorts of systems suffers with small datasets, but such systems are very effective when trained on large volumes of data. Liu and Hansen (2011) em-

[2]See: https://www.nist.gov

phasise the need to test technologies on smaller datasets, in the context of more niche recognition problems. They state that, in reality, we typically do not have access to large databases of relevant accents and dialects. It is more realistic to expect smaller datasets to work from.

In a number of ways, these large NIST SRE datasets are very good for exploring a range of different system architectures and various aspects of performance. The fact that these datasets are largely made up of telephone recordings means that, in some respects, these datasets are relevant to forensic applications. However, in reality, forensic casework can be very specific in terms of the type of data that is relevant to a particular case, and we know that the performance of automatic speaker recognition systems can suffer considerably when there is a mismatch in recording quality and/or recording type between the training and test samples (Alexander, Botti, Dessimoz and Drygajlo, 2004; Rajan, Kinnunen and Hautamäki, 2013). In fact, one of the reasons why there was uncertainty surrounding the evidence presented in the Slade & Ors v Regina [2015] case discussed above was because the recordings involved were captured inside a car, and testing of automatic speaker recognition systems is not standardly done on recordings of this very specific type. Developing technology and solutions for forensic applications therefore requires training and testing of systems on a broad spectrum of databases that represent different types of speech, speakers and dataset sizes. Considering technology and techniques for forensic applications, there have been researchers who have developed speech databases for very specific scenarios. An example of a specialised corpus is one collected and researched by Fecher (2014). The purpose of this work was to conduct a phonetic analysis of speech affected by various items of 'facewear' (e.g. balaclavas and motorcycle helmets). Again, considering forensic applications, it is important to discover the effects of these

conditions on speech samples to be able to more confidently comment on how an analysis might be affected by these kinds of complicating factors.

In addition to these rather niche datasets, a number of databases exist that were originally collected for sociophonetic research purposes that could also be of value to forensic casework. The NIST dataset largely offers North American English speech, which of course provides a very good resource in the context of casework involving North American English. However, we need a resource to account for the spoken variety that is involved in a given case, and so we can turn to various speech databases collected by sociolinguists. For this reason, Hughes and Wormald (2017) make a call for more collaboration between the two areas of sociolinguistics and forensic speech science. They refer to the "paradigm shift" in forensic science and the need for more relevant data to be able to conduct the data-driven analyses that are expected of forensic speech scientists.

There is one database in particular that has been repeatedly used for forensic speech science research. To simulate as closely as possible the most typical forensic scenario, Nolan, McDougall, de Jong and Hudson (2009) created the DyViS database. This corpus consists of samples of speech produced by over 100 young adult male speakers (the demographic most often encountered in forensic casework) in a number of mock (but forensically-realistic) scenarios. In particular, it includes recordings of the same speakers making a telephone call and answering police-style questions in a police interview room. This comparison task (between a telephone recording and a police interview) is typical of forensic casework. It is therefore important to conduct research on this very specific mismatch in recording type. Although the DyViS corpus offers a number of features that are useful to forensic research (both manual analysis and automatic analysis), one criticism of the corpus is the specific spoken variety

of the speakers. The young male informants were recruited from Cambridge University, and so a large number of these speakers probably do not represent the demographic of speakers that is regularly encountered in forensic casework.

The accent recognition experiments presented in this thesis make use of three different datasets which offer a variety of conditions that we can train and test the automatic accent recognition systems under. One key aim of this subsection is to illustrate that a cross-corpus investigation is important to be able to uncover the strengths and weaknesses of a system, so we can be as transparent as possible about the methodology. However, the datasets obviously cannot reflect all possible forensically relevant scenarios. It is therefore of interest to observe how a system or methodology transfers between different types of dataset to assess how performance is affected. This comparison of system performance on different corpora is discussed at numerous points throughout this thesis.

1.4 Accent Variation

This thesis aims to characterise accent variation using automatic methods. We know that all kinds of speech features can vary across speakers. When we are looking at accent variation, we want to just identify those that are characteristic of whole speech communities. A speech community is not defined by a specific factor that speakers in a community share, but rather a group of speakers share some trait to unite them in some way. This deliberately makes way for quite a broad interpretation of what makes an accent (i.e. accent is not limited to geographically defined speaker groups). Some examples of ways in which groups of speakers can form a speech community are given below in Section 1.4.1.

The variables that mark an accent variety are often segmental (for example, the presence or absence of non prevocalic-/r/ that distinguish rhotic accent varieties from non-rhotic ones). Academic studies of accents commonly include measurements and analyses of vowel formants, which indicate the 'quality' of vowels. In the context of English, linguists will often refer to Wells' (1982) keywords to pinpoint vowels of interest. For example, studies have been interested in how far forward in the vowel space some groups of speakers produce their GOOSE vowel (e.g. Haddican, Foulkes, Hughes and Richards (2013)). More specifically, measuring and comparing the second formant values from speakers' productions would allow us to observe whether the GOOSE vowel is fronted or not. By referring to keywords like GOOSE, linguists can include a number of words that contain vowels that fall into the same phoneme class. For example, we could focus on the vowels in *boot* and *tune* to obtain measurements for a speaker's GOOSE vowel.

Features of voice quality tend to be associated with individual speaker variation, one reason being due to the physiological differences between speakers. However, some work has attached voice quality features to whole speech communities. As Stuart-Smith (1999) highlights, some voice quality features are attributed to the physiological makeup of the speaker, while there are others that have been 'acquired' where speakers implement muscular settings. It is this latter case that could lead to certain voice quality features being linked to specific speech communities. Voice quality seems like a difficult aspect of speech to characterise, but the work of Laver (1980) has instigated ways of decomposing voice quality into numerous features, resulting in a method known as a *Vocal Profile Analysis* (VPA). Stuart-Smith (1999) suggests that Glasgow speakers share a particular constellation of voice quality features, based on a VPA. She makes further distinctions, however, within the Glasgow speaker

group, classifying speakers into working class or middle class groups. In her analysis, she found a number of voice quality differences between these groups of speakers, the most prominent being that working class speakers exhibit more whispery voice than middle class speakers. Beck and Schaeffler (2015) conducted a voice quality study of adolescent speakers of Scottish English, but reported that they did not find any significant voice quality differences between speaker groups determined by geographical origin. They did, however, find significant voice quality differences between male and female speakers (features that are not necessarily determined by obvious physiological differences between these two groups). Voice quality features have also been found to characterise Liverpool English (Knowles, 1973) and Leicester and Bradford Punjabi English (Wormald, 2016), for example. Depending on the accent groups that we are interested in, the literature indicates that voice quality features could also assist in distinguishing between accent groups.

There are also prosodic cues that could reveal a speaker's accent group (Peppé, Maxim and Wells, 2000; Clopper and Smiljanic, 2011). By measuring and analysing pitch contours throughout speakers' utterances, studies have uncovered categorical patterns between speaker groups.

Different types of automatic accent recognition system aim to capture these different features of accent variation so as to classify speakers, and they are expected to take advantage of different types of variable. We see some of this system range among automatic systems in Chapter 2. This subsection aims to outline the reasons for accent variation and to briefly review some of the sociolinguistic research literature.

1.4.1 Factors Responsible for Variation

Within the field of linguistics, the study of accent variation comes under the term *sociophonetics*. Foulkes, Scobbie and Watt (2010) highlight the challenge of defining sociophonetics because of the number of layers of variation and the angles from which it can be studied. Probably the first type of accent variation that comes to mind is regional variation. Regional variation is concerned with sociolinguistic variables that are shared by speakers of a specific geographical origin. This is the type of variation that is at the centre of this thesis. However, sociophonetic variation extends way beyond geographical boundaries. Social class is another key factor that has received a lot of research attention, as well as speakers' age and sex (Foulkes, Scobbie and Watt, 2010). Lots of other contributing factors have also been researched, however, such as ethnicity (Alam and Stuart-Smith, 2011), sexual orientation (Mack and Munson, 2012) and even political affiliation (Hall-Lew, Friskney and Scobbie, in press).

Contact varieties have also been a topic of focus in the sociophonetic literature. This is where an accent variety has developed as a result of two or more communities coming into contact. One example of this within British English is Multicultural London English (MLE) (Cheshire, Kerswill, Fox and Torgerson, 2011). Similarly, Wormald (2016) sociophonetically dissects varieties of British English that have been influenced by a heritage language. Specifically, she looks at a number of variables in the speech of Panjabi-English speakers in two English cities, Bradford and Leicester. Conducting a parallel analysis of these speakers in two cities allowed Wormald to comment on which variables used by first-generation and second-generation speakers of Panjabi-English are produced as a result of influences from the heritage language (Panjabi), and which are produced as a result of influences from the local regional variety.

In addition to these kinds of factors that separate the pronunciation of some speaker groups from others, we can also talk about *intra-speaker* phonetic variation. A number of factors (such as the speaker's emotional state or level of tiredness) can be responsible for causing an individual's speech production to vary from one point in time to another. We can refer to this kind of data as *non-contemporaneous*, and it is a type of variation that is very much a concern for forensic speech scientists.

There has also been research into how a speaker's pronunciation changes across his or her lifespan. This type of intra-speaker variation has featured within the forensic speech science research literature. Rhodes (2012) and Kelly (2014) consider the effects of aging on speakers' speech from a linguistic-acoustic perspective and from an automatic speaker recognition perspective. This is relevant to cases in which a considerable amount of time has passed between the creation of suspect and unknown recordings (as was the case with the John Humble case mentioned above in Section 1.2.1). It is acknowledged within the research literature that speech properties are expected to change throughout a speaker's lifetime, due to physiological changes. Xue and Hao (2003) showed that the volume of the vocal tract tends to increase with age. This kind of change naturally has repercussions on acoustic features produced. Xue and Hao measured an overall decrease in the formant frequencies of vowels of by older speakers. Within the sociolinguistic literature, Sankoff and Blondeau (2007) conducted a longitudinal study of /r/ production among Montreal French speakers. There are two possible phonetic realisations of /r/ in Montreal French that vary in their place of articulation: the alveolar trill [r], and the uvular trill [ʀ]. They collected speech data from the same individuals at two points in time, once in 1971 and again in 1984. They showed a definite change in /r/ production within these individuals over this 13-year time pe-

riod. An overall finding is that speakers increased their use of the [ʀ] variant over time, such that the mean proportion of this variant was 63.8% in 1971 and 77.8% in 1984. A study like this allows us to witness language change occurring within speakers. Naturally, some speakers adopt a change like this more than others due to a number of factors (speaker contact and mobility, for example). Again, this sort of intra-speaker variation can be an important consideration when conducting forensic casework.

The current subsection only gives a glimpse of some of the layers that contribute to the overall speech production of an individual. It should, however, introduce the complexity we can expect in the pronunciation of an individual. This thesis simply targets one of these factors: regional variation. It is impossible to account for all of the potential factors that might influence a speaker's production in accent recognition experiments, particularly when we wish to maximise the quantity of data we can collect to represent an accent group. However, it is important to keep these additional factors in mind as they can help to explain the variation within a single speech community and the extent of the challenge of the accent recognition problem.

1.4.2 Approaches to Sociophonetic Research

The early influential work of Labov (1963, 1966) prompted the dialectological research community to adopt a largely quantitative approach to dialectology. This usually involves the selection of one or a few linguistic variables (e.g. the centralisation of diphthongs in the case of Labov (1963)) that are hypothesised to vary across speaker groups, the collection of appropriate data to elicit tokens of these variables, and the empirical analysis of the variables across the speech communities of interest.

In much of the sociolinguistic research that is carried out, only a small number of linguistic variables are selected to compare the speech of different communities. The selection of variables is often based on auditory observation by the researcher or on findings reported in the existing research literature. Nerbonne (2009) advocates an "aggregate" approach, whereby we take whole collections of features to analyse sociolinguistic variation. He argues that single-feature analyses are very likely to miss important aspects of variation and, as a result, are more unreliable. In a way, the main accent recognition system tested in this thesis employs an aggregate approach to analysing sociolinguistic variation.

A lot of research takes measurements of speakers' speech production to analyse variation, but there is also research that makes use of the perceptions of human listeners as a measurement tool. Perhaps a more obvious use for listener perception is to uncover social judgements about speech communities (Campbell-Kibler, 2010; Giles and Billings, 2004). However, listener perception can also be used to analyse specific linguistic variables that might be distinctive of particular spoken varieties (e.g. Clopper and Pisoni, 2004).

We can view the automatic accent recognition systems presented in this thesis as further ways of conducting sociophonetic analyses. It might, however, be unclear which specific variables these systems use to distinguish between accents. At points throughout this thesis, we will look to the sociophonetic research that has been produced by these kinds of sociolinguistic methods to try to make the inner workings of these technologies more transparent.

1.5 Summary

This chapter has pointed to where the topic of accent recognition technology sits within the current research climate with regards to speech technology, forensic applications and sociophonetics. In a nutshell, the large proportion of automatic accent recognition research has not considered factors that might concern forensic applications. This thesis aims to counteract this trend.

With the call for transparency in forensic methodologies from regulatory organisations, it is important to understand our analytical processes, to a high level of detail, and so getting familiar with the relevant sociophonetic literature could help us to do this for accent recognition systems. By making reference to the sociolinguistic literature, we can gather an idea of what we can expect from an accent recognition system, and try to determine the strengths and weaknesses of these technologies.

CHAPTER 2

A Comparison of Automatic Accent Recognition Systems on Geographically-Proximal Accents

2.1 Introduction

The experiments in this chapter aim to challenge and compare a range of different types of automatic accent recognition system in the context of geographically-proximal accents. As discussed in Chapter 1, much of the past automatic accent recognition research has been motivated by the automatic speech recognition application. One factor which is problematic to successfully recognise speech in automatic speech recognition is the great variation in a single language. It is quite conceivable to think of numerous words which have very different realisations in different accent varieties of the same language, and the challenges all these variants bring to speech recognition models. For example, the English word *pot* in a typical North American English accent has a

similar realisation to *part* in a Standard Southern British English accent, phonetically transcribed as [pʰɑt]. For speech recognition applications, we can see how differences between quite distinct accents could cause problems. Indeed, if we try to overcome this great variation by first identifying the speaker's accent group, and adapting the models accordingly, speech recognition rates are seen to improve (Najafian, DeMarco, Cox and Russell, 2014).

For other applications it might be of interest to investigate accent recognition systems' performance on accent varieties which are not as distinct from one another as the varieties used in much of the accent recognition research. A tool for forensic applications is more likely to be useful if it can model a collection of differences which are more subtle, and perhaps less well-known in the linguistic literature, rather than simply distinguishing between speakers of North American English and Standard Southern British English, for example. Because of the low level of difficulty involved in this kind of distinction, developing a tool to group speech samples into one of these two categories would not be particularly useful to forensic experts, who would easily be able to make this kind of classification. We should therefore test systems on tasks that are expected to be more difficult to forensic analysts. This is the main motivation behind the experiments in this chapter.

In style, the experiments shown here closely simulate those presented in Hanani, Russell and Carey (2013). In their study, they test a number of automatic accent recognition systems on the same accent corpus, the *Accents of the British Isles* (ABI) corpus (D'Arcy, Russell, Browning and Tomlinson, 2004). The ABI corpus contains recorded data from speakers in 14 locations across the breadth of the British Isles. For each location, they collected data from 10 male speakers and 10 female speakers. The recorded data consist of speech samples of the speakers reading the same prompts, so spoken content

is comparable and controlled across samples, providing a good experimental basis to compare systems. Using these data, Hanani, Russell and Carey (2013) compared the capacity of a number of different systems to classify speakers into one of the 14 categories. The range of systems they tested contained both *text-dependent* and *text-independent* systems. These two terms will be defined and discussed in more detail in Section 2.3.1, but for now, text-dependent systems are those which require a transcription of the spoken content as input to accompany the speech sample, whereas text-independent systems do not. The experiments presented in the current work take four of the accent recognition systems which were tested in Hanani, Russell and Carey (2013) and develop similar versions to then test them on a corpus of accents which is expected to be more challenging in that the varieties are predicted to be more similar to one another than those in the ABI corpus. The corpus chosen for this purpose is the *Accent and Identity on the Scottish/English Border* (AISEB) corpus (Watt, Llamas and Johnson, 2014), which will be described in detail in Section 2.3.2 below. It is a corpus of *geographically-proximal* varieties that we are assuming are more similar to one another than the accent varieties in the ABI corpus. We should, however, emphasise that geographically-proximal accents are not necessarily always more similar to one another than geographically non-proximal varieties. Likewise, accent varieties that are geographically further away from each other can be phonologically similar. For example, Ulster Scots and Scots varieties are thought to be similar due to their shared history of the same broad group of settlers (Montgomery, 2001). Two other systems from other studies (Wu, Duchateau, Martens and Compernolle (2010) and Najafian, Safavi, Weber and Russell (2016)), have influenced two further systems developed and applied for comparison in this chapter. This is to provide a substantial range of different accent recognition system architectures

for comparison using the same corpus of geographically-proximal accents. A total of six systems have been built and tested.

2.1.1 Outline

This chapter will first review past approaches to automatic accent recognition taken in other studies. We will then specify the details of the experiments presented in this chapter, which include descriptions of each of the systems developed and tested (in Section 2.3.1) as well as the AISEB corpus (Section 2.3.2). Before giving the results obtained by each of these systems on the AISEB corpus, Section 2.3.3 will first provide results that have been generated by similar systems in past studies which use different corpora. These results gathered by past studies provide a context in which to analyse and discuss the results generated by the systems using the AISEB corpus. Section 2.4 will then discuss the overall findings that we can draw from the experiments presented in this chapter.

2.2 Past and Current Approaches to Automatic Accent Recognition

In the *Introduction* above, it was mentioned that past automatic accent recognition studies have involved accents with a large number of linguistic differences between them. This has meant that certain past system architectures have been reasonably successful in distinguishing between these accents. This section provides an overview of a number of different system types developed in past studies, as well as the nature of the accent data that some of these systems have been tested on. To guide this discussion, this section broadly

divides system types into two categories. We first review *phonotactic systems* in Section 2.2.1 and then move on to *acoustic systems* in Section 2.2.2.

2.2.1 Phonotactic Systems

Earlier accent recognition systems incorporated methods from Language Identification (LID) technologies. Zissman (1996) was a key study in exploring a phonotactic approach to the task of LID, which looked at *Phone Recognition followed by Language Modelling* (PRLM) systems. Given an unknown utterance, a PRLM system first estimates its phone sequence using a phoneme recogniser. Making use of this phone sequence estimation (as well as the number of occurrences of each phone), the likelihood of this sequence appearing in each of the reference languages in the system is calculated. PRLM systems heavily rely on the different candidate languages having sequences and distributions of phones which are distinctive enough to be able to discriminate languages. For this reason, we can expect that using phone sequences to distinguish between different accents might be much more challenging than it is in a LID task. Depending on the selection of accent varieties we are interested in, it is not likely that there will be many sequential differences to separate different varieties. We might expect accent features like rhoticity to be picked up on, because in rhotic accents, /ɹ/ occurs in specific contexts that it does not in non-rhotic accents. A PRLM system might therefore be able to distinguish between rhotic and non-rhotic accents. However, often, accents are characterised by the realisations of certain phonemes, rather than the sequences they occur in. For example, it might be helpful to determine the precise quality of /ɹ/-production for a given accent classification task.

There have been some studies into the prospect of using phonotactic ap-

proaches in dialect recognition. For example, Biadsy, Soltau, Mangu, Navratil and Hirschberg (2010) show the performance of these kinds of phonotactic approaches on the task of distinguishing between four dialects of Arabic: Iraqi Arabic, Gulf Arabic, Levantine Arabic and Egyptian Arabic. They claim that these dialects are distinguishable by the varieties' phone sequences. Because of the distinctive nature of these Arabic varieties, Biadsy *et al* achieved an Equal Error Rate (EER) of 6.0% on this classification task. While these PRLM systems might be successful on this particular Arabic dialect classification task, we cannot assume it will be successful when applied to other datasets of dialect varieties.

The main focus of this chapter is to test systems on geographically-proximal accents, where heightened levels of similarity between the varieties are assumed. It is reasonable to expect that this phonotactic encoding has little to offer an accent recognition task of this kind. It is predicted that the phone sequences themselves are too similar and this type of difference would be too subtle to confidently discriminate accents. We should also be mindful of the errors that the first stage of phone recognition would also inevitably bring to the process, possibly further diluting the already subtle or scarce accent differences. We might therefore suggest that more attention should be devoted to the phonetic realisational differences between very similar varieties when building a system for this purpose. Acoustic systems might be more appropriate to uncover these phonetic realisational differences.

2.2.2 Acoustic Systems

To be able to model phonetic realisational differences, we need to move towards acoustic approaches, rather than simply estimating which phones are

present or absent in an utterance's phone sequence (as suggested in the section above). This subsection therefore describes some of the tools and techniques which are, or have been, employed to acoustically characterise and classify speakers' accents. We can broadly look at acoustic accent recognition systems in three main stages. In order, these are *feature extraction*, *accent modelling* and *classification*. Each of these stages are discussed separately in turn below.

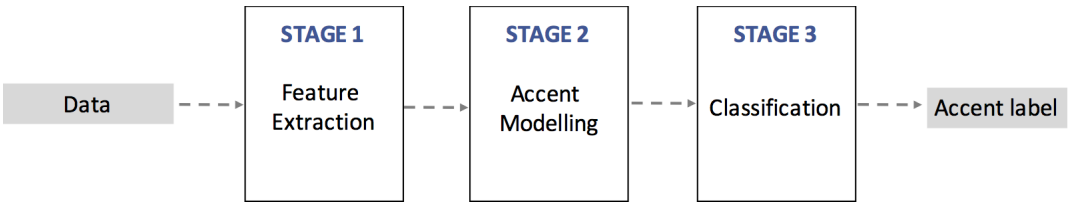

Figure 2.1: A broad illustration of the three main stages of automatic accent recognition using an acoustic approach.

1) Feature Extraction

The first step in an acoustic system is information reduction of the raw signal, for both the training data and the testing data. We therefore need to extract acoustic features to represent the signal. The most common type of acoustic features are Mel Frequency Cepstral Coefficients (MFCCs) (Davis and Mermelstein, 1980). These are acoustic feature vectors which are widely used across speech technology as a whole. They were initially intended for speech recognition, but are also used in areas like speaker recognition and speaker sex recognition.

To generate MFCCs, the following steps are taken. As a preprocessing stage, we apply *preemphasis* to the signal to overcome *spectral tilt*. Spectral

tilt is simply the natural distribution of energy across the frequencies of voiced sounds. There is less energy at the higher frequencies, so to ensure we do not overlook potentially useful information at these higher frequencies, we can apply preemphasis to boost the energy.

In most systems, MFCCs are extracted right across the speech sample at overlapping intervals. A standard configuration for MFCCs is to extract them from 25ms windows of speech, every 10ms. To extract the spectral information from each of these windows of speech, a Discrete Fourier Transform (DFT) can be applied to each window[1]. A DFT enables the extraction of the magnitude of different frequency components. Because acoustic information at all available frequency bands is not necessarily useful for characterising segmental information, applying a mel-spaced filterbank to the window would include a higher concentration of phonetically informative values in the overall feature vector. This is because more segmental information is found at the bottom of the spectrum and mel spacing brings more extraction points at these lower frequencies. The magnitude can then be extracted from these mel-spaced intervals. The distribution of a mel-spaced filterbank is demonstrated in Figure 2.2.

[1]In many versions of the process, a *Fast Fourier Transform* is applied as an alternative transform for this purpose.

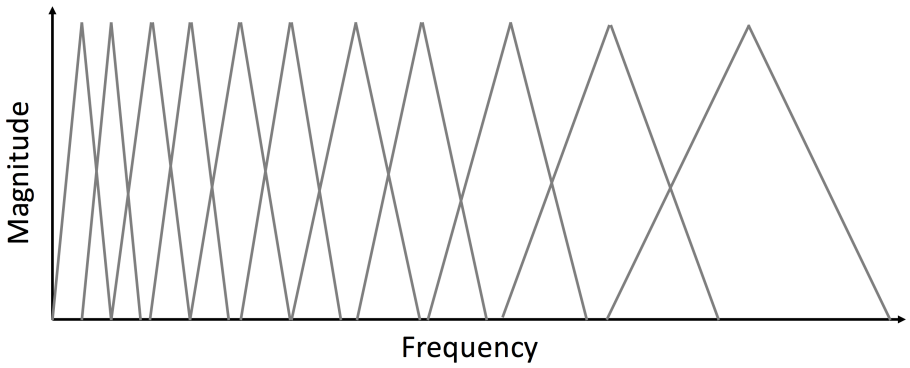

Figure 2.2: Illustration of the distribution of mel-spaced filterbanks.

A mel-spaced filterbank means that more information will be extracted from lower down in the spectrum and fewer extraction points exist higher up in the spectrum. This reflects the distribution of useful segmental information and is also thought to be an approximation of the human perceptual system. This is because of the "resolving power" of human ears throughout this distribution of frequencies (Holmes and Holmes, 2001: 160). The logarithm of these energy values is then taken before a Discrete Cosine Transform (DCT) is applied to output the *cepstrum*. The cepstrum organises the acoustic information into two categories: acoustic information determined by the filter (i.e. the shape of the vocal tract manipulated by the articulators) and information determined by the glottal source (Jurafsky and Martin, 2009: 335). Since it is the filter information which provides segmental information, we only take the values in the cepstrum associated with filter information. The first 12 values are therefore taken to form the resulting MFCC vector. It is thought that the values beyond this reflect the acoustic information which is determined by the source. The overall MFCC extraction process is summarised in Figure 2.3:

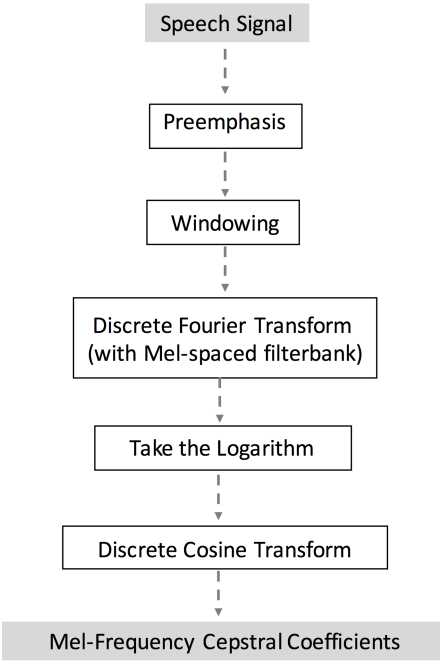

Figure 2.3: Flow diagram of the processes involved in MFCC extraction.

In addition to the 12 filter-based values described above, we can introduce dynamic information to the MFCC representation. We can add *delta* coefficients which effectively track the change in the coefficients. To do this, delta coefficients simply log the difference between corresponding coefficients in neighbouring windows of the signal. *Delta delta* coefficients can also be introduced to add further dynamic information. In a similar way, these capture the difference between corresponding delta coefficients in neighbouring windows of the signal.

MFCCs are not the only feature type that we could use in our systems. Perceptual Linear Predictive coefficients (PLPs) (Hermansky, 1990) are an alternative feature type which have been seen in a number of studies (e.g. Müller and Mertins, 2011), to outperform MFCCs under certain conditions,

mainly under more noisy conditions. Shifted Delta Cepstra (SDCs) are another example of an alternative feature type used in speech technology. In fact, Behravan, Hautamäki and Kinnunen (2015) chose to use SDCs in their accent recognition work. SDCs are effectively a concatenation of delta cepstra to place a greater emphasis on the dynamic information encoded in speech. Torres-Carrasquillo *et al* (2002) demonstrated that SDCs can be an effective feature in Language Identification tasks. As a parallel in the forensic phonetic literature, dynamic formant measurements have been of interest and have served to generate reliable results (e.g. McDougall and Nolan (2007)). It is thought that the dynamics of speech can be very informative in speaker comparison casework, and could add value to evidence beyond that of just taking midpoint formant values. Therefore, in the context of forensic speech technology, it might be worth considering these features (SDCs) which aim to offer more dynamic information, but this is a consideration for future research since this thesis does not focus on a comparison of feature vectors.

While investigating different features predominantly seeks to improve overall performance, or to equip systems to perform better under certain conditions, other motivations behind feature research might be to lower the overall computational cost of a system. For example, while Poblete *et al* (2015) have aimed to develop a new high-performing feature for speaker recognition (Locally-Normalized Cepstral Coefficients (LNCCs)), they also highlight that LNCCs come with a lower computational cost, which is of course an advantage.

These are just a few examples of alternative features that could be used, but many more types and variations are possible. Throughout the experiments presented in this thesis, MFCCs will be used as these are in line with much of the past accent recognition research. Once acoustic features have been extracted, they are then used for accent modelling, which is described and

discussed in the next part of this section.

2) Accent Modelling

The modelling stage involves taking the extracted features and using them to form a representation of the speech signal, which aims to capture the aspects relevant for the particular purpose of a system. For example, for a speaker recognition system, we want to form a model of the signal that best represents the distinguishing features of a speaker. Likewise, for accent recognition systems, we want a model that best represents distinguishing features of an accent. However, features that are relevant to other speech characteristics that may be embedded within the speech signal (e.g. speaker sex and age factors) can interfere with developing effective models. Because MFCCs are expected to be particularly effective at capturing segmental information, it is reasonable to expect that they are an appropriate measure to use for accent recognition. This subsection introduces the three modelling methods applied in the experiments presented in Section 2.3: Gaussian Mixture Models (Reynolds and Rose, 1995), i-vectors (Dehak, Kenny, Dehak, Dumouchel and Ouellet, 2011) and ACCDIST matrices (Huckvale, 2004, 2007).

a) Gaussian mixture models

An extremely common model in speech technology, and one employed in three of the accent recognition systems tested below, is the Gaussian Mixture Model (GMM). GMMs are probabilistic models that characterise data where it is assumed that we have multiple normally distributed subpopulations. Taking in feature vectors (in this case, MFCCs), a GMM can produce a typical representation of the data we are trying to characterise. A GMM is a collection

of gaussian distributions (together, forming multivariate gaussians) representing the distributions of the cepstral coefficients. They are composed of the individual means for each coefficient and a covariance matrix.

In the context of forensic speech technology, GMMs have been the central component of much of the automatic speaker recognition systems used. In fact, Kinnunen and Li (2010: 20) refer to it as the "de facto reference method in speaker recognition". This is in relation to widely implemented GMM-UBM (Gaussian Mixture Model Universal Background Model) systems, which is where the GMM models are adapted from a UBM which is typically a very broad representation of speech trained on masses of data.

Because of their success in speaker recognition, GMMs have naturally been applied to the task of automatic accent recognition. Chen, Huang, Chang and Wang (2001) were targeting the problem that the great variation among Mandarin Chinese dialects brings to automatic speech recognition systems. To approach this particular problem, they built and tested variants of a text-independent GMM-based system. They focussed on separate results for male and female speakers, because speaker sex is known to affect the GMMs (because this is another speaker property that is embedded in the speech signal). They observed error rates of 15.5% and 11.7% for males and females respectively. Considering these GMM-based systems do not depend on pre-defined segmental units, as they are text-independent systems, these error rates seem quite promising.

b) i-vector

i-vector-based systems are now widespread in automatic speaker recognition (Dehak, Kenny, Dehak, Dumouchel and Ouellet, 2011) and language identification (Dehak, Torres-Carrasquillo, Reynolds and Dehak, 2011). An i-vector-

based system uses a distinctive modelling approach to a problem. An i-vector is a compressed low-dimensional representation of a speech sample, which is then used for the scoring or classification stage of a process. Further details of how i-vectors are extracted are given below in Section 2.3.1. This subsection reviews the performance of i-vector-based systems in past research.

Having witnessed the relative success that i-vector systems have had in automatic speaker recognition, it is reasonable to assess an i-vector-based approach to accent recognition. Behravan, Hautamäki and Kinnunen (2013) tested different variants of an i-vector accent recognition system on speech samples drawn from two different corpora, while also comparing i-vector technology against a more traditional GMM-UBM system. The two different corpora aimed to address two slightly different types of accent recognition task. The first corpus used for their experiments was the *CallFriend* corpus (Canavan and Zipperle, 1996), which provides two native dialects of different languages (i.e spoken varieties that native speakers of that language produce, rather than non-native varieties). They selected English, Mandarin and Spanish to run experiments on. The second corpus was the Finnish national foreign language certificate corpus (FSD), which provides recordings of 'foreign-accented' Finnish speech from speakers of various different native languages. The aim of an accent recognition task using this corpus is to identify the native language of a test speaker. We can treat this as a slightly different task in accent recognition, as the second task of 'foreign-accented' recognition involves an extra factor of linguistic proficiency. It is assumed that the speakers in this corpus speak Finnish at differing levels. There are more factors that come into play with regards to non-native accented speech. This topic is dealt with in further detail in Chapter 6 of this thesis. Behravan *et al*'s (2013) findings showed that in both types of accent recognition task, the i-vector system outperformed the

GMM-UBM system. The i-vector system achieved Equal Error Rates (EER) of 15.06% and 20.01% for the *CallFriend* and FSD corpora, respectively, whereas the GMM-UBM system achieved 18.73% and 24.13% EER on the same tasks. The additional factor of language proficiency in the FSD data is likely to affect automatic accent recognition performance by perhaps introducing more variability within the models, although it is not necessarily clear in what respect it might do so. A difference in the size of the datasets could also contribute to the discrepancy in performance between the two corpora, but it is unclear whether this was a factor in this study. Behravan, Hautamäki and Kinnunen (2015) look into the effect of language proficiency on accent recognition in more detail. They found that higher language proficiency does indeed tend to lead to a lower likelihood of an unknown speaker being correctly classified by native language. Presumably, a higher level of proficiency removes features that are indicative of a speaker's native language. Chapter 6 of this thesis further considers classifying non-native accents with regards to another type of automatic accent recognition system.

Bahari, Saeidi, van Hamme and van Leeuwen (2013) also compared different types of system, including i-vector-based systems, on an accent recognition task. The data they used were from the National Institute of Standards and Technology (NIST) Speaker Recognition Evaluations (SRE). Again, using the same terminology as Behravan, Hautamäki and Kinnunen (2013), this was on a foreign-accent recognition task, where speakers of five different language backgrounds were recorded speaking English. When coupled with a Support Vector Machine classification mechanism (discussed further below), they found that an i-vector-based system generates a higher classification rate (56% correct) than when using Gaussian Mean Supervector as the modelling technique, which generated a classifcation rate of 53% correct on this particular task.

When coupled with other types of classification mechanisms, however, it is worth noting that this hierarchy of performance is not necessarily the case on their particular accent classification task. It does not appear that we can necessarily expect that the state-of-the-art i-vector modelling approach will outperform GMM-based systems when other kinds of classification techniques are implemented.

Other work has focussed on native accent recognition using i-vector-based systems. DeMarco and Cox (2012) tested different i-vector systems on the ABI corpus (described in Section 2.1 above). On the 14-way accent recognition task between different accents across the British Isles, their best i-vector system achieved an accuracy of 68%.

While a number of the i-vector-based accent recognition studies discussed so far have explored the performance of different variants of i-vector systems working alone, more recent research has investigated the accent recognition performance of i-vector systems fused with other system types. System fusion is when we can combine the output scores of a number of different systems working in parallel to obtain an overall score for a trial (Brümmer *et al*, 2007). Najafian, Safavi, Weber and Russell (2016) fuse an i-vector-based system with a phonotactic-fused system and test this overall fused system on the ABI corpus. This fused system outperforms either system type alone, with an overall accent recognition accuracy of 84.87% on the 14-way recognition task. This kind of fused approach to system building aims to take advantage of the different strengths of different systems. It is possible that some types of system make correct classifications in cases where others do not. The i-vector-based system developed for the purposes of the experiments on the AISEB corpus detailed below has been influenced by the one in Najafian *et al* (2016). They combine their i-vector approach with a Support Vector Machine classifier (Sup-

port Vector Machines are introduced further below). When it is not fused with other systems, their i-vector system is reported to achieve 76.76% on the 14-way ABI accent recognition task.

c) ACCDIST-based modelling

This thesis has a particular focus on an alternative accent modelling method to GMMs and i-vectors: ACCDIST-based modelling. ACCDIST (Accent Characterization by Comparison of Distances in the Inter-segment Similarity Table) was first introduced by Huckvale (2004). It has been found to outperform GMM-based models in certain accent recognition tasks (Hanani, Russell and Carey 2011, 2013). However, unlike most i-vector-based systems and GMM-based systems, with past ACCDIST-based systems (Huckvale, 2004, 2007; Hanani *et al*, 2011, 2013) come two fundamental practical limitations:

1. Past ACCDIST-based systems have required the spoken content of the reference data used to train the system to be the same as the spoken content (the exact same string of words) of the unknown speech sample. For most applications, including forensic ones, we cannot expect that the content of training data and unknown data will match. This therefore greatly limits the number of applications that past ACCDIST-based systems could be used for.

2. Related to point 1. above, ACCDIST is text-dependent, because it requires a transcription to accompany the unknown speech sample in order for it to be processed and classified. This therefore requires a more laborious preprocessing stage as part of the overall classification task, and many applications will not accommodate this. When we would like to attach an accent recognition system to a speech recognition system's front

end, for example, there of course will not initially be a transcription available for the purpose of accent recognition. This would obviously remove the objective of developing a speech recognition system. For forensic applications, on the other hand, there may be instances where a transcription is available for an analysis, particularly when accuracy and precision of the outcome takes priority, well above convenience.

Section 2.3 will describe how the specific ACCDIST-based system implemented and represented here, Y-ACCDIST, overcomes the first of these limitations in the development details. This is the key way in which Y-ACCDIST has progressed from past ACCDIST-based studies. However, the second limitation of requiring a transcription still remains.

3) Classification

Classification procedures take the modelled data and make the decision surrounding the category of the unknown data. Like feature extraction and data modelling, numerous ways of classifying data exist and it is possible to trial different combinations of these different techniques.

Likelihoods

One way to do this, which will be employed in some of the systems in Section 2.3, is to calculate the likelihood that the acoustic features extracted from an unknown speaker's sample belong to the same group/class represented by a GMM. The highest likelihood value generated between an unknown speaker's sample and a model results in a class label. Calculating likelihoods in this way indicates the degree of similarity between an unknown and a reference model.

Correlation

Another way to calculate degree of similarity (in relation to ACCDIST-based modelling), and one seen in one of the systems introduced in Section 2.3, is by calculating correlation. Again, higher values indicate higher degrees of similarity. Various distances metrics (e.g Euclidean distance) could operate in the same way. The fifth system described in the experiments below employs Pearson's r product-moment correlation for this purpose (as per Ferragne and Pellegrino 2007, 2010).

Support Vector Machines

A widely used classification technique found across many machine learning applications is the Support Vector Machine (SVM) (Vapnik, 1998). This is a classification mechanism incorporated into most of the systems described in Section 2.3.1. To broadly illustrate how it works, we can refer to plotting training speakers, in their processed and modelled form, in multi-dimensional space. An optimal hyperplane is calculated and formed between the different categories of speakers. The category label of an unknown speaker can then be determined by where the unknown speaker's model falls in relation to the hyperplane. Further details and an illustration are given in Section 2.3.1.

Table 2.1 below summarises the systems and results in accent recognition studies discussed so far.

Table 2.1: Summary table of past accent recognition systems, databases and results discussed so far.

	DATABASES					
METHOD	**Arabic dialects**	**Mandarin Chinese**	**CallFriend**	**Finnish foreign language corpus**	**NIST**	**ABI**
Phonotactic	6.0% EER					
GMM-UBM		15.5/11.7% EER	18.73% EER	24.13% EER		
Gaussian Mean Supervector					53 % correct	
i-vector-svm			15.06% EER	20.01% EER	56% correct	76.76% accuracy
i-vector						68% accuracy
i-vector-phonotactic fused						84.87% accuracy

2.3 Experiments

The experiments in this chapter compare a number of systems, using the different modelling and classification mechanisms discussed in Section 2.2 above. In particular, this chapter will reveal the performance of a variation on ACCDIST-based modelling, which has been termed Y-ACCDIST (the York ACCDIST-based automatic accent recognition system). So far, variant systems of Y-ACCDIST have only been tested on databases that other automatic accent recognition systems have not. This chapter compares the performance of six different accent recognition system architectures on the same accent database:

- GMM-UBM
- GMM-SVM
- i-vector-SVM
- Phonological-GMM-SVM
- Y-ACCDIST-Correlation
- Y-ACCDIST-SVM

These can be compared with accent recognition systems presented in previous studies. The text-independent GMM-UBM and GMM-SVM systems are close to those compared in Hanani, Russell and Carey (2013). The i-vector system developed here is based on that seen in Najafian, Safavi, Weber and Russell (2016), a study discussed above in Section 2.2.2. We will also draw comparisons between the Y-ACCDIST systems developed here and the ACCDIST-based systems developed in Hanani *et al* (2013) as the most recent and high-

performing alternative ACCDIST-based systems. To create some cohesion between the text-independent GMM-based systems and the text-dependent ACCDIST-based systems, a text-dependent Phonological-GMM-SVM system (similar to the one found in Wu, Duchateau, Martens and Compernolle (2010)) has also been included in these experiments.

The main purpose of this chapter is to bring all of these system types to the same classification task on geographically-proximal accents. It is expected that the nature of the data, and the degrees of similarity that exist among the accents, will affect the success rates. The experiments in this section will therefore involve training and testing the six different automatic accent recognition systems on the same corpus of geographically-proximal accents. While the text-independent GMM-based systems and ACCDIST-based systems in Hanani, Russell and Carey (2013) were tested on the ABI corpus of 14 distinct accents spoken across the British Isles, the text-dependent phonological GMM-SVM system was tested on a different corpus by Wu *et al* (2010). They claim that the corpus of five Flemish varieties they tested it on is a more challenging task. Even though Flanders is known to have very distinct dialects for such a small geographical space, when speakers speak the standard language, differences between speaker groups are less prominent. Their experiments used recordings of speakers speaking the standard language. This chapter will allow us to compare Wu *et al*'s system with other system types on the same set of geographically-proximal accents. The results these past studies presented will be reproduced in the results section to compare these with the results produced by the systems when testing them on the same set of geographically-proximal varieties. The corpus of geographically-proximal accents used in these experiments is the *Accent and Identity on the Scottish/English Border* (AISEB) corpus (Watt, Llamas and Johnson, 2014). This is described in further detail

in Section 2.3.2 below.

The following subsections give details of aspects of the experiments run in this chapter. First, each of the different automatic accent recognition systems is described in turn, then the corpus of geographically-proximal accents, the AISEB corpus, used to train and test the systems is outlined. The results produced by each of these systems and an evaluation of performance are then given.

2.3.1 Automatic Accent Recognition Systems

The discussion in Section 2.2 above divided accent recognition systems into two types: phonotactic systems and acoustic systems. The systems developed and implemented in these experiments can be separated in a different way into two categories of system. These categories are assigned according to the systems' text-dependency. The first three systems described below are *text-independent systems*, whereas the following three are *text-dependent systems*. Due to discrepancies in the definitions of these terms between studies, it is important to define what is meant by these system types in the context of this study. Text-independent systems are systems which do not require any kind of transcription or knowledge of the spoken content of the speech data being processed. Text-dependent systems, on the other hand, require a transcription to accompany the speech data to help to estimate the linguistic units making up the spoken content. These linguistic units will then play a part in the modelling phase of the accent recognition process. Often, when text-dependent systems are tested and presented, the spoken content is identical between training and test data. While for the experiments in this chapter this is largely the case, the main intention behind the text-dependent systems used

in this thesis is that they can be applied to problems where the spoken content is known, but not necessarily where the spoken content is exactly matched between training and test data. This is in line with what might be feasible in forensic analyses and so this possibility is investigated in later chapters in this thesis. In potential forensic applications, we might know the spoken content of an utterance we want to analyse, but it is highly unlikely that it will match the training data. Using these text-dependency terms to subdivide our systems, each system is described below.

Text-Independent Systems

This part describes the three text-independent systems trained and tested in this chapter (i.e. the systems that do not require transcriptions as input).

System 1: GMM-UBM

This system architecture is frequently used for speaker recognition tasks. However, in this work it has been modified (following specifcations laid out by Hanani, Russell and Carey (2013)) to be tested for a classification task.

First, a Universal Background Model (UBM) is trained on multi-accent multi-speaker speech data (with 64 mixture components) through the Expectation Maximisation algorithm (Bilmes, 1988). Section 2.3.2 will describe the dataset used for the experiments presented in this chapter. It will describe a subset of a corpus (120 speakers out of a total of 160) that is used for training and testing the different systems. 40 speakers were discarded based on factors to do with the quality of the reading passage recording (in anticipation of the text-dependent experiments). The speech data of these 40 speakers have therefore been used to train the UBM in the systems that require a UBM.

Speech data of the corpus were manually processed to divide the data into stretches of speech between natural pauses, in effect aiming to remove silences from the recording. MFCCs (defined and explained in Section 2.2.2) are extracted throughout the UBM training speakers' speech samples. In this study, the Hidden Markov Model Toolkit (HTK) (Young *et al* 2009) was used to extract MFCCs. These are composed of 12 coefficients, plus energy, and in addition, delta and delta delta coefficients were included in the overall vector, totalling to 39 elements. These were extracted from 25ms windows of speech at overlapping 10ms intervals.

Accent-specific multi-speaker speech data for each accent in the corpus are then introduced to the training process as *enrolment data*. For each set of enrolment data (one set will contain only data for a single accent), MFCCs are extracted and maximum a-posteriori (MAP) adaptation (Gauvain and Lee, 1994) is applied to adapt an accent-specific model: a representative accent-specific GMM. The GMMs were made up of 256 mixture components. To classify a test speaker, the likelihood of the test speaker's acoustic features belonging to each of the adapted models is calculated. The highest likelihood indicates class membership.

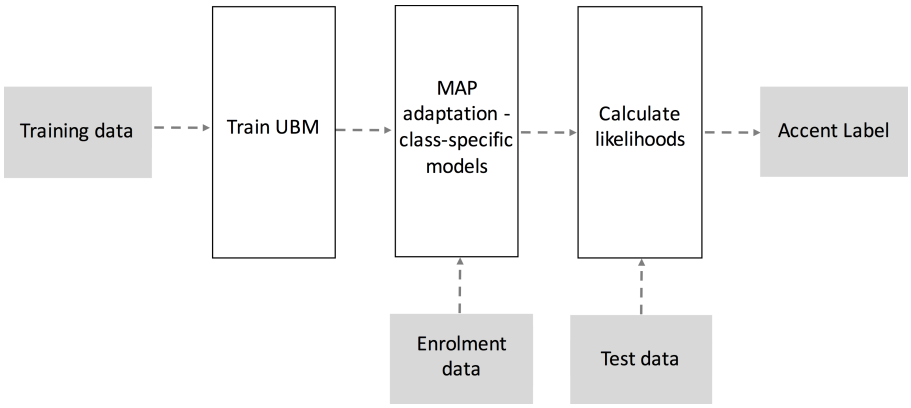

Figure 2.4: The processes involved in the GMM-UBM system.

System 2: GMM-SVM

In the same way as System 1, a UBM is trained using multi-accent, multi-speaker speech data. In the case of this system, the enrolment data are speaker-specific, but are still independent of spoken content. Instead of adapting one single model to represent one accent, a model is adapted for each of the speakers in the enrolment data (again, using MAP adaption). This leaves multiple adapted GMMs representing each accent, one per speaker. Taking each of these speaker-specific GMMs, the means are taken and concatenated to form a supervector to represent that speaker. These speaker vectors are then fed into a SVM classifier, which is effectively plotting these speakers in high-dimensional space. For each accent class we have in our corpus, we form a 'one against the rest' binary configuration. Each accent becomes the 'one', while all other speakers are collapsed into 'the rest'. The configuration rotates, in order to enable each accent to become the 'one'. Each time, the aim is to find a hyperplane that separates the accent class in question from 'the rest'. A

number of hyperplanes are likely to be possible, but a SVM aims to calculate the hyperplane with the largest (optimal) margin between the two groups[2]. Figure 2.5 below offers a simplified illustration of a Support Vector Machine classifier within 2-dimensional space.

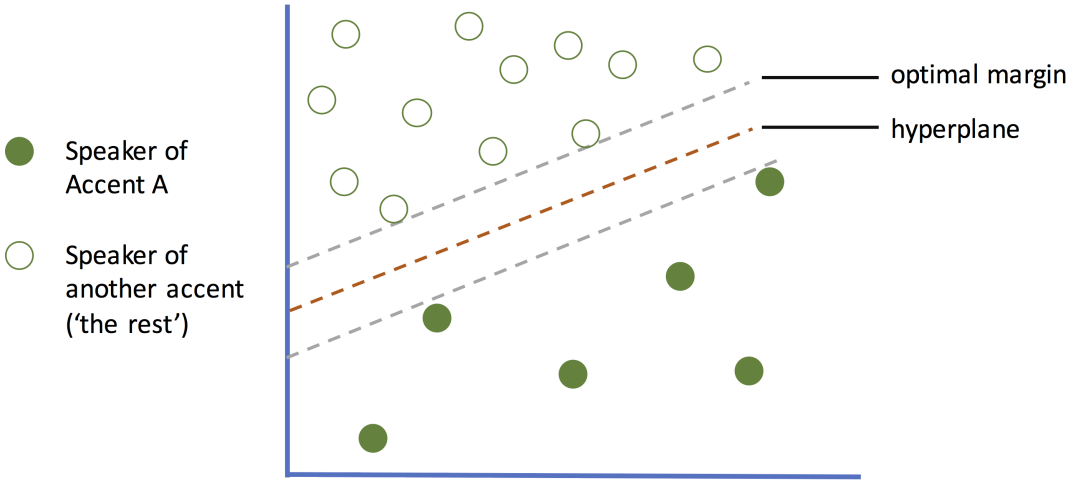

Figure 2.5: A simplified illustration of a Support Vector Machine classifier.

When classifying an unknown speaker, the speech sample is adapted from the UBM to form a GMM representing the speaker. The means of this model are fed into the SVM on each rotation. The sample is classified according to the clearest margin it forms relative to the hyperplane, indicating accent class membership.

[2]The SVMs implemented in the systems in this thesis have been implemented using the *scikit-learn* machine learning Python package. URL: scikit-learn.org.

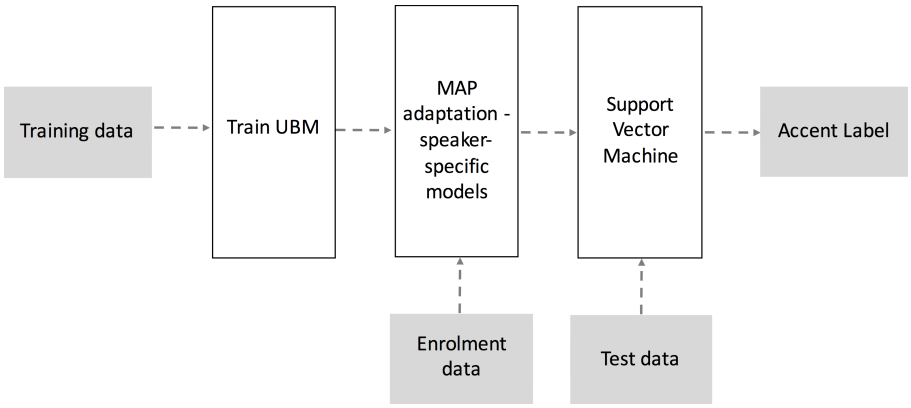

Figure 2.6: The processes involved in the GMM-SVM system.

System 3: i-vector-SVM

To extract i-vectors, the MSR identity toolbox (Sadjadi, Slaney and Heck, 2013) was used. Training an i-vector system begins in the same way as a GMM-UBM system and a GMM-SVM system. A UBM is first trained, again using the Expectation Maximisation algorithm. Using the enrolment training data for each accent, we can form a GMM supervector to represent each speech sample. From these supervectors, we can calculate i-vector models to represent each sample. We can look at this compression from supervector to i-vector through the equation,

$$\mathbf{M} = \mathbf{m} + \mathbf{T}\mathbf{w},$$

where $\mathbf{M}$ is the GMM supervector, $\mathbf{m}$ is the UBM supervector, $\mathbf{T}$ is the low-dimensional matrix and $\mathbf{w}$ is the i-vector. The aim is to capture the between-sample variability through applying a *total variability matrix*. An i-vector for a speech sample, $\mathbf{w}$, is calculated using Baum-Welch statistics and the

UBM. Behravan, Hautamäki and Kinnunen (2015: 120) summarise the i-vector process well when they describe it as "a mapping from high-dimensional GMM supervector space to a low-dimensional i-vector that preserves most of the variability". We can specify the number of dimensions that the total variability matrix has. In this work, we use 400, which is a standardly used number of dimensions in i-vector systems.

Once we have an i-vector representing each of our samples, our aim is to capture the between-accent variation, not the between-speaker variation which is what the i-vector models do up to this point. For the specific i-vector system used in this study, the mechanism used to classify speakers is the Support Vector Machine classifier. For this system, a polynomial kernel was used. This i-vector-SVM system configuration closely follows that seen in the accent classification experiments in Najafian, Safavi, Weber and Russell (2016).

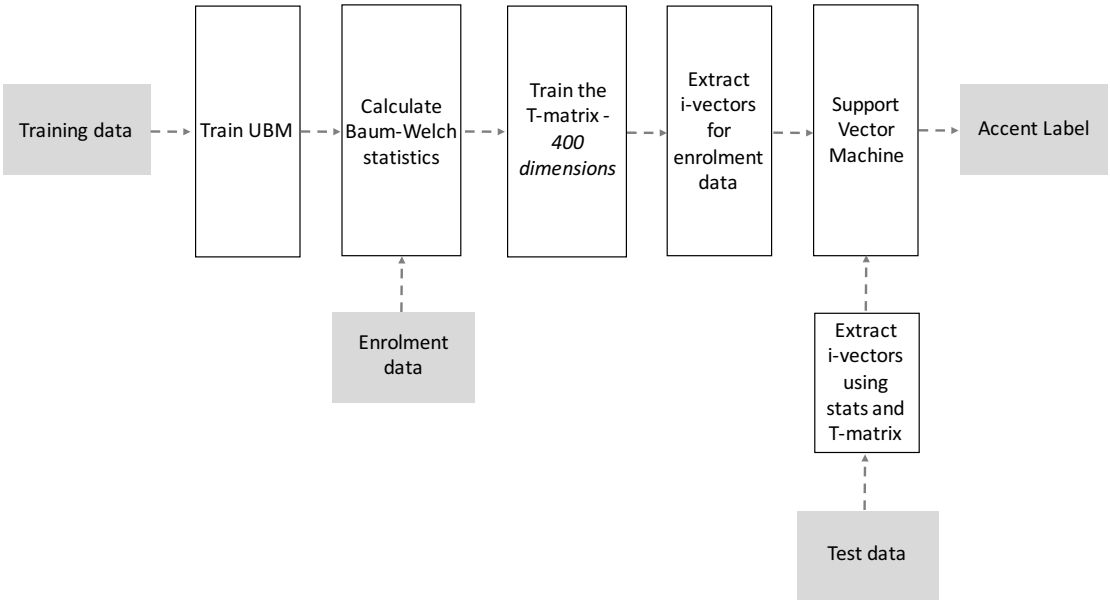

Figure 2.7: The processes involved in the i-vector-SVM system.

Text-Dependent Systems

The next three systems described are termed *text-dependent*. As noted above, these systems do not necessarily require the testing data to be composed of the exact same spoken content as the training data. However, the experiments in this chapter do use matching spoken content. Subsequent chapters make use of data where the spoken content is not directly comparable.

These text-dependent systems require a phoneset and pronunciation dictionary (or lexicon) to represent the phoneme segments that form the accent models. This is because each of these text-dependent systems relies on forced alignment as the first step in its processes. The forced aligner used was developed using the HTK 3.4 toolkit (Young *et al*, 2009). To align data for each speaker, the aligner was iteratively trained on the data itself (a more low-

resource solution to forced alignment). The specific phoneset that has been used for the experiments in this chapter is detailed further below in Section 2.3.3.

System 4: Phonological GMM-SVM

This system is based on that seen in Wu, Duchateau, Martens and Compernolle (2010). A number of speakers' speech samples, along with their transcriptions, for each of the accents involved, is taken and passed through a forced aligner. Using the output time alignments, a GMM is trained to represent each phoneme for an individual speaker. All the GMM means for each phoneme are concatenated to represent the speaker's pronunciation system in one long supervector. In the same way as the GMM-SVM system described above, each training speaker's representative vector is fed into a SVM classifier. To classify an unknown speaker, the speech sample and transcription are force aligned and subsequently used to train phoneme-specific GMMs. The means of these GMMs are concatenated into a supervector and introduced to the SVM to assign an accent label.

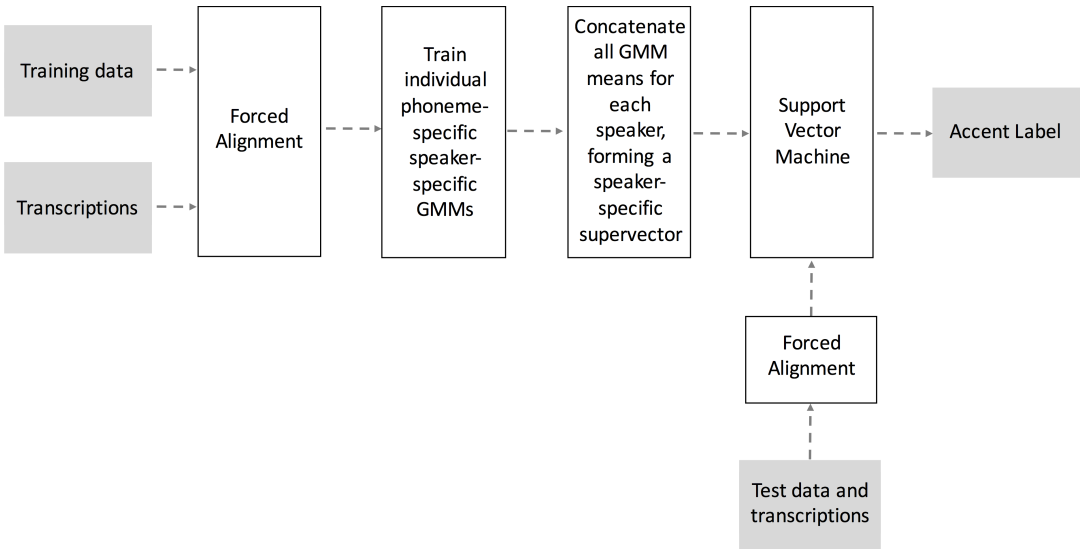

Figure 2.8: The processes involved in the Phonological-GMM-SVM system.

System 5: Y-ACCDIST-Correlation

For each speaker in the training set, a speech sample and a transcription are passed through a forced aligner. For each vowel phone token, the midpoint 12-element MFCC vector is extracted. These MFCCs are clustered into their respective phoneme categories and an average MFCC vector is calculated to represent each vowel phoneme[3]. As a result, each speaker's vowel phoneme inventory is represented by a series of average MFCC vectors, one per phoneme. These representative MFCC vectors form the foundations of a matrix and the Euclidean distance is calculated between all possible phoneme-pair combinations. The resulting table of distances, the Y-ACCDIST matrix, represents

[3]For now, the experiments will only involve vowel segments to follow the work of Huckvale (2004, 2007) and Hanani, Russell and Carey (2013). However, this thesis will move towards modelling consonants as well in subsequent chapters.

the speaker's pronunciation system. To explain how this provides a model of accent, we can refer to a specific example from British English. The vowels in FOOT and STRUT for Northern English English speakers are typically realised the same (both as [ʊ]). For Southern English English speakers, they are typically realised differently (FOOT contains [ʊ], whereas STRUT contains [ʌ]). When the Euclidean distance between the MFCCs for these two vowels is calculated, it is expected that a smaller value will be generated for a Northern speaker, and a larger one for the Southern speaker. A whole table (matrix) of these phoneme-pair distances should capture a collection of these accent-specific features. A Y-ACCDIST matrix is illustrated in Figure 2.9.

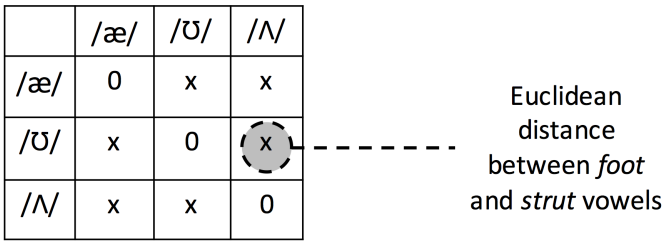

Figure 2.9: An illustration of part of a Y-ACCDIST matrix.

All the Y-ACCDIST matrices of the training speakers are pooled together according to accent category and an average Y-ACCDIST matrix is calculated to represent each accent. These average Y-ACCDIST matrices form our reference system.

To classify an unknown speaker, the speech sample and transcription are passed through a forced aligner and converted into a Y-ACCDIST matrix in the same way that is described immediately above. The Pearson r product-moment correlation is then calculated between the unknown speaker's matrix

and each of the average reference Y-ACCDIST matrices (one average matrix per accent). Using correlation here is an indicator of similarity between matrices. The higher the correlation, the more similar two matrices are. The unknown speaker's matrix is therefore assigned the same accent label as the reference matrix with which it generates the highest correlation.

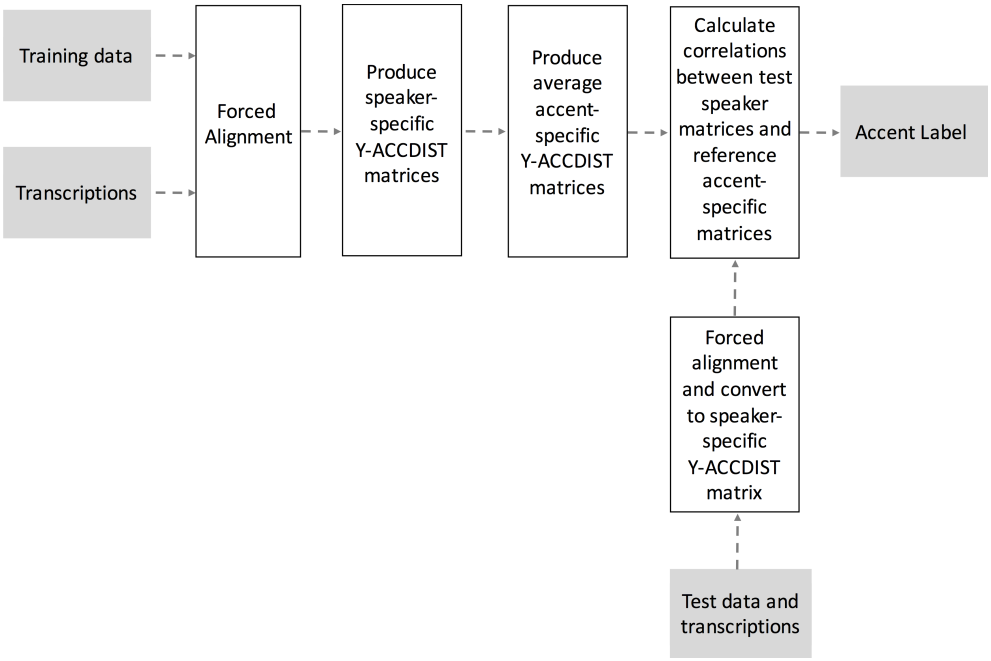

Figure 2.10: The processes involved in the Y-ACCDIST-Correlation system.

System 6: Y-ACCDIST-SVM

All speakers' speech samples in the training set are converted into speaker-specific Y-ACCDIST matrices in the same way described in the subsection immediately above for the Y-ACCDIST-Correlation system. The difference between the Y-ACCDIST-Correlation system and the Y-ACCDIST-SVM sys-

tem lies in the classification process. Rather than producing an average matrix for each accent category, the Y-ACCDIST matrices belonging to each accent category remain speaker-specific. These speaker-specific matrices are then used as input feature vectors for a SVM and the same classification process takes place as that explained for the other SVM systems described above.

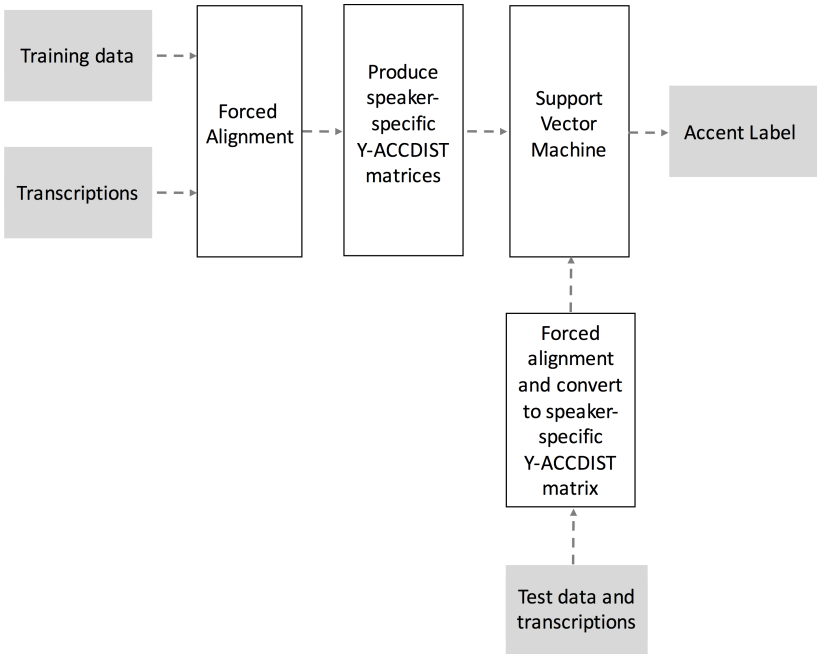

Figure 2.11: The processes involved in the Y-ACCDIST-SVM system.

The experiments presented here compare the six systems described above on the same accent corpus. First, details and some background information will be given about the accent corpus used in this chapter (the Accent and Identity on the Scottish-English Border (AISEB) corpus (Watt, Llamas and Johnson, 2014)).

2.3.2 The AISEB Corpus

The data used for the experiments run in this study are from the *Accent and Identity on the Scottish/English Border* (AISEB) corpus (Watt, Llamas and Johnson, 2014). This corpus contains speakers from four locations across the Scottish-English Border: Berwick-upon-Tweed, Eyemouth, Carlisle and Gretna. The AISEB corpus was not primarily collected for speech technology or forensic research, but for sociolinguistic purposes. These four communities provide an interesting case study for sociolinguistic research in that they sit in pairs very closely to one another (approximately 10 miles) either side of a political border between Scotland and England. In total the corpus contains speech samples from 160 speakers (40 per geographical location). Not only were recordings from speakers in these four towns collected to analyse the phonetic variation among them, but also to research the perceptual categorisations that the AISEB speakers make themselves about the speech of these communities (Llamas, Watt and MacFarlane, 2016). To achieve this, the corpus contains production data, attitudinal data and a perceptual strand, in the form of several experiments.

Sociolinguistic analysis of the AISEB corpus is ongoing, but there are some points about the linguistic features of these varieties presented in the literature so far. The accent varieties in the AISEB corpus are of sociolinguistic interest, particularly as there is a dimension of how national identity can interact with the linguistic features of speakers in the communities that sit along a political border (Llamas, 2010). In this specific instance, it is of interest to discover which linguistic features, if any, that are typically associated with Scottish English or English English are produced by speakers in the border communities. Llamas (2010) considers this in relation to coda-/r/, which is of course

a typical feature of Scottish English, a rhotic variety. In an empirical investigation of coda-/r/ production among speakers of the four AISEB varieties, Llamas shows a distinct difference between the localities either side of the border, where the coda-/r/ is used much more by the speakers in the two Scottish communities compared with speakers from the two English communities.

Llamas, Watt and Johnson (2009) also make predictions about the variety spoken in Berwick-upon-Tweed, based on their observation that it is "clearly a hybrid of Scottish and Northumbrian varieties". In doing so, they predict that Berwick speakers produce the second vowel in LETTER (when referring to Wells' (1982) keywords) in a similar way to how Tyneside speakers are known to produce it. A more open vowel [ɐ] is more likely to be produced than [ə].

Independent of work specifically on the AISEB corpus, there has been some account of the spoken variety in Carlisle. Jansen (2013) acknowledges that the Carlisle variety is often confused for Newcastle English or even Scottish English. She provides an overview of Carlisle English, drawing on features it shares and does not share with Newcastle English. Within her account of Carlisle English, Jansen notes that the DRESS and KIT vowels are raised, while the TRAP vowel is low ([a], rather than [æ]). Among younger Carlisle speakers, the possible realisations of the vowels in FACE and GOAT are reported to be the same as those produced by Newcastle speakers. However, older Carlisle speakers have an alternative possible realisation of [eə] for FACE.

A subset of AISEB was used for the experiments in this chapter, taking 30 speakers from each location and using the recorded reading passage as a speech sample for each. A subset was used, rather than the whole corpus, because of the quality of some of the reading passage recordings. For various reasons, some recordings were judged unsuitable for these experiments. In addition, we wanted a balanced number of speakers to represent each accent group. Within

each group, we have a range of ages (15 speakers in the 'younger' category (approx. 14-27 years) and 15 in the 'older' category (approx. 54-93 years). We attempted to maintain a balance of male and female speakers per accent group. This division was not perfectly even, but it was approximately equal within the groups. Like the experiments by Hanani, Russell and Carey (2013), the spoken content for each speaker is the same. The recordings are of good quality - a sampling rate of 44.1kHz was used - and last approximately three minutes per speaker.

2.3.3 Phoneset

The phoneset defines the speech segment categories to which we will assign our individual phone tokens in the speech data (which is of relevance to the text-dependent systems). In linguistic terms, we can think about it as a kind of phoneme inventory we choose to apply to the data. We know that a number of different phoneme inventories can be relevant to a single language, dependent on the accent varieties in that language. Even though we are working with Scottish and English varieties in this chapter, we will be using a North American English phoneset, based on the *VoxForge* phoneset[4]. In the early development stages of these text-dependent systems, this North American English phoneset was implemented and compared with when a British English phoneset was implemented[5]. The North American phoneset and pronunciation dictionary were discovered to yield the best results for these data. A key difference between the North American pronunciation dictionary and the British English pronunciation dictionary is that the North American one accounts for rhoticity in speech. This is likely to be a key distinguishing feature between

[4]The VoxForge website can be found here: http://www.voxforge.org/home

[5]Neither phoneset includes the full repertoire of phonemes for the varieties in question.

the Scottish and English speakers in the AISEB corpus. Experiments later in this thesis will use slightly different pronunciation dictionaries, which are determined by the type of data that are being used. For reference, the phoneset that we have used for these AISEB experiments is presented in Tables 2.2 and 2.3 below:

Table 2.2: The vowel phoneset symbols used for the AISEB experiments alongside their corresponding IPA symbols.

Phoneset Vowel Symbol	AE	AA	AO	IY	UW	EH	IH	UH
IPA Symbol	æ	ɑ	ɔ	i	u	ɛ	ɪ	ʊ
Phoneset Vowel Symbol	AX	EY	AY	OY	OW	AW	ER	AH
IPA Symbol	ə	eɪ	aɪ	ɔɪ	əʊ	aʊ	ɜ	ʌ

Table 2.3: The consonant phoneset symbols used for the AISEB experiments alongside their corresponding IPA symbols.

Phoneset Consonant Symbol	P	T	K	B	D	G	CH	JH
IPA Symbol	p	t	k	b	d	g	tʃ	dʒ
Phoneset Consonant Symbol	F	V	S	Z	TH	DH	SH	HH
IPA Symbol	f	v	s	z	θ	ð	ʃ	h
Phoneset Consonant Symbol	zh	L	R	W	Y	M	N	NG
IPA Symbol	ʒ	l	ɹ	w	j	m	n	ŋ

2.3.4 Results

This section compares results obtained by each type of system from past studies with those obtained by each system tested on the AISEB corpus. First, the results of the previous studies are described and presented, and then the results produced by similar systems on the AISEB corpus are presented.

Previous Studies

As repeated above, we can loosely compare the results generated by the systems developed and tested here with those of similar architectures tested on corpora in previous studies. Those which were conducted by Hanani, Russell and Carey (2013) and Najafian, Safavi, Weber and Russell (2016) were 14-way accent classification tasks using the ABI corpus, whereas the result generated by Wu, Duchateau, Martens and Compernolle (2010) was a five-way classification task on Flemish varieties. For reference, these past results are presented in Table 2.4.

Table 2.4: Recognition rates for six accent recognition systems from past studies for reference.

System	Corpus	No. classes	% Accuracy
GMM-UBM	ABI	14	61.13
GMM-SVM	ABI	14	76.11
i-vector-SVM	ABI	14	76.76
Phon-GMM-SVM	Flemish	5	63.2
ACCDIST-based Correlation	ABI	14	93.17
ACCDIST-based SVM	ABI	14	95.18

While these results are not the main focus of this chapter, they provide a good reference point to then discuss the results presented below. What is important to note is the overall hierarchy of performance of these different systems. This will be discussed further in relation to the results given in the next subsection.

Results generated using the AISEB corpus

A three-fold cross-validation experimental setup was put in place to generate results using the AISEB corpus. The total pool of 120 speakers was split into three groups, 40 in each (10 speakers per accent). Each system was then trained on two of these groups (80 speakers), and tested on the remaining 40 (the UBM remained unchanged, and it was only the enrolment data that changed in the systems that made use of a UBM). The groups of speakers would rotate round for each group of 40 to become the test batch, and eventually all 120 speakers would be tested.

The overall recognition rates for each of the six systems are shown in Table 2.5 below.

Table 2.5: Recognition rates for six accent recognition systems when tested on the AISEB corpus.

System	**Recognition rate (%)**
GMM-UBM	37.5
GMM-SVM	47.5
i-vector-SVM	40.8
Phon-GMM-SVM	68.3
Y-ACCDIST-Correlation	76.7
Y-ACCDIST-SVM	86.7

We can observe a great spread of results across the six systems, as we would expect of a combination of text-independent and text-dependent systems. We can draw some conclusions when comparing the above results with those presented in past studies (reproduced in Table 2.4). Hanani, Russell and Carey (2013) and Najafian, Safavi, Weber and Russell (2016) report results from a 14-way accent recognition task between the relatively more geographically-distant accent groups found in the Accents of the British Isles (ABI) corpus. In the context of a 14-way accent recognition task, we would predict a chance recognition rate of 7.14%. All results they report show performances well beyond chance expectations. When we compare the performance of very similar systems on the AISEB corpus, we can see how much more robust ACCDIST-based systems are to discriminating geographically-proximal accents. In contrast, the lowest-performing result (produced by the GMM-UBM system) only sits marginally above the chance level of 25% for this task. Although a drop in the performance of text-independent systems is perhaps expected, we can observe the extent to which the text-independent systems suffer when they are faced with a more challenging task of distinguishing between accents with greater degrees of mutual similarity. To lend greater support to this observation, it would of course be useful to test the exact systems built in this study on the ABI corpus, rather than draw conclusions from a comparison between very similarly developed systems.

To take a closer look at the performances of each system, the confusion matrices for each individual system are presented below. The absolute number of correctly classified speakers is presented in each cell, indicating which accent label the test speakers were assigned.

Table 2.6: GMM-UBM confusion matrix (37.5%).

Accent	Ber.	Car.	Eye.	Gre.
Ber.	**12**	12	5	1
Car.	6	**13**	8	3
Eye.	3	6	**19**	2
Gre.	13	11	5	**1**

Table 2.7: GMM-SVM confusion matrix (47.5%).

Accent	Ber.	Car.	Eye.	Gre.
Ber.	**14**	3	3	10
Car.	6	**12**	4	8
Eye.	4	1	**16**	9
Gre.	7	5	3	**15**

Table 2.8: i-vector-SVM confusion matrix (40.8%).

Accent	Ber.	Car.	Eye.	Gre.
Ber.	**18**	7	1	4
Car.	11	**8**	1	10
Eye.	9	5	**12**	4
Gre.	11	6	2	**11**

Table 2.9: Phonological GMM-SVM confusion matrix (68.3%).

Accent	Ber.	Car.	Eye.	Gre.
Ber.	**18**	3	3	6
Car.	1	**28**	0	1
Eye.	5	0	**23**	2
Gre.	4	7	6	**13**

Table 2.10: Y-ACCDIST Correlation confusion matrix (76.7%).

Accent	Ber.	Car.	Eye.	Gre.
Ber.	**28**	2	0	0
Car.	2	**17**	2	9
Eye.	0	0	**26**	4
Gre.	0	6	3	**21**

Table 2.11: Y-ACCDIST SVM confusion matrix (86.7%).

Accent	Ber.	Car.	Eye.	Gre.
Ber.	**30**	0	0	0
Car.	2	**23**	1	4
Eye.	0	0	**28**	2
Gre.	0	5	2	**23**

The confusion matrices offer the opportunity to observe whether the systems struggle to classify speakers of some accents over others. Some such observations are discussed in Section 2.4 below.

2.4 Discussion

In line with the results of past studies (Table 2.3), the results generated using the AISEB corpus (in Table 2.4) are largely consistent with the expected hierarchy in performance. The Y-ACCDIST-SVM system, followed by the Y-ACCDIST-Corrleation system, achieve the highest recognition rates, while the GMM-UBM system achieves the lowest. It was mentioned above that even though the Y-ACCDIST systems, as well as the three text-independent systems, display a drop in performance relative to previous studies, it seems that the drop in performance is less substantial for the Y-ACCDIST systems, particularly the Y-ACCDIST-SVM system. Based on this observation and comparison with other studies, we can consider the Y-ACCDIST-SVM as a state-of-the-art text-dependent accent recognition system.

One thing to bear in mind is that the performances of all the systems are likely to improve with an increased number of training speakers, particularly in the case of the text-independent systems. Because of the way Y-ACCDIST works by calculating intra-speaker segmental distances, voice quality characteristics, like those which correlate with the speakers' age and sex, are expected to play a much smaller role in the model than in i-vector or GMM-based models, and instead, the emphasis is on the speakers' vowel systems. Particularly as the AISEB dataset provides a range of speaker ages and speakers of both sexes, Y-ACCDIST is probably at an advantage. In the other types of system, we can expect that these voice-quality characteristics are more prevalent in

the models, and so the accent variation (the specific aspect we are interested in for this task) is overshadowed on more occasions in classification. Taking a segmental approach in the first instance of course increases the Y-ACCDIST systems' chances of achieving greater recognition rates. Increasing the training dataset is likely to allow the accent variation to be modelled more effectively by the text-independent systems. Alternatively, it might be that these particular accent varieties are not suitable for classification by these kinds of systems. A larger dataset would be required to establish answers to these kinds of questions.

In a sense, the facts that the text-dependent systems can cope with the corpus size we are using, and the text-independent systems do not, are worth knowing. For forensic applications, it is unlikely that we will have access to vast amounts of data to train a single system on. It could therefore be worth further investigating more 'low-resource' approaches to accent recognition and trying to determine how many training speakers are sufficient for a reliable analysis to take place. It might be the case that the text-dependent approaches can offer us this kind of low-resource capability.

One prominent difference that is obvious between the results of the past studies, and the results from the AISEB experiments is the performance of the GMM-SVM and the i-vector-SVM systems. Using the ABI corpus, Hanani, Russell and Carey (2013) and Najafian, Safavi, Weber and Russell (2016) present very similar performances of 76.11% and 76.76%, respectively. However, when both are tested on the AISEB corpus, more of a performance gap emerges between the two types of system, with the i-vector-SVM system achieving the lower score of 40.8% and the GMM-SVM system achieving 47.5%. Because both systems use an SVM classifier, it seems that the i-vector modelling procedure is not the best for this specific task on this dataset. Again,

it could be down to the quantity of data, whereby i-vector-based systems can create more effective models with greater quantities of data. Alternatively, it might be that i-vector models are not as effective when discriminating varieties distinguished by more subtle accent differences. Applying these systems to a greater number of corpora of varying sizes and degrees of similarity would allow us to get a grasp of how these different systems perform on different types of dataset.

Turning our attention to the individual confusion matrices for each of our systems, we can see that some patterning is in evidence. With the number of speakers that have been used, it is important to note that these are fairly speculative observations which would require further investigation. It is reasonable to expect that some accents are naturally more distinct than others. Some will have a greater number of differences, or larger degrees of difference, from the rest of the varieties in the dataset. We can look for these kinds of differences among the accents in the confusion matrices for the systems. Broadly speaking, Gretna tends to be the lowest-performing group overall (with a total of 84 correct classifications out of a possible 180 across all six systems). From a sociolinguistic point of view, this is unsurprising. Historically, Gretna is the newest of the four towns in the AISEB corpus and was formed very suddenly in the First World War as workers of a new munitions factory came to settle in the area (Watt, Llamas and Johnson, 2014). It is therefore plausible that the variety spoken in Gretna is not as established or distinct as the other varieties because the town has not been long established itself. It might therefore be understandable that the Gretna accent is more confusable than other varieties, which is what we can generally observe among the confusion matrices.

At the other end of the spectrum, Eyemouth speakers are correctly classified on the most occasions, with a total of 124 correct classifications out of a

possible 180 across all six systems. Eyemouth is consistently in the top two most correctly classified varieties across all the systems' results. Simply due to the nature of accents, we can assume that some accents are more likely to be correctly classified than others, and accent recognition performance is not solely down to the systems.

Related to this, it could be that some varieties are more suited to being classified by a particular type of system. Carlisle, as a variety, seems to be inconsistent across systems in terms of its individual success rate relative to the other accents. When processed by the i-vector-SVM system it is the lowest-performing accent variety, whereas the Phonological-GMM-SVM system correctly classifies Carlisle speakers on more occasions than the other varieties. Interestingly, though, when it comes to both Y-ACCDIST-based systems (the highest-performing systems), the Carlisle variety performs relatively poorly. It could be the case that the Carlisle accent is not as well suited to Y-ACCDIST modelling as the other varieties, and the distinguishing factors in a Carlisle accent are better presented through other modelling procedures. This is where system *fusion* may be able to play a part in achieving an overall improved accent recognition performance (as seen in Hanani *et al*, (2013) and Najafian *et al*, (2016)). Fusion can move us towards combining the strengths of different types of system to produce a single improved result. Future research could move in this direction.

When we are considering applying automatic accent recognition technologies to forensic tasks, we should expect challenging data types. The experiments presented in this chapter were conducted on content-controlled reading passage data, which is of course not a data type we can normally expect in forensic tasks. The experiments presented in the next chapters in this thesis will therefore involve spontaneous speech data so as to move closer towards

the kinds of data found in forensic casework, and to challenge an automatic accent recognition system in this respect.

2.5 Summary

This chapter has tested and evaluated six different automatic accent recognition systems on a corpus of geographically-proximal accents, the AISEB corpus. With reference to similar systems developed in past studies, we have compared the performances of different systems in relation to the specific dataset chosen for this purpose. Performance was evaluated in terms of the dataset size and the specific accents that were involved in the experiments. A number of possible further research directions exist in the form of running similar experiments on a whole range of corpora, as well as combining the different systems in order to take advantage of the strengths of different system types.

CHAPTER 3

Automatic Accent Recognition on Spontaneous and Degraded Speech Data

3.1 Introduction

For controlled accent recognition experiments, using spoken content which is the same across speakers (i.e. producing the same reading passage or prompts) has its advantages. Controlling this condition can allow us to more directly compare recognition performance across different accent varieties and individual speaker performance. It also makes it possible to make direct comparisons across individual segments found in the same phonological environments, if we were interested in how particular speech segments contribute to a task. However, when considering these systems for real-life forensic applications, matching controlled spoken content across speakers is not a data feature we can rely on. To make an accent recognition tool as versatile as possible, it is essential that the tool can make accent recognition decisions based on what will be termed *content-mismatched* speech data (i.e. spontaneous speech sam-

ples), as well as *content-controlled* speech data (i.e. speakers producing the same reading passage or prompts). In the phonetics literature, differences have been studied between the reading mode and the spontaneous mode of speech. For example, Howell and Kadi-Hanifi (1991) ran a small study in which they asked a number of speakers to describe a room (spontaneously) while being recorded. This recording was orthographically transcribed, and the same speakers were asked to read the transcription of their own recording. The reading was also recorded, resulting in two recordings of the same spoken content, by the same speaker, one spontaneously produced and the other in reading mode. Among their observations, they found differences such as speakers pausing in different places and stress being placed differently. They also comment that the rate of speech is higher in spontaneous speech than it is in read speech. Consequently, not only does working with spontaneous spoken data introduce content mismatch, but it also presents a different quality of speech, and this quality might affect the performance of an automatic system. Hanani, Russell and Carey (2013), a study regularly referenced throughout this thesis, ran their automatic accent recognition experiments on read speech data, like in the experiments presented in Chapter 2 above. However, Hanani *et al.* speculate in their discussions of the experiments that spontaneous speech might actually increase the prominence of a speaker's accent features, in turn increasing the chances of a successful accent classification.

In the context of automatic accent recognition systems, it is not clear whether the spontaneous spoken mode might be advantageous to overall classification performance, or detrimental to it. It might be the case, particularly in the context of text-independent systems, that spontaneous speech increases the chances of an unknown speaker being correctly classified. Because the text-independent systems we investigated in Chapter 2 (GMM-based systems

and i-vector-based systems) capture an overall acoustic representation of the speech samples, rather than making phoneme-on-phoneme comparisons, they rely on having enough accent-specific features among speakers in a certain accent category. If Hanani *et al*'s (2013) hypothesis is correct, then text-independent systems might benefit from spontaneous speech data.

In contrast, by taking a more segmental approach, as we do with Y-ACCDIST, we might expect recognition rates to suffer when classifying content-mismatched speech data, in comparison to a Y-ACCDIST system's performance using controlled read speech. This is for a number of reasons. First, the content-mismatched nature of data is likely to bring problems to a unit-dependent approach like this one. This is because of the varying co-articulatory effects on a single phone. In the case of Y-ACCDIST, which makes average representations of phonemes, the variation in the contexts that the individual phones are found will affect the stability and comparability of these phoneme representations across speakers. Another factor which might contribute to Y-ACCDIST's performance when processing spontaneous speech is the impact the spontaneous data has on the forced alignment quality. It is reasonable to expect that spontaneous speech would increase the segmentation error of the aligner relative to read speech. The expected increase in articulation rate and connected speech processes that come with spontaneous speech may reduce the quality of the alignment (Goldman, 2011). If this is the case, it will also contribute to less stable phoneme representations in the modelling stage of Y-ACCDIST, and consequently, more 'noise' in the overall recognition process.

As discussed, there are reasons why spontaneous speech data might challenge a Y-ACCDIST-based system. The key purpose of this chapter is therefore to show Y-ACCDIST's accent recognition performance on content-mismatched speech data and to outline key considerations when processing such variable

data.

The main way in which Y-ACCDIST deviates from the other ACCDIST-based systems reported in previous studies (Huckvale 2004, 2007; Hanani, Russell and Carey 2011, 2013) is that it is designed to be able to process content-mismatched (spontaneous) speech. Past ACCDIST-based systems have required the spoken content of the training and testing data to match. Chapter 2 contained the system descriptions of two Y-ACCDIST-based systems. The key difference between Y-ACCDIST and past ACCDIST-based systems is the modelling stage of these two systems, and it is this difference which makes Y-ACCDIST applicable to content-mismatched speech data. All ACCDIST-based systems are reliant on effectively making comparisons between the same segmental units across speakers. The default version of Y-ACCDIST, for example, will compare each representation of a test speaker's vowel phonemes with the vowel phoneme representations in the trained model. The ACCDIST systems in Huckvale (2004, 2007) rely on word-specific vowels, so the vowel in *cat* and the vowel in *trap* could not be compared against one another. They are treated as separate vowels. The vowel in *cat* can only be compared with other instances of the vowel in *cat*. The ACCDIST-based systems in Hanani *et al* (2011, 2013) are similarly context-specific in the comparisons they are restricted to. These systems make triphone-specific vowel comparisons. Like the system in Huckvale (2004, 2007), Hanani *et al*'s systems cannot make comparisons across *cat* and *trap*. However, unlike the system by Huckvale, Hanani *et al*'s systems could compare the vowel in *cat* with the first vowel in *cattle*. The Y-ACCDIST systems collapse all of these vowels into one, and can therefore make comparisons across them, making it possible to work across content-mismatched speech data. As already speculated, making these segmental collapses is likely to form less stable accent models, but the purpose of

this chapter is to determine how an ACCDIST-based system performs when processing content-mismatched data, rather than only processing highly controlled spoken content.

In addition, this chapter will also present results on degraded data. Again, when aiming to make further steps towards testing and evaluating an automatic accent recognition system on forensically relevant speech data, it is important to test a system on data of a quality lower than what we have so far been testing it on. A large proportion of a forensic speech scientist's work involves analysing telephone call recordings (French, Harrison, Kirchhübel, Rhodes and Wormald, 2017). We can simulate telephonic data by artificially degrading these recordings. In doing so, we can make a direct comparison between recording qualities using exactly the same recordings, providing a means by which to conduct a controlled experiment. It is important to keep in mind, however, that artificial degradation does not simulate genuine telephone recordings with complete accuracy. Byrne and Foulkes (2004) demonstrated this via a study in which they collected recordings of a small number of speakers producing a short reading passage over a mobile phone. While this mobile phone recording was being made, another microphone positioned in front of the speaker was recording the same event, resulting in the same speech signal being captured by a mobile phone and a microphone. Having analysed a number of vowel formants, they showed a clear increase in Formant 1 (F1) values in the mobile phone recording relative to the studio microphone recording of the same speech production. Artificially degrading a speech sample will not necessarily accurately replicate these kinds of effects, and so this should be kept in mind when analysing the results from the degraded data. Chapters 6, 7 and 8 of this thesis will reveal system performance using genuine telephone recordings, but these cannot be directly compared with good-quality

recordings in the way shown in this chapter.

3.1.1 Outline

To try to satisfy both of the present chapter's key aims, which revolve around testing an automatic accent recognition on more forensically realistic data, this chapter takes the following two steps:

1. Test the highest-performing system observed in Chapter 2, the Y-ACCDIST-SVM system, on spontaneous speech data.

2. Test the Y-ACCDIST-SVM system on the same data, having artificially degraded it.

These two objectives are addressed in the experiments presented in Section 3.2.

In connection with these research aims, this chapter will also take a closer look at the speaker models (the individual speaker Y-ACCDIST matrices) that are produced using the content-mismatched dataset. In Section 3.3, we will explore the degree of similarity that exists between the speaker models in the dataset of different accents to see how the models sit in relation to each other. We would expect to see speakers fall closest to other speakers in the same accent class. We will do this with both good-quality data and artificially degraded data to observe the effects of degradation on the individual speaker models. This is done in an effort to better understand how Y-ACCDIST models content-mismatched data in their high-quality and degraded forms.

To pursue the research objectives, the experiments in this chapter use speech data from a different accent corpus from the one used in Chapter 2. More details about this new corpus will be given in Section 3.2.1 below. The

change in corpus has been motivated by the large amount of transcribed conversational data per speaker, allowing for a more detailed inspection of the effects of this kind of data on accent recognition performance.

3.2 Experiments

This section outlines the experiments carried out so as to test Y-ACCDIST on spontaneous speech. First, details about the data will be given in Section 3.2.1 and then the procedure and results with some analysis will follow. Details about the Y-ACCDIST-SVM system being used for these experiments were given above in Section 2.3.1 (System 6).

3.2.1 The Data

To thoroughly explore the condition of spontaneous speech in automatic accent recognition, a different corpus has been selected based on the quantity of orthographically transcribed speech data per speaker it contains. While the AISEB corpus used for the experiments in Chapter 2 does contain a spontaneous speech component, it does not have enough transcribed conversational speech per speaker to sufficiently address the research objectives set out in this chapter[1]. A subset of the *Language Change in Northern Englishes* corpus (Haddican, Foulkes, Hughes and Richards, 2013) (henceforth, the *Northern Englishes corpus*) has been processed for these experiments. This corpus contains recordings of British English speakers from Derby, Manchester, Newcastle and York, and the part of the corpus that this study is interested in is the

[1]It would, however, be preferable to conduct a more direct comparison of content-controlled and content-mismatched data types on the same data.

conversational data produced in pairs between the speakers. For each pair of speakers, there is approximately an hour of orthographically transcribed conversation (on topics guided, but not necessarily enforced, by the researchers). These recordings are sampled at 44.1kHz, as good-quality recordings.

The Northern Englishes corpus was originally collected for sociolinguistic research purposes, like the AISEB corpus used for the experiments in Chapter 2, and not specifically for speech technology or forensic research. Each location contains speakers of a 'younger' age band (ranging between ages 16-27). There was a slightly different demographic focus in the Manchester group. Because of a particular sociolinguistic interest in the Manchester accent variety by Haddican *et al* (2013), the Manchester sample contains two additional age groups of speakers. These can be broadly labelled as 'middle-aged' and 'older'. The intention behind this stratification of speakers is to monitor and analyse changing variables in the Manchester spoken variety. Even though this additional categorisation of speakers is available in the corpus, the experiments and analysis in this chapter will mostly focus on accent groups from different geographical locations. This is because the number of speakers per age group is not expected to be sufficient to train and test the Y-ACCDIST-SVM system.

For most of the experiments reported in this chapter, then, only a subset of the Northern Englishes corpus has been used. Section 3.3, however, does make use of these additional groups of Manchester speakers. The number of speakers per accent group in the corpus is uneven. In order to achieve equal numbers of speakers per group, Derby speakers have been excluded from this study, as fewer than 15 speakers are available to use. Recordings of 15 speakers, all falling within the 'younger' age group (approximately aged 16-27), along with their orthographic transcriptions, from Manchester, Newcastle and York, have been pre-processed and processed by the Y-ACCDIST-SVM

system for the experiments presented in this section. Where possible, equal proportions of male and female speakers have been included per accent group. However, occasionally the speakers did not provide enough conversational data for the experiments in this thesis. There are therefore some slight imbalances in the proportions of male and female speakers in the accent groups. Particularly when using smaller datasets like this one, these kinds of imbalances in the speaker population are likely to affect accent recognition performance for many systems. Text-independent systems, like i-vector-based systems, are likely to capture speaker variation in multiple ways, including voice quality characteristics that correlate with speaker properties like speaker sex. These characteristics are likely to 'distract' from distinctive accent features. Y-ACCDIST-SVM aims only to capture and characterise accent-specific features of speakers, and so such imbalances are expected to be less of an issue for a text-dependent system like this. For each speaker, 10 minutes of net speech was obtained. To do this, the speech was manually preprocessed to only get turns of speech spoken by a single speaker, effectively separating out the pairs of speakers in the conversational data. The orthographic transcriptions were chunked and labelled along with their corresponding snippets of recordings. Once the speech data of the speakers were separated, and the appropriate orthographic transcriptions assigned to the relevant chunks of speech recording, the data were ready to pass through Y-ACCDIST-SVM.

Compared to the AISEB varieties that were in focus in Chapter 2's experiments, the accent varieties in this chapter are relatively well documented. There are a number of segmental features that might play a part in distinguishing them. Hughes, Trudgill and Watt (2012) list a number of linguistic features that are characteristic of Manchester speech. Of these, they note the production of [g] that follows /ŋ/ in words like *sing* and *tongue*. /l/ is typically

realised as 'dark'-/l/ in both onset and coda positions (rather than 'light'-/l/ in the onset position, which is typical of most varieties of English). Among the vowel features, Manchester speakers typically have a more open and further back realisation of the second vowel in LETTER (more like [ʌ] or [ɒ], rather than [ə]). Hughes, Trudgill and Watt also point out that the NURSE and GOOSE vowels are usually fronted, where the GOOSE vowel is even sometimes realised as a diphthong. This list presents us with a constellation of features that make up a typical Manchester accent.

Turning to what we might expect from the Newcastle data, Watt (2002) shows that what we might have previously considered as typical features of Tyneside English are likely to have changed. More traditional realisations of the FACE and GOAT vowels are the diphthongs /ɪə/ and /ʊə/ respectively. However, in Watt's study of vowel productions from Tyneside speakers found monophthongal realisations, more like /eː/ and /oː/. Watt and Allen (2003) provide an overview of phonetic features of Tyneside English. One notable distinguishing feature of Tyneside English is the glottalisation of stops /p, t, k/ in certain contexts. The example Watt and Allen give is the /t/ in the word *carter*. It might more appropriately be transcribed as [ʔt] in a typical Tyneside production as a sort of double articulation.

Compared to Manchester and Newcastle, the variety spoken in York is not as well documented. York English falls in line with other Northern varieties of English. In contrast to Newcastle English, Haddican, Foulkes, Hughes and Richards (2013) show the diphthongization of the FACE and GOAT vowels (the reverse of what was reported for Tyneside English by Watt (2002)). This could be a point of distinction in the experiments presented in this chapter. Similarly to Manchester, however, GOOSE is shown to become more front in York English.

The sociophonetic literature has provided us with a number of linguistic features we can perhaps depend on for distinguishing between the accent varieties in the Northern Englishes dataset.

3.2.2 Phoneset

Because we are working with different varieties of English, these experiments use a slightly different phoneset and pronunciation dictionary from the experiments run on the AISEB corpus in Chapter 2. The pronunciation dictionary used in Chapter 2 adopted North American features (such as rhoticity), which is not relevant to the Northern Englishes data. Instead, the experiments in the present chapter will take on a pronunciation dictionary and phoneset that is based on British English. This means that the phoneset we are using for these experiments is slightly larger than that used for the experiments presented in Chapter 2. For reference the phoneset used for the Northern Englishes experiments is presented in Tables 3.1 and 3.2:

Table 3.1: The vowel phoneset symbols used for the Northern Englishes experiments alongside their corresponding IPA symbols.

Phoneset Vowel Symbol	AE	AA	AO	AX	IY	UW
IPA Symbol	æ	ɑ	ɔ	ə	i	u
Phoneset Vowel Symbol	EH	IH	UH	EY	AY	OY
IPA Symbol	ɛ	ɪ	ʊ	eɪ	aɪ	ɔɪ
Phoneset Vowel Symbol	OW	AW	ER	AH	EA	IAX
IPA Symbol	əʊ	aʊ	ɜ	ʌ	ɛə	ɪə

Table 3.2: The consonant phoneset symbols used for the Northern Englishes experiments alongside their corresponding IPA symbols.

Phoneset Consonant Symbol	P	T	K	B	D	G	CH	JH
IPA Symbol	p	t	k	b	d	g	tʃ	dʒ
Phoneset Consonant Symbol	F	V	S	Z	TH	DH	SH	HH
IPA Symbol	f	v	s	z	θ	ð	ʃ	h
Phoneset Consonant Symbol	ZH	L	R	W	Y	M	N	NG
IPA Symbol	ʒ	l	ɹ	w	j	m	n	ŋ

We should also note here that for the results that follow, all vowels and consonants were included in the construction of the Y-ACCDIST matrices. This differs from the Y-ACCDIST experiments presented in Chapter 2 that only used vowels (minus schwa) in the construction of the models. This was in a move to more closely simulate previous ACCDIST-based studies which have only used vowel-based units. However, as noted below, in the case of the

Northern Englishes corpus, an *all-phonemes* setting achieved a higher recognition rate than a *vowels-only* setting. All phonemes will therefore be used for the experiments here. This point will be relevant to subsequent chapters in this thesis.

3.2.3 Results

The results presented here were produced by a *leave-one-out* cross-validation experimental setup, where, in turn, each speaker in the dataset became the test speaker and the system was trained on the remainder of the dataset. This was to maximise training data, as a reduced number of speakers, only 15 per class, is used in these experiments compared with the experiments shown in Chapter 2. In this configuration, where each speaker is modelled using a total of 10 minutes of net spontaneous speech, Y-ACCDIST achieves a recognition rate of 86.7% correct[2]. In a three-way classification task like this one, we would of course expect a rate of 33.3% if the system was working by chance. The confusion matrix to accompany this result is given in Table 3.3 below:

[2]When we use a *vowels-only* setting instead of the *all-phonemes* setting, we achieve a lower recognition rate of 77.8% correct. We will return to this comparison in result in Chapters 4 and 6.

Table 3.3: Confusion matrix for an accent recognition task using the Y-ACCDIST-SVM system on spontaneous speech data from the Language Change in Northern Englishes corpus (86.7% correct).

Accent	Manc	Newc	York
Manc	**13**	0	2
Newc	1	**13**	1
York	1	1	**13**

Noticeably, unlike in the accent recognition tasks undertaken in Chapter 2, the different accent groups appear to be recognised at the same rate (i.e. for each accent group, 13 of 15 speakers were correctly classified). From these results, it is not obvious whether speakers of a particular group are more likely to be recognised than speakers of another group. The low number of speakers per accent class used for these experiments also contributes to this.

Sample Duration

These results are of course unrealistic for a number of reasons. One of the main criticisms of the results is the quantity of speech used to model each speaker. This thesis considers results in the context of forensic applications, and 10 minutes of speech for an analysis is an unrealistic quantity to expect in forensic casework. The same experiments were re-run, therefore, on varying quantities of speech for each speaker.

Until now, experiments have made use of a fixed phoneset for each trial that has been run to construct the Y-ACCDIST matrices for that trial. In the context of the AISEB experiments, where a controlled reading passage was produced by all speakers, the same phones were produced by the different

speakers, and the reading passage covered nearly all the phonemes in the phoneme inventory. In the Northern Englishes experiments above, 10 minutes of speech per speaker was available, and so coverage of the phoneme inventory was not a significant concern. However, with shorter speech samples, it is quite possible that a full coverage of the phoneme inventory is not present in the test speaker's sample. To account for this, the phoneset that makes up the Y-ACCDIST matrices for each trial is determined by the phonemes that are present in the test sample.

In increments of 30-seconds, the graph below shows the recognition rates of the Y-ACCDIST-SVM system when it processes different quantities of speech. The shorter portions of the total 10 minutes available for each speaker are the first temporal portions which come up in the available 10 minutes (i.e. anchored on 00:00).

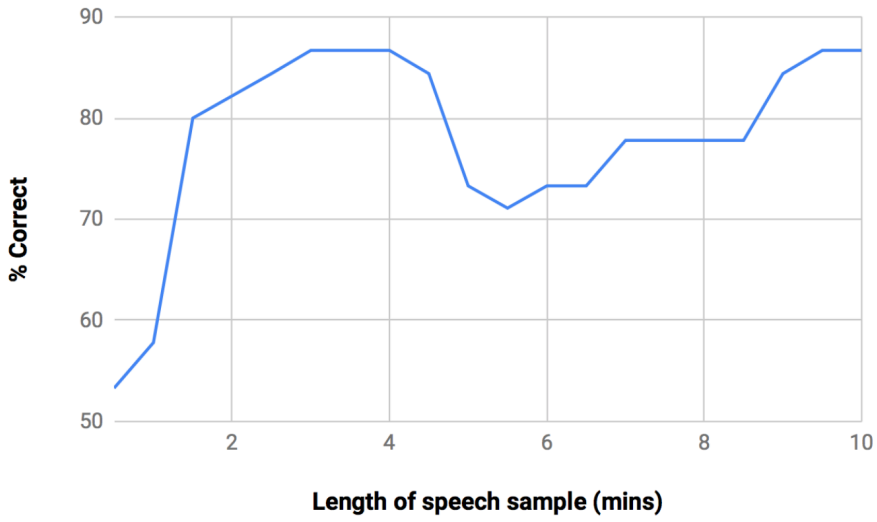

Figure 3.1: Y-ACCDIST-SVM recognition rates when processing varying lengths of speech sample.

It is no surprise that we can observe a general increase in recognition rates from 30-second speech samples up to 10 minutes. Presumably, more information relevant to speaker accent is available in longer stretches of speech, therefore increasing recognition rates. Also, longer stretches of speech mean that we have more tokens of a single phoneme, leading to more stable representations in the accent models. We seem to see a steep increase in recognition rate up to 1.5 minutes of speech (80.0% correct). However, we do witness inconsistencies in recognition rate beyond this point. This is likely to be partly attributed to the number of speakers we have used for these experiments. The drop between the performance at 4 minutes and 5.5 minutes is only a difference of 7 speakers. Using only 45 speakers to test the system naturally means that losses and gains of correctly classified speakers will result in greater impacts on the overall percentage. Nonetheless, it is still interesting to witness instability in performance as speech sample length increases.

Accent recognition on degraded speech data

Another way in which the experiments shown in this thesis so far are unrealistic is that they have all been conducted using good-quality recordings. Of course, in forensic casework, it is highly unlikely that analysts will work with recordings which are of the same quality as recordings collected for a corpus intended for sociolinguistic research. It is therefore a natural next step to find out how Y-ACCDIST performs when processing degraded data. To do this, and to attempt to make a direct comparison with the results presented above, the same speech recordings from the same speakers have been artificially degraded. Although in some ways it might be better to work with recordings which are genuine low-quality forensic recordings, we can make stronger com-

parisons with Y-ACCDIST's performance on good-quality data if we use the same material that has been degraded. The data were downsampled to 8kHz, bandpass-filtered to 250-3500Hz[3], and a-law compression was applied[4], resulting in recordings which resemble telephony. The same experimental configuration employed in the experiments above was implemented in these experiments. 10 minutes of net speech per speaker was used to model all the speakers (both training and testing), and a leave-one-out cross-validation configuration was applied. Under this degraded condition, Y-ACCDIST achieved a result of 64.4% correct. The accompanying confusion matrix for the task is given in Table 3.4 below:

Table 3.4: Confusion matrix for an accent recognition task using the Y-ACCDIST-SVM system on artificially degraded speech data from the Language Change in Northern Englishes corpus (64.4% correct).

Accent	Manc	Newc	York
Manc	**7**	4	4
Newc	0	**13**	2
York	2	4	**9**

As we might expect, degrading the data has resulted in a reduction in the overall recognition rate, from 86.7% correct to 64.4% correct. By degrading the data in the way we have, we are likely to lose information that is telling of a speaker's accent. An overall reduction is therefore unsurprising. One

[3]A wider bandwidth (upper frequency limit of 4kHz) would be more representative of modern telephony, but the specifications implemented here still produce data that are close to telephone quality.

[4]a-law has been used for these experiments because these is the companding algorithm used in Europe.

interesting observation we can gather from the confusion matrix is that the even distribution of correct classifications that we see across the accent groups in Table 3.3 has been lost. Of course, it is important to note that when only 15 speakers per accent have been used, we can only speculate. However, it is interesting to see that the number of Newcastle speakers correctly classified remains approximately the same, while we witness a drop in the number of correctly classified speakers from the other two accent groups. This might be an indication that the Newcastle accent variety is more distinct overall among these three accent varieties. It seems that more information has been lost in the degradation that is important to characterising Manchester and York speakers, relative to the Newcastle sample.

3.3 Individual Speaker Similarity

The results so far in this chapter have simply categorised individual speakers based on models trained on a collection of speakers. However, it might be of interest to observe how individual speakers sit in relation to other speakers in the whole dataset. This might reveal more about how Y-ACCDIST matrices characterise speech samples. This section aims to take a deeper look into how individual speaker Y-ACCDIST matrices model a speaker's speech sample by looking at the degree of similarity between these matrices. This section does this through two types of visualisation: *swarmplots* and *multidimensional scaling*.

3.3.1 Swarmplots

Taking inspiration from the success of a similarity metric deployed in one of the systems developed and tested in Chapter 2, the outputs in this section will make use of Pearson r product-moment correlation to measure similarity between individual speakers' Y-ACCDIST matrices. We saw in Chapter 2 that Pearson r product-moment correlation, when attached to Y-ACCDIST as the classification mechanism for the AISEB corpus, achieved a recognition rate of 76.7%. From this, Pearson r product-moment correlation appears to be a suitable method for measuring similarity between Y-ACCDIST matrices. These measurements will then be visualised through swarmplots.

Method

Each speaker's 10-minute speech sample (sampling rate of 44.1kHz) in the dataset used in the experiments above in this chapter was processed and modelled as a Y-ACCDIST matrix in the usual way, making use of only the vowel phonemes. Only vowel phonemes have been used because it seems that we can witness clearer separation of speakers in this setting. The Pearson r product-moment correlation was then calculated between every possible pair of speakers available in the dataset.

Outputs and Analysis

The resulting speaker-pair correlation values are presented in the 'swarmplots' below. These swarmplots show the degree of similarity between different categories of speaker and individual speakers. Each speaker is taken individually and the correlation value is subtracted from 1. Each speaker's 'swarm' in the swarmplots therefore shows the more similar speakers lower down on the y-axis

(i.e. points closer to the labelled x-axis indicate a higher similarity measurement)[5]. For the purposes of clarity, the outputs below show the individual speakers in each category in three separate plots. The Manchester speakers' swarmplot is shown in Figure 3.2, the Newcastle speakers' swarmplot is shown in Figure 3.3, and the York speakers' swarmplot is shown in Figure 3.4.

To decode the speaker ID labels in the figures in the remainder of this chapter, Table 3.5 clarifies what each of the component parts correspond to:

Table 3.5: Table to decode speaker ID labels in the Northern Englishes corpus.

Part of Speaker ID label	**Corresponding speaker property**
The first three letters **MNC**YMABC	The geographical accent group of the speaker MNC = Manchester, NCL = Newcastle, YRK = York
The fourth letter MNC**Y**MABC	Age group of the speaker Y = younger, M = middle-aged, O = older
The fifth letter MNCY**M**ABC	Sex of the speaker M = male, F = female
The final three letters MNCYM**ABC**	Correspondence to the speaker's pseudonym assigned during data collection

[5]The subtraction from 1 reinforces the mapping between distance shown on the swarmplot and similarity

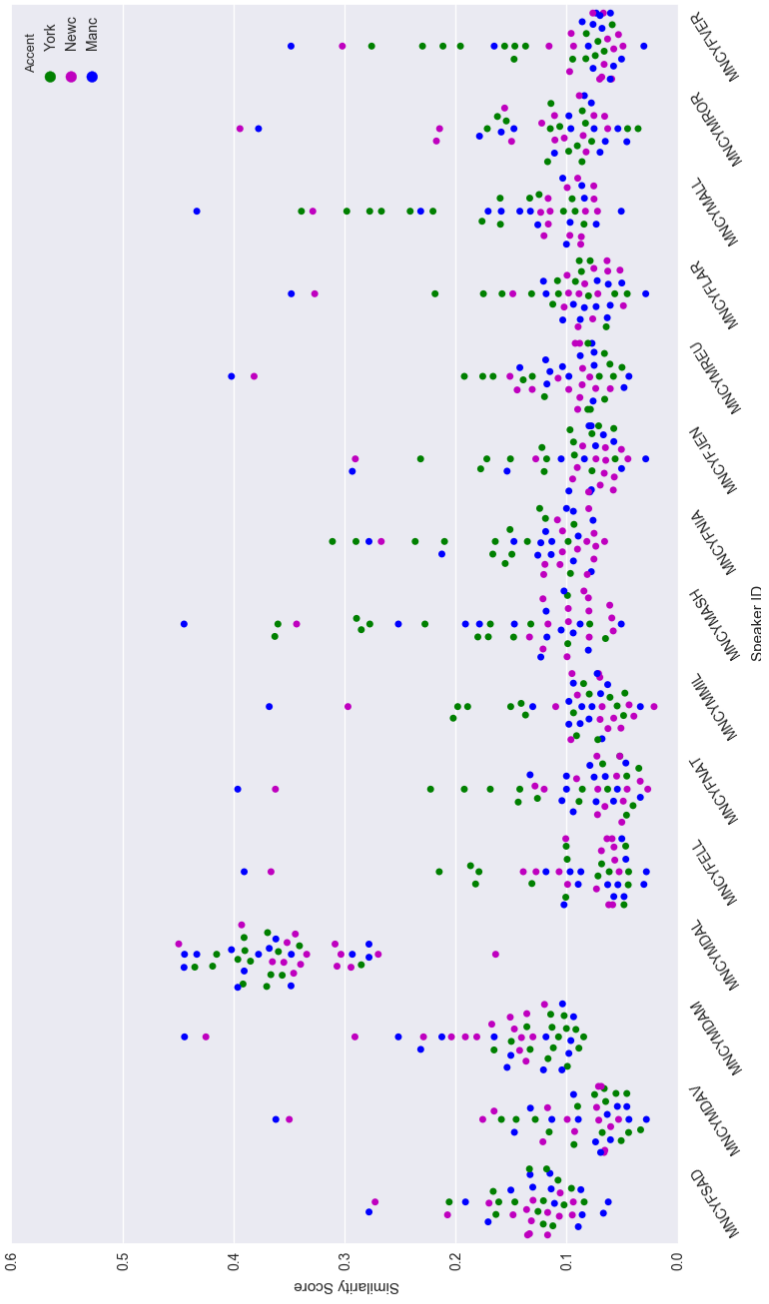

Figure 3.2: Swarmplot showing correlation values between individual young Manchester speakers and the rest of the 44 speakers in the dataset.

Figure 3.3: Swarmplot showing correlation values between individual young Newcastle speakers and the rest of the 44 speakers in the dataset.

Figure 3.4: Swarmplot showing correlation values between individual young York speakers and the rest of the 44 speakers in the dataset.

The swarmplots above do not appear to display the clear banding of colours we might expect. In the case of Figure 3.2, for example, we would expect to find a band of blue points lower down on the y-axis to show that these points are more similar to the individual Manchester speakers labelled on the x-axis. We would then expect to see the points representing Newcastle and York speakers falling higher up on the y-axis to reflect a lower degree of similarity between these speakers and the individual Manchester speakers. Instead, we see more of a mix of these speaker points, but more specifically, we appear to see a combination of Manchester and Newcastle speakers sitting closer to the individual Manchester speakers, with York speakers generally sitting further away and sharing a larger range of similarity. Because we do not see the clear bands of speaker categories we might expect, this suggests that the individual Y-ACCDIST models do not carry a strong accent characterisation alone. This could be because of the instability that spontaneous, content-mismatched speech data brings to an accent model, and so a number of matrices to represent an accent group is required to construct a strong enough accent representation. Overall, these swarmplots demonstrate the high degree of similarity that exists between the Northern Englishes speaker models, whether they are members of the same accent group or not.

These swarmplots also reveal individual speakers who seem to be significantly more dissimilar from the rest of the speakers in the dataset. Among the young Manchester speakers in Figure 3.2, we see this for speaker MNCYMDAL, and in the case of young Newcastle speakers, speaker NCLYMCHR behaves similarly. Both speakers stand out among the dataset, being seemingly dissimilar from the rest of the speakers. This could be for a few reasons. One could be that these speakers are not typical of any of these Northern varieties of English, to the point of being distant from all other speakers. Another rea-

son might be that these speakers generated a particularly high segmentation error rate in the forced alignment process, meaning that the segmental representations which form the foundations of the Y-ACCDIST matrix are more inaccurate than those for other speakers.

Individual Speaker Similarity using Degraded Speech Data

We can also generate swarmplots of these speakers once we have artificially degraded their speech samples. These are presented below for individual speakers in each of the three accent groups.

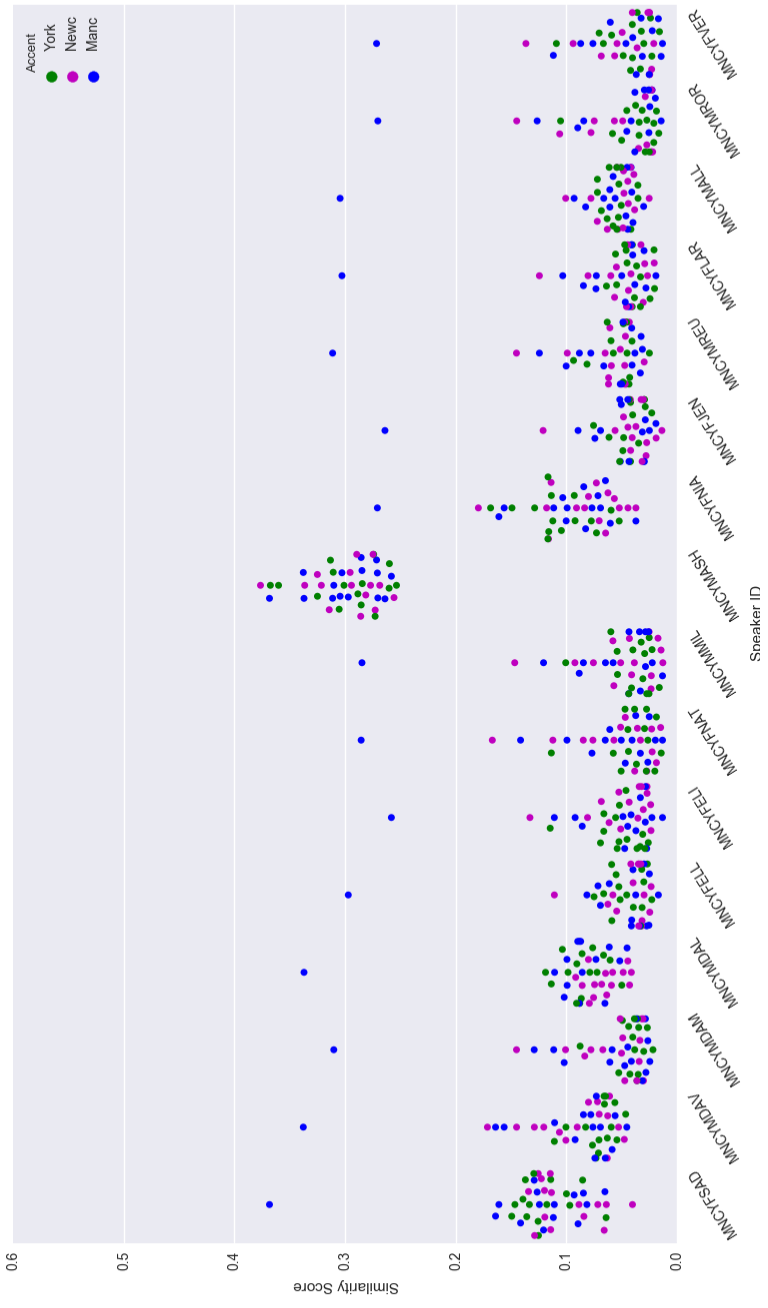

Figure 3.5: Swarmplot showing correlation values between individual young Manchester speakers and the rest of the 44 speakers in the dataset with artificially degraded speech samples.

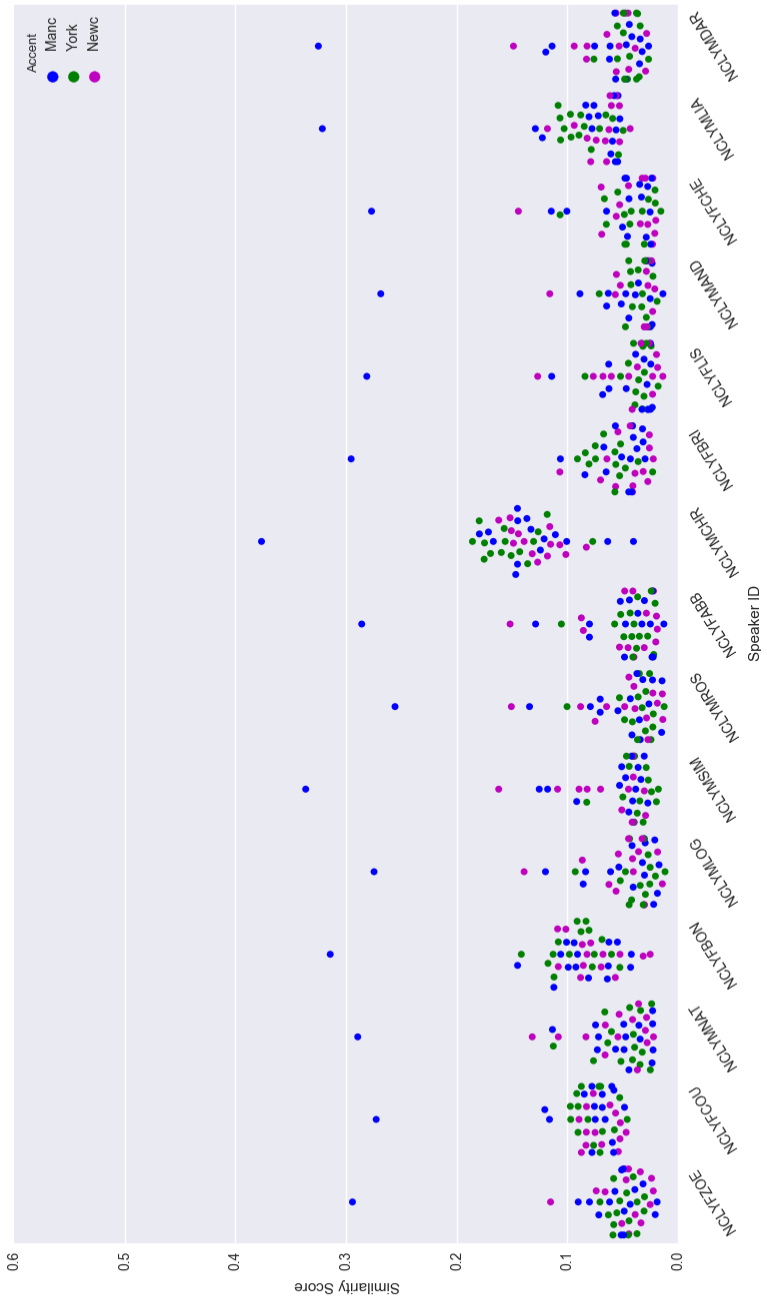

Figure 3.6: Swarmplot showing correlation values between individual young Newcastle speakers and the rest of the 44 speakers in the dataset with artificially degraded speech samples.

Figure 3.7: Swarmplot showing correlation values between individual young York speakers and the rest of the 44 speakers in the dataset with artificially degraded speech samples.

These swarmplots generated using degraded data reveal some interesting effects that degrading data has on the Y-ACCDIST matrices, not just the overall recognition rate that is outputted after the whole classification process. Firstly, what stands out about the swarms using degraded data, compared with the swarms generated from good-quality recordings shown in Figures 3.2, 3.3 and 3.4, is the apparent overall increase in similarity between speakers in the dataset. We also see that the range in which the speakers fall seems to have been decreased. Overall, with the artificial degradation of the speech data, we observe an increase in similarity among the speakers, which suggests that the degradation is reducing the amount of variation represented among these speakers. It seems the Y-ACCDIST matrices are made more similar by reducing the amount of information which distinguishes speakers in the degradation. An alternative hypothesis might have been that degradation decreases the degree of similarity among the speakers, because the resulting segmentation error from the forced alignment might increase. This would mean that the segmental representations might destabilise, therefore causing a more random distribution. However, we are witnessing the opposite effect in these swarmplots.

Additionally, it appears that individual speakers are affected differently by the degradation. For example, we no longer see speaker MNCYMDAL as being much more distant from the other speakers. However, a different speaker (MNCYMASH) is seen to be much more dissimilar from the rest of the speakers under the degraded data condition. From listening to these recordings (among other recordings from this dataset), there did not appear to be anything particularly unusual about these specific recordings. It is interesting to see that the recording quality affects individual speakers in different ways. It might be that degradation has increased the forced alignment segmentation error for

MNCYMASH, or that less information that distinguishes MNCYMASH from the rest of the speakers in the dataset has been removed by the degradation.

3.3.2 Multidimensional Scaling

Another way we can observe how these speakers are modelled is through Multidimensional Scaling (MDS). Ferragne and Pellegrino (2010) conducted a MDS analysis, combined with ACCDIST-based models, to observe similarity between the 14 accent groups represented in the Accents of the British Isles (ABI) corpus (D'Arcy, Russell, Browning and Tomlinson, 2004). This was done using highly controlled wordlist data. We can also conduct an MDS, combined with ACCDIST-based models, to observe the degree of similarity between individual speakers. In a similar way to the swarmplots above, this should shed some light on how Y-ACCDIST models the data we have used in this chapter. It could also reveal points of interest about the data.

Because it is easier to visualise a large amount of data with MDS than it is with the swarmplots, the dataset has been increased for this subsection to include additional Manchester speakers. We will therefore use the Y-ACCDIST models of 15 younger Manchester speakers, 15 younger Newcastle speakers and 15 younger York speakers (these are the speakers that were used in the experiments above in this chapter), as well as 10 older Manchester speakers and 10 middle-aged Manchester speakers. By including extra speaker groups of a different sort, we are more likely to see how Y-ACCDIST modelling works.

Once each speaker's speech sample has been processed and speaker-specific Y-ACCDIST matrices have been generated, we calculate the Euclidean distance between all the possible pairs of speakers so as to grasp the degree of similarity between them. MDS then seeks to project our data points (indi-

vidual speakers) in a low-dimensional space, staying as true as possible to the original degrees of similarity computed between the speakers via Euclidean distances. To conduct this reduction, a loss function, *stress*, is used to measure how well a projection of the data points in a low-dimensional space (e.g. two dimensions) maps on to the original high-dimensional space (McDougall, 2013: 168). The resulting 2-dimensional scatterplot for these Northern Englishes speakers can be found in Figures 3.8 and 3.9 below. The first is for the good-quality data condition, and the second is for the artificially degraded data condition.

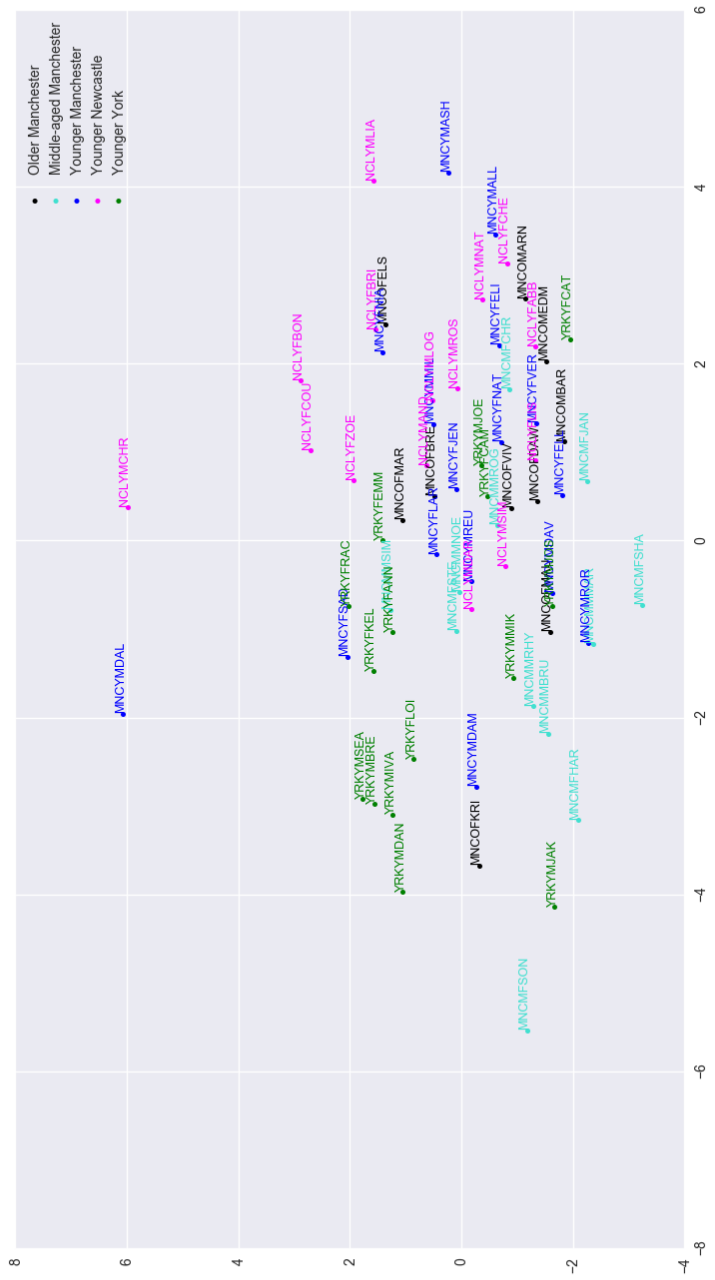

Figure 3.8: Multidimensional scaling output of Northern Englishes speakers, having modelled each speaker's speech sample (good-quality condition) as a Y-ACCDIST matrix.

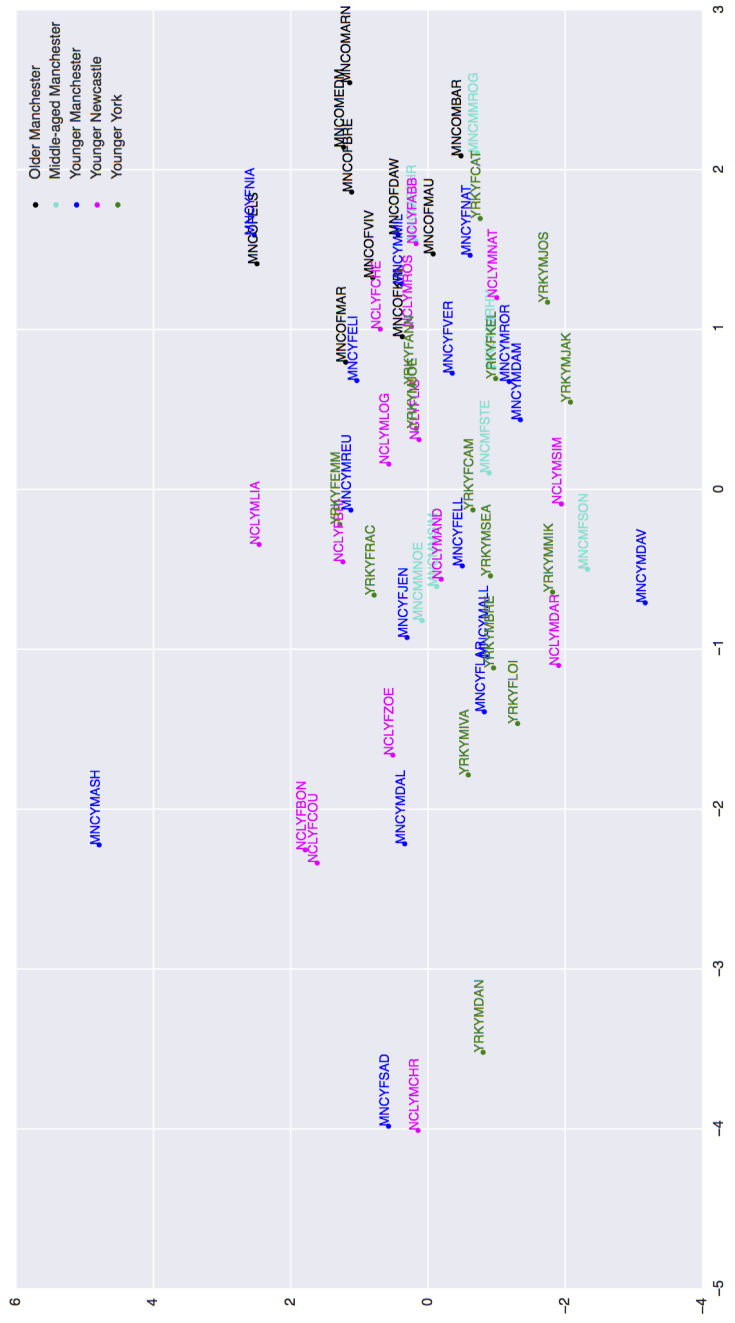

Figure 3.9: Multidimensional scaling output of Northern Englishes speakers, having modelled each speaker's speech sample (artificially degraded condition) as a Y-ACCDIST matrix.

These MDS plots show some correspondence with the swarmplots, but there are also some things that they expose that the swarmplots did not.

Firstly, we see the overall difference in degree of similiarity between the good-quality condition and the artificially degraded condition. By simply observing the x and y axes, we can see that different scales are being used. A smaller range is used for the artificially degraded data. This corroborates our observations from the swarmplots in Section 3.3.2, in which we saw a greater degree of similarity among speakers once their recordings had been artificially degraded. While we can see a large amount of overlap between the groups, the good-quality recordings (Figure 3.8) seem to provide clearer clusterings of the accent groups of the sort expected. For example, the York speakers (green labels) and the Newcastle speakers (pink labels) seem to cluster at different sides of the population, with only a small amount of overlap. Such a clear grouping is lost in the case of artificially degraded recordings in Figure 3.9.

Secondly, we see the same speakers that were identified in the swarmplots as particularly distinct from the rest of the population as rather distant from the rest of the population in the MDS plots. In the case of the good-quality data, these were MNCYMDAL and NCLYMCHR. In the case of the artificially degraded recordings, the speaker identified as particularly distinctive was MNCYMASH.

The MDS plots, however, more clearly indicate the general degree of speaker similarity within each accent group by using two values (each dimension), rather than just using one (which is what was implemented to form the swarmplots). It is interesting to observe that the older Manchester speakers occupy a much smaller space as a group in this plot than do the other groups. The middle-aged Manchester speakers similarly occupy a small area (but a larger one than the older category), whereas the three accent groups of younger

speakers appear to occupy larger areas. This could be an indication of the difference in degree of variation within different speaker age groups. That is, there could be a greater degree of variation in the pronunciation systems of younger speakers. It would be of interest to conduct a phonetic analysis of the data to observe whether this observation is confirmed.

While there are ways in which the swarmplots and MDS plots correspond, there are perhaps more details revealed by the MDS plots. These visual outputs have allowed to us to better understand how Y-ACCDIST models speakers using good-quality and artificially degraded recordings.

3.4 Discussion

The Northern Englishes corpus has provided a means to test our ACCDIST-based system (Y-ACCDIST) on spontaneous (content-mismatched) speech data. Past ACCDIST-based systems have been restricted only to tasks involving content-controlled data, whereby the speakers produce read prompts or word lists. As already described, what separates Y-ACCDIST from past ACCDIST-based systems is that it collapses speech segments into their phoneme classes, enabling the possibility of processing content-mismatched data. The experiments in this chapter do indeed suggest that Y-ACCDIST can work on spontaneous speech by exceeding chance expectations on a three-way task. However, in an ideal world, we would prefer to run these experiments on the same speakers and varieties producing reading passage data. This would allow us to make a more direct performance comparison between content-controlled and content-mismatched data. The Northern Englishes corpus was selected based on the quantity of transcribed spontaneous speech per speaker that it provides, so allowing the experiments in this chapter to be run. Future re-

search could make a more direct comparison between content-controlled and content-mismatched data.

The overall accent recognition results showed the expected drop in performance between using good-quality recordings to train and test the system, and using artificially degraded data. Taking a closer look at the swarmplots between the good-quality recordings and the artificially degraded recordings, we can grasp the kind of effect that data degradation has on the Y-ACCDIST models, rather than simply on the overall recognition rate. Degradation of the data appears to increase the degree of similarity among speakers. We can suppose, then, that it is this increase in similarity that we see among the models that is responsible for the increase in errors in the degraded data condition. It would be interesting, and potentially of use in forensic applications, to test the system on mismatched data conditions and to see the effects of this. By this, we mean that the system is trained on data of a certain quality (e.g. good quality) and tested on another quality type (e.g. degraded). In the context of automatic speaker recognition technology for forensic applications, this kind of data mismatch has been researched to observe how robust speaker recognition systems are to it (e.g. Alexander, Botti, Dessimoz and Drygajlo (2004)). Botti, Alexander and Drygajlo (2004) build on this and suggest ways of compensating for data mismatch in speaker recognition systems. Carrying out this line of research with automatic accent recognition systems might also be worthwhile when considering them for forensic tasks. However, it is more likely that a tool like Y-ACCDIST is more suited to an application like LADO (as indicated previously in this thesis), where we are more likely to have more control over the kind of data we use. The work in this thesis therefore focusses on the matched condition.

We have also observed how individual Y-ACCDIST matrices model speak-

ers. Used alone, it does not seem that speaker-specific matrices are particularly useful at categorising speakers. In the similarity outputs (swarmplots and MDS plots) we did not see clear bandings or groupings of speakers belonging to the same accent class. It would appear that to develop a strong model of accent, multiple speakers from the same accent group need to be used to form that model (and this is why we find good recognition rates in an accent classification task). Despite this finding, we are able to make interesting observations about individual speakers or groups of speakers via these outputs.

3.5 Summary

This chapter has reported the performance of the Y-ACCDIST-SVM system when challenged by spontaneous (content-mismatched) speech data rather than reading passage (content-controlled) recordings, showing that the system still performs well above chance level on these particular Northern English accent varieties (86.7% correct). To challenge the system further, Y-ACCDIST-SVM was tested on the same speech recordings once they had been artificially degraded to specifications similar to telephony. An expected decrease in performance (to 64.4%) was observed in this condition. These experiments were run in a bid to test this accent recognition system on more forensically relevant data, compared to the experiments run in Chapter 2 of this thesis. The chapter then progressed towards taking a closer look at the Northern Englishes data, and how Y-ACCDIST modelling appears to characterise speakers' speech samples. This was done by processing the samples in the usual way and computing an individual Y-ACCDIST matrix to represent each speaker. Pearson r product-moment correlations between all the possible speaker pairs were calculated to gauge the degree of similarity between speakers to produce

swarmplots, and Euclidean distance was used as part of the Multidimensional Scaling analysis. Based on these speaker-on-speaker comparisons, it seems that Y-ACCDIST matrices alone do not appear to form a strong enough representation of a speaker's accent. Indeed, they seem to be extremely similar to one another, regardless of accent group. Interestingly, when we look at these individual speaker similarity measures once we have artificially degraded the data, it appears to raise the degree of similarity among the speakers, and reduces the variation represented in our dataset.

CHAPTER 4

Incorporating feature selection into automatic accent recognition

4.1 Introduction

The term *feature selection* is used to describe the process of determining a 'good' subset of 'features' from a larger feature set in order to conduct a given task. Feature selection usually refers to automatic, or objective, methods of achieving this. Feature selection is implemented across a range of research domains. Guyon and Elisseeff (2003) discuss it in the context of gene selection and text categorisation, as just two example areas that commonly apply feature selection. Feature selection is a means to overcome the so-called *curse of dimensionality*, which acknowledges the problems that come with using high-dimensional data (Keogh and Mueen, 2011; Bellman, 1957). When we use high-dimensional data to work on problems, we run the risk of including information that does not necessarily contribute to a given task. In fact, some data features (dimensions) might add 'noise' to a process, which may divert

conclusions away from an accurate outcome.

Feature selection is usually associated with 'big data' problems (e.g. Bolón-Canedo, Sánchez-Maroño and Alonso-Betanzos, 2015). However, it is proposed here that feature selection could be useful to forensic tasks. A forensic speech practitioner may well be asked to conduct an analysis involving any linguistic variety. The number of 'relevant' populations (i.e. datasets) available to the forensic practitioner tends to be limited. It is likely that a dataset containing a relatively small number of speakers is available for a given case. However, the number of potential features available for analysis does not change. Feature selection techniques can therefore be useful in this kind of scenario by reducing the entire feature set to a smaller subset of features useful to a specific task. Also, the objectivity of feature selection techniques is seemingly advantageous. Rather than an analyst making assumptions about which features might be most valuable in a given task (this might be based on the phonetics literature or auditory judgment), we can also estimate, using statistical or computational methods, which features are most valuable using feature selection. Such an approach could help to avoid some useful features being overlooked. There are also likely to be some advantages surrounding the amount of labour involved in analysing a dataset. Computational methods can of course save a considerable amount of time for the analyst.

This chapter deals with feature selection in relation to automatic accent recognition. More specifically, we look at whether integrating feature selection into the Y-ACCDIST-SVM system (the highest-performing system shown in the experiments in Chapter 2) will improve performance further. This is a development on past studies which have looked at the performance of ACCDIST-based systems. A number of assumptions have been placed on these past systems and feature selection could help to remove these assumptions. The

ACCDIST-based systems developed and tested in Huckvale (2004, 2007) and Hanani, Russell and Carey (2013) focussed on only vowel-based units. In Huckvale's system these were word-specific vowels, whereas Hanani *et al*'s system used triphone-based vowel units (this difference was more comprehensively described in Chapter 3 in Section 3.1). To match these systems for a closer comparison, the Y-ACCDIST experiments that were run in Chapter 2 only included vowel phonemes. To include only vowel segments is a decision made by the developers of these systems. While a lot of sociophonetic research does indeed focus on vowel differences between accents, and we can expect vowels to differ between accents, we know that consonant realisations also vary across accents. By excluding consonants from these systems, we are possibly ignoring valuable discriminatory information in an accent recognition task. We noted in Chapter 3 that in the case of the Northern Englishes corpus that using an all-phonemes setting in the Y-ACCDIST-SVM system, we achieve a higher accent recognition rate than we do using the vowels-only setting. However, in the case of accent recognition experiments on the AISEB corpus, the reverse is the case (the vowels-only setting - 86.7% correct - outperforms the all-phonemes setting - 80.8% correct). This contrast in performance demonstrates that the optimum segmental configuration for accent recognition using the Y-ACCDIST system is data-specific. Another assumption that was made in Huckvale (2004, 2007) and Hanani *et al* (2013) was that including schwa was not likely to assist with the classification task, and so schwa-based units were also excluded from the analysis. Contradicting this assumption, there are indeed some accents where schwa might offer discriminatory power (Watt, Llamas, French, Braun and Robertson, 2016), and so we perhaps should not overlook these speech segments when conducting these tasks. Feature selection could help us to move away from these assumptions and to estimate a

combination of both vowels and consonants that are better at discriminating between a specific set of accents in a dataset, rather than just choosing a vowel-based system. Of course, we might also want to use this kind of technology on linguistic varieties that we have little knowledge of, and therefore would not know which phonemes would be best to include. Feature selection is a way of overcoming a need for prior knowledge. These are key motivations behind conducting the experiments in this chapter.

As already mentioned, feature selection is incorporated across a range of disciplines. Even when we just look at the area of speech technology, we can see this in language recognition, emotion recognition and speaker recognition (e.g. Richardson and Campbell (2008), Vogt and André (2005), and He, Wornell and Ma (2016)). Some of these applications will be discussed in the following section. This chapter not only looks at whether feature selection can contribute to the performance of the Y-ACCDIST-SVM system, but it also compares the outputs based on the two different accent corpora (containing different sets of British English accents) used in this thesis. By making a cross-corpus comparison like this, we can observe whether the same features are highlighted as valuable for both sets of accents or if we can see corpus-specific tendencies. One consideration related to this is whether the phonemes that are naturally more frequent in the data are better accent discriminators (because a higher number of tokens presumably leads to stronger phoneme representations in the Y-ACCDIST models). Comparing two corpora allows us to speculate with greater confidence about these kinds of factors. Particularly when considering forensic applications, comparing the analysis of more than one corpus is important to do. If feature selection is not necessarily appropriate to use on all sets of accents, we should find out which datasets it is appropriate for, and consider why this might be the case.

4.1.1 Outline

Based on the discussion above, this chapter has four key objectives:

1. Compare the performance of two feature selection methods when they are incorporated into the Y-ACCDIST-SVM system. This comparison largely takes place in Section 4.3.2.

2. Compare the effects of these feature selection methods on two different accent corpora to monitor how feature selection transfers between datasets. This comparison is mainly conducted in Section 4.3.3.

3. Inspect and compare the rankings of features produced by the feature selection methods across the two corpora. We raise points of interest in relation to these rankings across both sections 4.3.2 and 4.3.3.

4. Explore the role of phoneme frequency on the estimated contribution a single phoneme makes to a given accent recognition task. This takes place Section 4.3.4.

Before experiments are run and outputs are produced to address these four objectives, Section 4.2 reviews past work on feature selection.

4.2 Previous Work on Feature Selection

Feature selection has multiple benefits:

Lowers computational cost

One function of feature selection is to lower computational cost. If we can effectively select only the features that contribute to the task at hand, we should achieve at least the same level of performance, while doing it using less information. He, Wornell and Ma's (2016) research into 'low-power' speaker recognition is motivated by achieving computational efficiency, and one of the things they focus on to do this is feature pre-selection so as to avoid taking unnecessary signal information through the whole speaker recognition process. In the context of security systems or similar commercial applications (e.g. building access applications), this is well-motivated research to make these sorts of technologies suitable for a wider range of applications. These kinds of applications cannot afford to have long processing times once a speech sample has been submitted. When developing a system for forensic casework, lowering computational cost might not necessarily be the priority aim for integrating feature selection, but improving performance would be. This potential benefit is covered by the following point.

Removes 'noisy' features

As mentioned in Section 4.1 above, feature selection can be used to remove 'noisy' features. These features are unlikely to benefit an analysis by including them, and can even prevent reliable analyses taking place. Often we can expect the removal of 'noisy' features to improve the overall performance of a

system.

Provides a ranking of useful features

While the above two points aim to make contributions to system performance in two different ways (increases in efficiency and success rates), we can also produce helpful outputs from a feature selection task, in the form of a ranking of features. This might simply be informative, but it could be useful to different applications, as we will explain in the context of forensic speech science further below.

4.2.1 Feature Selection in Speech Technology

This subsection reviews literature, specifically within speech technology, that looks at the effects of feature selection for different types of problem. It finally arrives at feature selection with specific reference to the task of automatic accent recognition.

We first look at automatic emotion recognition through speech samples, because of the amount of feature selection literature in this area. On the basis of acoustic information derived from a speech sample, can we classify the emotional state of the speaker? The area of emotion recognition via the speech signal has received attention in relation to feature selection because of the scarcity and small sizes of the relevant databases available (Rong *et al*, 2009). Forensic speech science shares some of the challenges faced by emotion recognition in that the number of relevant speech databases for a given case might be small, as well as the size of the databases themselves. When using high-dimensional data, having a small number of training samples is a problem. This is because *overfitting* can happen, whereby the high-dimensional data can

perfectly fit to the small number of data points, but does not classify unseen data samples well. If we can identify just a subset of useful features, we avoid this problem of overfitting on small datasets.

Automatic emotion recognition usually focusses on the classification of speech samples into one of a few classes. Schüller *et al* (2004), for example, ran experiments classifying speech samples into one of seven classes: anger, disgust, fear, joy, neutral, sadness and surprise. Emotion recognition experiments by other researchers tend to use a similar set of categories. Rong, Li and Chen (2009) apply different feature selection methods to an automatic emotion classification system. The original number of acoustic features was 84, consisting of values to do with pitch and MFCC coefficients, for example. They found that by reducing the number of acoustic features to 16 (those that are estimated to be the 16 most useful) improved emotion recognition rates by 2.01%.

Pohjalainen, Räsänen and Kadioglu (2015) compared feature selection methods on other kinds of paralinguistic classification tasks. They used a range of corpora that allowed them to conduct a number of binary classification tasks. They used the *Speaker Likeability Database* (SLD) (Burkhardt, Schüller, Weiss and Weninger, 2011), which contains recordings that have been rated by listeners according to how likeable they thought the voices in the recordings were. From this rating, Pohjalainen computed whether each speaker was rated as either likeable or not overall. In a similar way, the NKI CCRT Speech Corpus (van der Molen *et al*, 2012) was used to conduct a binary classification task of assigning speakers to either an *intelligible* class or an *unintelligible* class. A third corpus was used to form binary classifications for five different personality traits of speakers: openness to experience, conscientiousness, extraversion, agreeableness and neuroticism. This was the *Speaker Personality Corpus* (Mo-

hammadi and Vinciarelli, 2012). Again, a process using listeners was used to determine whether speakers' speech samples were conscientious-sounding or not (for example). When applying their feature selection techniques to each of these classification tasks, they were reducing the feature set size from all possible features available to between 2% and 6.2% of the entire original feature set. This is a huge reduction in the number of features, and they found that classification performance was either approximately the same as using the entire set of features, or performance was marginally improved. In the case of these results, then, we can at least lower the computational cost of a process, even if we do not see an improvement in performance.

Wu, Duchateau, Martens and Compernolle (2010) compared methods of feature selection on an automatic accent recognition task. They used two different modelling strategies (involving GMMs), three different types of classifier, and trialled three different feature selection methods in different combinations. System 4 in Chapter 2 of this thesis (the Phonological GMM-SVM system) is based on the system used in Wu *et al*'s study, but without a feature selection step integrated. The focus of their paper was to compare these methods, but they also looked at 'hybrid' methods that combined these different feature selection techniques.

Overall, Wu *et al.* found that incorporating feature selection into a text-dependent automatic accent recognition system does improve recognition rates, with 10% of the total number of available features achieving the best performance.

Like Wu *et al*, this chapter incorporates and evaluates feature selection methods in text-dependent automatic accent recognition. More specifically, we are interested in whether feature selection can be an advantageous step when integrated into the Y-ACCDIST-SVM system, the highest-performing

system from Chapter 2 (which outperformed the accent recognition system that was replicated from Wu *et al* (2010)). The experiments in this chapter first compare two feature selection methods that were shown to be beneficial in Wu *et al*'s (2010) study (Analysis of Variance and Support Vector Machine Recursive Feature Elimination) using the AISEB corpus, the larger corpus used so far, to establish whether feature selection might have something to offer the overall Y-ACCDIST-based process. We then apply Y-ACCDIST-SVM with these feature selection methods to the Northern Englishes corpus to see how the outcome alters when we use it on different data that present different challenges. These different types of problem will be emphasised below.

4.3 Experiments

The following experiments investigate how two feature selection methods affect overall accent recognition rates. First, this section will introduce the two feature selection methods being used in this chapter, and will then move on to their performances when integrated into the Y-ACCDIST-SVM system. These variant systems will be trained and tested on two different corpora in turn: the AISEB corpus (Watt, Llamas and Johnson, 2014) (already described in Chapter 2 in Section 2.3.2) and the Northern Englishes corpus (Haddican, Foulkes, Hughes and Richards, 2013) (already described in Chapter 3 in Section 3.2.1). The purpose of conducting these experiments on two different corpora is to observe whether feature selection has the same effect across different sets of accents.

4.3.1 Feature Selection Methods

The experiments in this chapter compare two feature selection methods that were used in the automatic accent recognition experiments in Wu *et al* (2010): *one-way ANOVA* and *Support Vector Machine Recursive Feature Elimination* (SVM-RFE). Within the Y-ACCDIST-SVM system, the feature selection method being tested will be integrated into the recognition process after the speakers have been modelled as Y-ACCDIST matrices. Feature selection will identify a subset of the Y-ACCDIST matrix elements and the reduced matrices are then taken through the classification process. The features we are therefore analysing are the individual Y-ACCDIST matrix elements (phoneme-pair distance values).

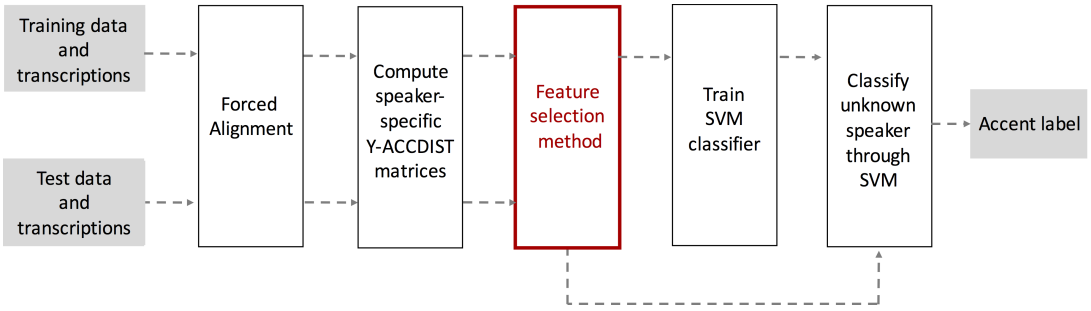

Figure 4.1: Y-ACCDIST-SVM system diagram with feature selection integrated.

Each feature selection method is introduced below:

1) One-way Analysis of Variance (ANOVA)

One-way ANOVA is a statistical technique that, in this application, will take the speaker Y-ACCDIST matrices for each accent, and determine the mean for each of the matrix elements. Using these means and the variance of these values, we can calculate a p-value that indicates whether a given feature is significantly different across the accent groups. The smaller the p-value, the more distinctive that feature is estimated to be between these accent groups. Based on these p-values, we can generate a ranking of all the phoneme-pair distances. We can then specify a number of features to include in the analysis and take the top-ranked features up to this number.

One problem with ANOVA is that it does not take into account the features that correlate with one another. It takes each matrix element independently and calculates a p-value for that element, disregarding the values generated for other matrix elements and how the matrix elements work together as a set. Correlating features are features where the value of one could be used to predict another. This means that taking features that correlate with each other is not necessarily going to contribute more to the overall analysis. In fact, including correlating features could introduce 'noise' to the process. Ideally, we should combine features that do not correlate. In the context of Y-ACCDIST, there are a lot of features that are likely to correlate. Within the Y-ACCDIST matrices we have distance values between all the phoneme-pair combinations possible in the phoneme inventory. This means that we have a lot of features that are based on the same phonemes. This is one reason why ANOVA may not be the best feature selection method for a Y-ACCDIST system, but we nonetheless observe its benefits to an analysis in the experiments below.

2) Support Vector Machine Recursive Feature Elimination (SVM-RFE)

Support Vector Machine Recursive Feature Elimination (SVM-RFE) begins by taking all of the available features and, one-by-one, assesses each one's contribution to the training speakers' separation within the SVM. The one that is estimated to be the least beneficial to a task is removed, and the process starts again by assessing all the features in the remaining set, one-by-one. In the case of Y-ACCDIST, the system will take speaker models consisting of all of the available matrix elements (phoneme-pair distances between all the vowels and consonants in the phoneme inventory). The feature that is assessed to contribute the least to the task is removed, and the elimination process starts again, gradually reducing down the number of matrix elements.

The problem of correlating features is expected to be less of a concern when using SVM-RFE, compared to using ANOVA. While ANOVA will take each Y-ACCDIST matrix element at a time and determine its significance to the task independently, SVM-RFE initially takes all of the matrix elements together and assesses them as a set. Looking at the features as a set, rather than looking at each feature independently, reduces problems around correlating features.

One downside to SVM-RFE, however, is that it is computationally much more expensive. Another is that we expect it to require a large amount of data for it to work effectively. It may well be the case that the datasets that we are using here are not large enough for this method.

4.3.2 Experiments on the AISEB corpus

This first set of experiments will allow us to observe and compare the performances of the two feature selection methods on the AISEB corpus. The

experiments presented in Chapter 2 involved the AISEB corpus, and so more detail about the data can be found in Section 2.3.2. The data from the AISEB corpus used for these experiments consisted of speech samples produced by 30 speakers from each of four locations along the Scottish-English border. Each speaker's speech sample is a recording of the reading passage task the informant was asked to carry out. For the purpose of these experiments, this means that phonemic coverage (that is, the extent to which each phoneme in the system is represented) will be roughly equal for all speakers, such that a controlled comparison can take place.

Effect on Performance

For each of the two feature selection methods, the following experimental process was applied. Using all the vowels and consonants available in the phoneme inventory, 120 speakers from the AISEB corpus (30 from each of the 4 locations) were represented by Y-ACCDIST matrices using the reading passage recording. In a leave-one-out cross-validation setup, each speaker became the unknown test speaker, while the rest of the speakers were used to train the Y-ACCDIST-SVM system. For each rotation (i.e. each time a speaker became the unknown speaker), the feature selection method was applied so as to reduce the total number of features used to train the SVM, ready for classification. At this set number of features, we reduce the number of Y-ACCDIST matrix elements to the ones that have been ranked as the top n matrix elements from the training data, and we generate a recognition rate with this number of features. In increments of 5[1], we establish the recognition rate at the given number of features. Figure 4.2 below shows the effect on performance when a

[1]Increments of 5 have been used, rather than increments of 1, due to the high computational cost of running the SVM-RFE feature selection method.

certain number of the top-ranked features (Y-ACCDIST matrix elements) is used to model the speakers, which then go on to train the SVM and classify an unknown speaker.

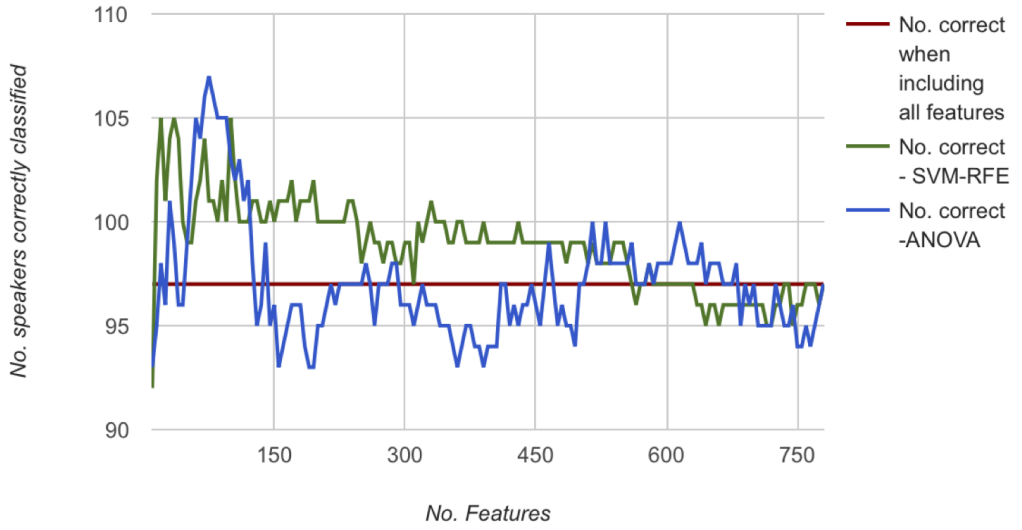

Figure 4.2: The effect of the number of top-ranked features on accent recognition performance by two feature selection methods: ANOVA and SVM-RFE.

The horizontal red line indicates the level at which the system performs when we use all the phonemes available (both vowels and consonants) to act as a baseline to show us whether the feature selection methods benefit performance or degrade performance. This baseline level translates to 80.8% correct.

The main observation we see is that using a smaller number of features improves the performance of the recognition system. The ANOVA method seems to achieve the highest recognition rate overall, where the 80 top-ranked features were used. This achieved a recognition rate of 89.2% correct. This also exceeds the recognition rate we achieve when we implement the original

vowels-only setting, which is 86.7% correct.

Even though ANOVA seems to achieve the highest recognition rate overall, it seems to be the more 'volatile' of the two methods. Firstly, ANOVA seems to be the method with the greater propensity to bring recognition rates below baseline recognition rate. Although SVM-RFE does not quite reach the same recognition rate that ANOVA does, it much more consistently produces recognition rates above the baseline, and the rate appears to more gradually rise as we decrease the number of features used. Additionally, SVM-RFE appears to select better sets of features when fewer than 80 are included. In the case of ANOVA, it appears to peak at 80 features, and then performance rapidly drops. SVM-RFE, on the other hand, still maintains reasonable recognition rates where lower numbers of features are used. This difference in effect on performance could be due to the fact that ANOVA deals with each Y-ACCDIST matrix element independently, whereas SVM-RFE looks at the collection of features as a set and then gradually removes each feature, while observing the effect that the remainder of the set has on performance. However, this consistency in performance comes at a greater computational cost, compared to when ANOVA is used.

Feature Ranking

We can also observe more closely how the specific Y-ACCDIST matrix elements are ranked by each feature selection method. This is shown by the heatmaps below in Figures 4.3 and 4.4, for which all speakers in our dataset have been modelled as Y-ACCDIST matrices and each feature selection method has been applied, resulting in an overall ranking of matrix elements (phoneme-pair distances). For ANOVA, this is based on the p-values generated indepen-

dently for each Y-ACCDIST matrix element. The smaller the p-value, the higher the ranking. For SVM-RFE, the ranking is based on the order in which the individual matrix elements are removed from the set to achieve the best performance overall. The matrix elements that are removed first are those which yield a low (worse) ranking. Only a simple ranking system has been used for each method. It is of course possible to generate a heatmap of all of the ANOVA p-values, so we can observe just how significant our individual matrix elements are. However, for the purposes of comparing the two methods, only a simple ranking system has been used, where the top-ranked element for each method is ranked *1*, and so on.

The following heatmaps form the basis of a Y-ACCDIST matrix. Only the lower triangle is shown because the matrix is symmetric. The darker the cell, the higher that matrix element is ranked.

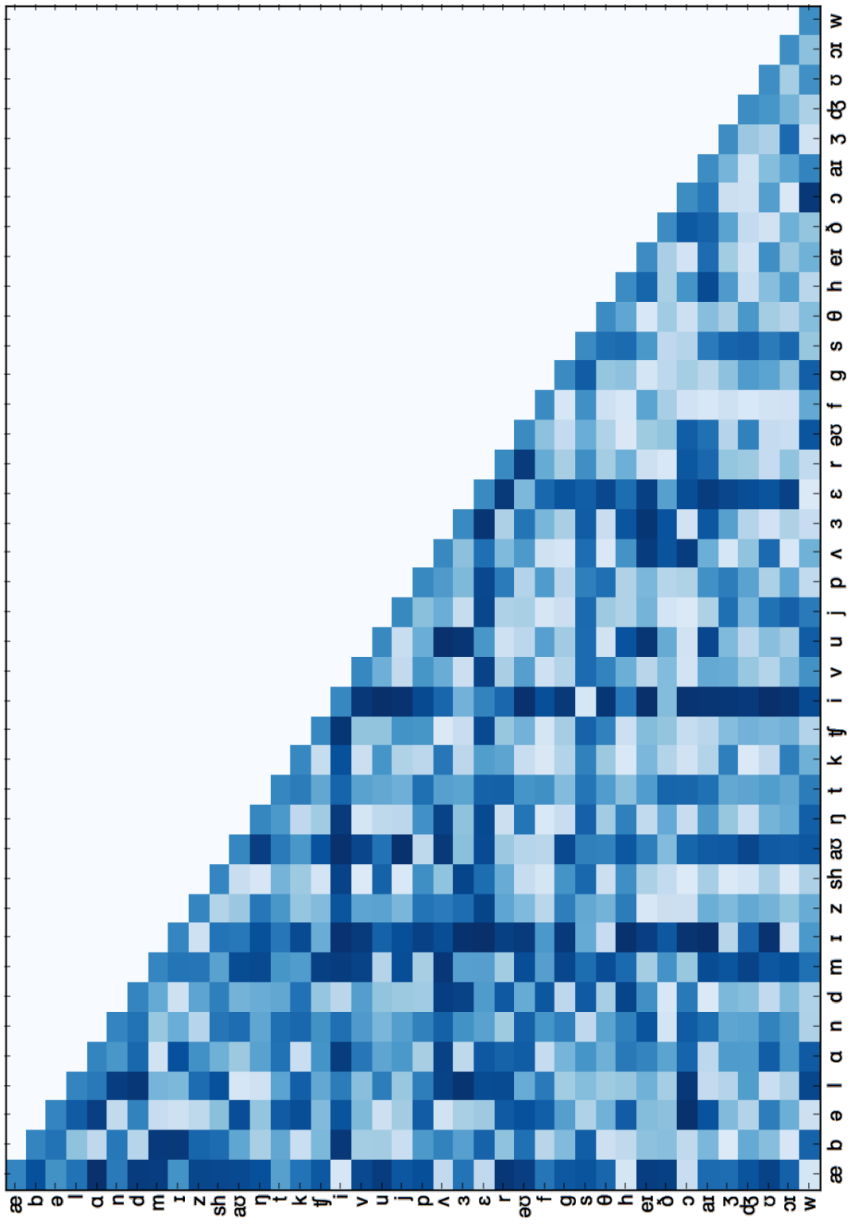

Figure 4.3: Heatmap of the ranking of Y-ACCDIST matrix elements having applied ANOVA as the feature selection method. The darker the matrix element, the more highly ranked that phoneme-pair distance.

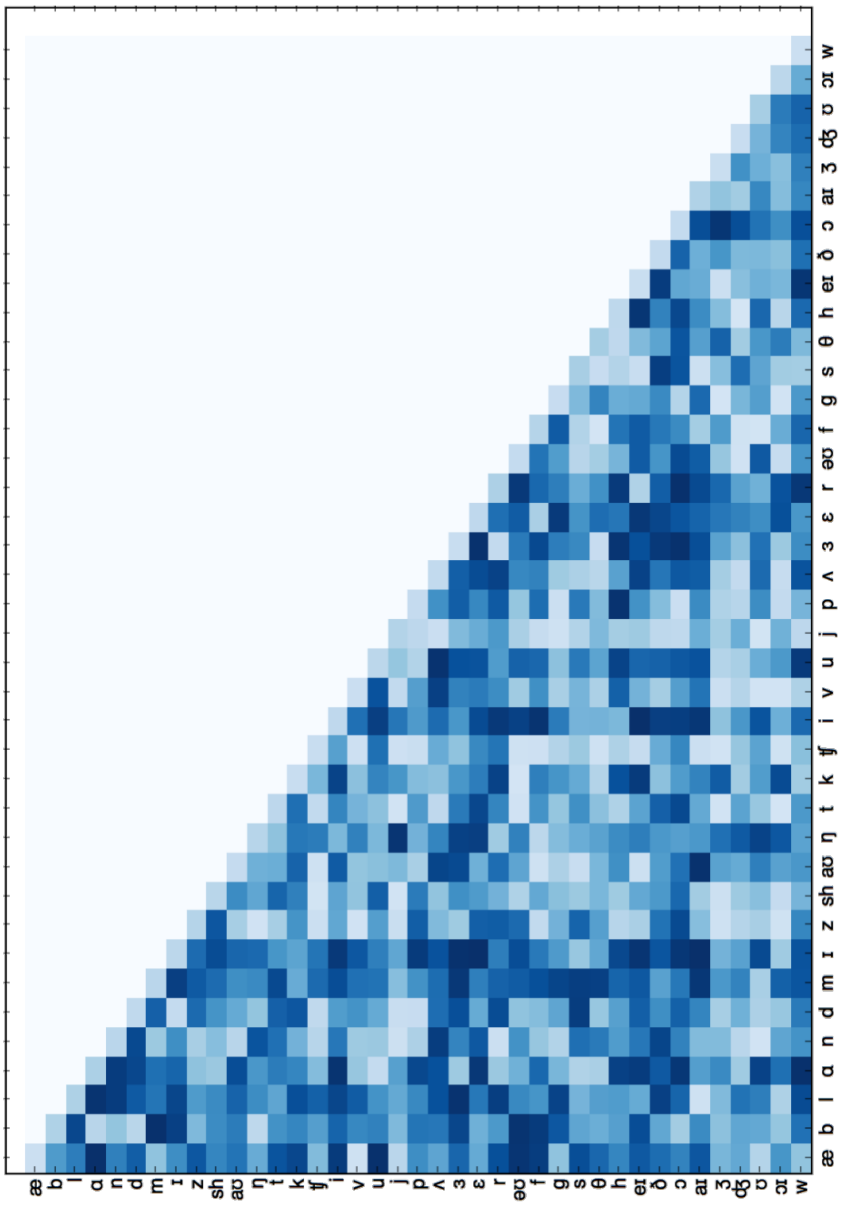

Figure 4.4: Heatmap of the ranking of Y-ACCDIST matrix elements having applied SVM-RFE as the feature selection method. The darker the matrix element, the more highly ranked that phoneme-pair distance.

When comparing these heatmaps, we can observe both similarities and differences. First of all, both feature selection methods rank the same Y-ACCDIST matrix element as the highest-ranking feature. This is the distance between 'iy' and 'ey' (/i/ and /eɪ/ or FLEECE and FACE vowels when referring to Wells' (1982) keywords). It is interesting how the two methods corroborate in this way, suggesting that this phoneme pair distance (or indeed, these two phonemes) are particularly valuable for distinguishing between these specific accent varieties. We might expect /i/-production to vary among these varieties. With reference to the AISEB corpus, Llamas, Watt and MacFarlane (2016) point out that the Scottish Vowel Length Rule (SVLR) is likely to condition different productions of /i/ among AISEB speakers. When SVLR is in place, there are durational differences between /i/ in certain phonological conditions, traditionally among Scottish English speakers. This may therefore be an expected distinguishing feature among these varieties. Even though Y-ACCDIST is not expected to capture durational segmental differences directly (because it only uses midpoint MFCC vectors to represent each segment), there may be quality differences within the segments that are a result of the increase in duration that could be reflected in the models (Lindblom, 1963).

It is interesting to see that a diphthong forms part of this highest-ranking feature. One criticism of the system in its current form is that it does not account for dynamic information within the sound segment. Recall from Chapter 2 (in the description of the Y-ACCDIST-based systems) that these phoneme-pair distances are distances between MFCC vectors extracted from the temporal midpoint of the segment. Not only this, but the MFCC vectors that are extracted exclude any delta or acceleration features that would also have accounted for dynamic information in the sound (i.e. information that characterises diphthongs). However, it appears that even just the midpoint of this

sound is distinctive.

From the heatmap of the ANOVA method (Figure 4.3), we can see that some rows and columns of darker cells are more evident than others. In particular, the 'iy' segment (/i/) seems to be showing this pattern more than any other segment. This suggests that this segment is particularly useful in distinguishing between the four AISEB varieties. The SVM-RFE heatmap (Figure 4.4), however, does not appear to show these kinds of categorical rankings to the same extent. This absence of relatively sharp distinctions is likely to be the result of the different ways in which the two methods work, as was discussed further above. As explained before, ANOVA takes each Y-ACCDIST matrix element independently and calculates a p-value indicating how valuable it is to the task. SVM-RFE, on the other hand, looks at all of the available elements as a set and selects an element to remove, in a one-by-one fashion, based on how the remaining elements work together as a set. This latter strategy removes the likelihood of using combinations with more correlating features (and is probably why we witnessed the effect on performance we saw in Figure 4.2).

Although there are performance advantages to using SVM-RFE, it might be that ANOVA sheds more light on which specific phonemes are good distinguishing features for a given set of accents through exposing these darker rows and columns. This in itself might be a useful feature of the system, especially with forensic applications in mind. Given a set of accents, we can perhaps get an idea in advance of which segments are most useful to the task of distinguishing between them. Forensic casework can require knowledge about any accent variety, and while there is a body of sociophonetic literature on a whole range of linguistic varieties, it is of course very difficult to account for all varieties that might be relevant to casework. This might either be be-

cause a variety has not been researched at all or it might be that a variety has not been researched and documented for a length of time. Due to sound change, we cannot assume that sociophonetic research conducted 20 or more years ago is still relevant today. Conducting feature ranking like this could be a way of efficiently screening a dataset of accents to identify which segments might be key to distinguishing between them. In addition, it is a data-driven and repeatable methodology. These are key properties of a method that are encouraged by forensic science regulators.

Equally, and for a similar purpose, feature selection could be a useful tool for sociophonetic research. A lot of sociophonetic research involves the selection (manually by the researcher) of a small number of linguistic variables to focus on in more detail. This selection could be based on the sociophonetic literature or on auditory judgements. However, a feature ranking process, like the one demonstrated in this section, could provide a way of taking a large number of variables and more objectively assessing which ones might be of interest. The prospect of using Y-ACCDIST as a sociophonetic research tool is discussed in further detail in Chapter 9.

4.3.3 Experiments on the Northern Englishes corpus

These experiments involve the Northern Englishes corpus that was introduced in the discussion of the experiments presented in Chapter 3. There are three key reasons why it might be valuable to run the same experiments on a different corpus:

1. First, it is of interest to observe whether feature selection has the same effect on performance (i.e. recognition rates) across corpora of different sets of accents.

2. Secondly, it would be interesting to see whether the same Y-ACCDIST matrix elements (phoneme-pair distances) are highlighted as being most valuable in distinguishing between a different set of accents. In an accent recognition task of this kind, we perhaps might not expect the same elements to be highlighted. We would expect different matrix elements to be key in distinguishing between different sets of accents (according to the unique combination of characteristics of the accents involved). However, it is important to test this for confirmation. There could be other factors at play, such as the frequency of phonemes. More frequent phonemes in a language might lead to being better distinguishing features because they create more stable phoneme representations within the Y-ACCDIST models. This issue is dealt with more directly in Section 4.3.4.

3. Finally, it is of interest to run these feature selection experiments on a corpus of spontaneous conversational speech. The AISEB corpus provided us with the opportunity to run these experiments on content-controlled data, but for forensic applications, it is vital to observe and compare performance on data that are more relevant to the type of data found in forensic casework.

Like the AISEB experiments above, we will log the recognition rates for each feature selection method, with different numbers of features to include in the analysis being specified (again, in increments of 5 features). 15 speakers per accent group (Manchester, Newcastle and York) will be used in our recognition tasks in a leave-one-out cross-validation setup. We will then observe how the Y-ACCDIST matrix elements have been ranked for these data using heatmaps (Figures 4.7 and 4.8), while comparing them with the heatmaps generated for

the AISEB data.

Effect on Performance

The graph below (Figure 4.5) shows the performance of each feature selection method, combined with the Y-ACCDIST-SVM system, while specifying the number of features used for the analysis in increments of five. This graph is directly comparable with the one generated for the AISEB corpus above in Figure 4.2.

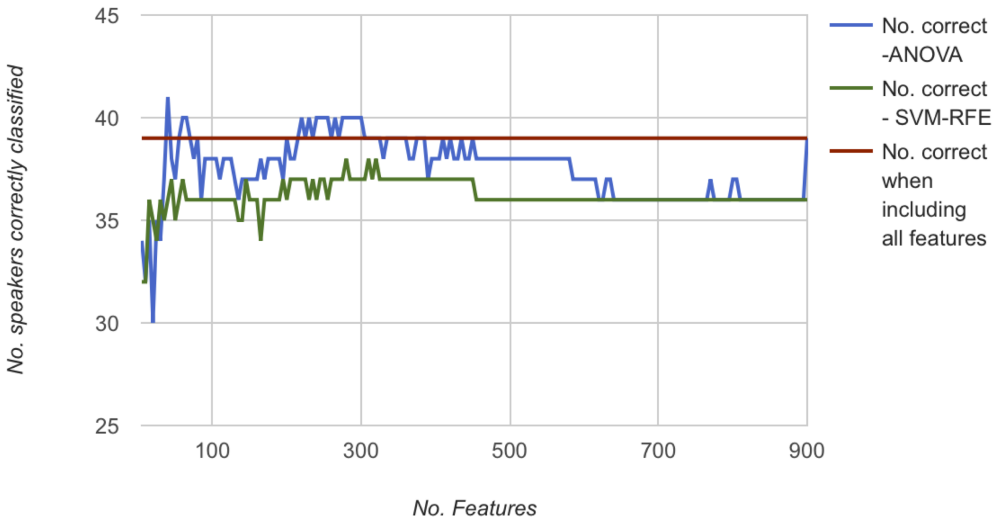

Figure 4.5: The effect of number of top-ranked features on accent recognition performance by two feature selection methods on the Northern Englishes corpus.

We can see a very different effect on performance that the two feature selection methods have on the Northern Englishes corpus. Firstly, baseline performance here equates to 86.7% correct. On the whole, we see that these feature selection

methods have a detrimental effect on baseline performance. Particularly in the case of SVM-RFE, this is likely to be down to the lower number of speakers we are working with from this corpus. Smaller datasets with a high number of features to work with prove problematic to feature selection techniques like this (Saeys, Abeel and de Peer, 2008). However, we also see that ANOVA largely does not have a positive effect on performance, and when it does, it only improves it by a very small margin. As well as the fact that we are using a smaller number of speakers, another reason for this might be because of the nature of these varieties. We might expect that these varieties are more distinct from one another than the varieties in the AISEB corpus. It could be that a larger number of phonemes are useful to the task of distinguishing between these varieties, and in removing these features, we are taking away more discriminative power from the system. In the case of a more similar set of accents (such as AISEB), feature selection is expected to benefit the task because we are more likely to be removing 'noisy' features, rather than useful ones. In a task concerning more distinct accents, we might expect to start closer to the 'ceiling' level of performance and removing features does not make much change to this, whereas for more similar accents, we can expect to start further away from the 'ceiling' level and removing 'noisy' features does make a difference. This demonstrates that feature selection might not always be appropriate for all sets of accents. However, we should remove the factor of dataset size to try to confirm this hypothesis.

In an attempt to better understand the effect of dataset size on the performance of these feature selection methods, we can reduce the number of speakers used for the AISEB experiments and re-run the trials. Figure 4.6 below shows exactly this where 15 speakers per accent in the AISEB corpus have been randomly selected (to match the number of speakers per accent group in

the Northern Englishes corpus).

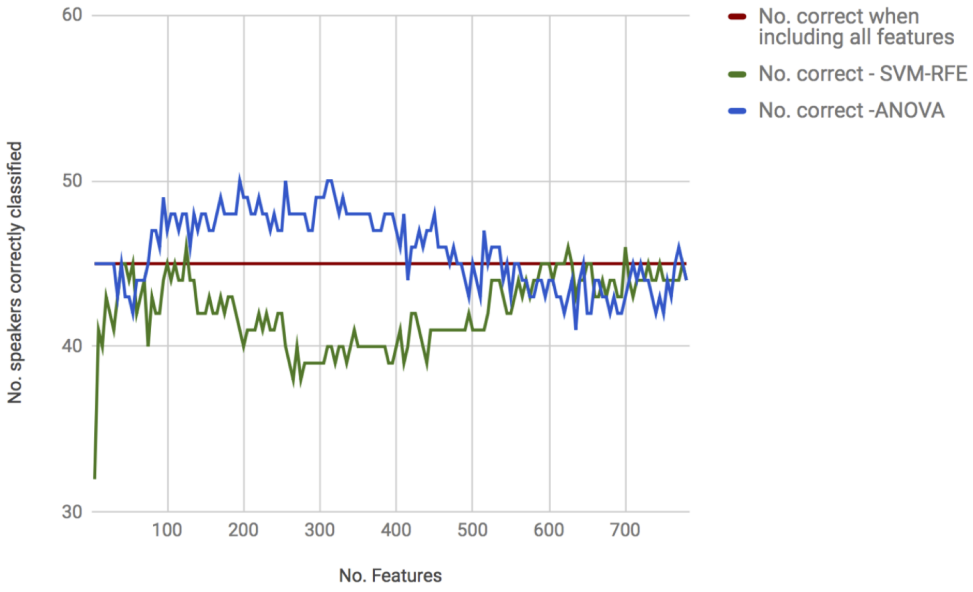

Figure 4.6: The effect of number of top-ranked features on accent recognition performance by two feature selection methods on the AISEB corpus, using 15 speakers per accent group.

This performance graph confirms our suspicions around the effectiveness of the SVM-RFE method on smaller datasets. While it improves performance overall on the AISEB corpus when 30 speakers per group are used, it is detrimental to performance when the number of speakers is halved. However, ANOVA still seems to bring some benefit to the AISEB varieties, despite having a smaller dataset. Because we see ANOVA having more of a benefit on the smaller AISEB dataset that we do on the Northern Englishes dataset, we can gather that the specific varieties involved has an effect on whether feature selection is beneficial to a task or not. It seems that we get around the best performance on the Northern Englishes corpus if we simply include the whole phoneme

inventory in our Y-ACCDIST models. For the AISEB corpus, on the other hand, some selection can bring about our best performance.

Feature Ranking

In the same way we observed how the two feature selection methods rank features on the AISEB data, we can observe how the Y-ACCDIST matrix elements are ranked for the Northern Englishes corpus. It is of interest to discover whether the same matrix elements are ranked in a similar order to those ranked for the AISEB corpus. It is expected that this will not be the case. It is expected that the ranking of elements will be dependent on the specific accents themselves, and the ranking will therefore be unique to the Northern Englishes corpus.

Before we analyse the outputs, there are a few ways in which this is a slightly different task to the one we conducted on the AISEB corpus, and therefore points to keep in mind. These are possible reasons for the differences between the AISEB outputs and the Northern Englishes outputs:

- One thing to keep in mind for the Northern Englishes corpus is the fact that the data are made up of spontaneous conversational speech, unlike the AISEB experiments which used reading passage (content-controlled) data. We can therefore expect more variability in the phoneme representations across speakers.

- Another factor is the fact that we have used a lower number of speakers per group for the Northern Englishes task (15 speakers per group), compared with the AISEB task (30 speakers per group). This could make the outputs less reliable than those generated for AISEB.

- Similarly, we have a smaller number of accent groups for the Northern Englishes task, compared with the AISEB task, and so it is a three-way classification task, rather than a four-way task.

- Finally, slightly different phonesets were used to construct the Y-ACCDIST matrices in the Northern Englishes task, and therefore slightly different phoneme inventories have been represented. While the AISEB task used a phoneset and pronunciation dictionary that aligns more closely to North American English, the Northern Englishes task made use of a Southern British English phoneset. These were the two options that were available during development. An American phoneset was selected for AISEB based on favourable preliminary experiments in the early development stages of the system. The same options were presented when the Y-ACCDIST system was being trained and tested for the Northern Englishes corpus. In this instance, a British English phoneset was shown to outperform a North American English phoneset. This means that the Northern Englishes heatmaps include the /ɪə/ ('iax') and /ɒ/ ('oh') phonemes, while the AISEB heatmaps do not.

Because of these differences between the two accent classification tasks, this is not a direct comparison, but it is still important to discover how feature ranking might differ between two different types of dataset. Below are the feature ranking heatmaps for ANOVA and SVM-RFE for the Northern Englishes accent recognition task.

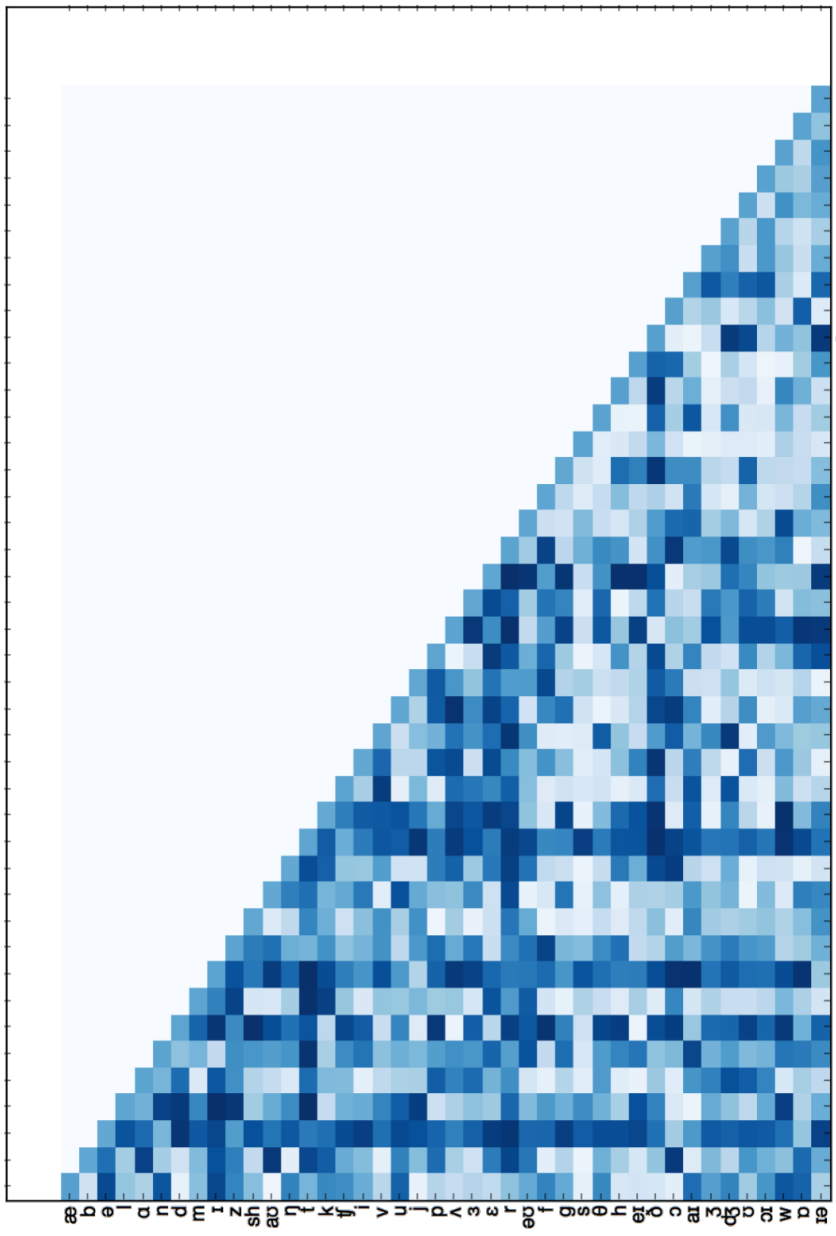

Figure 4.7: Heatmap of the ranking of Y-ACCDIST matrix elements having applied ANOVA as the feature selection method to the task of distinguishing between accents in the Northern Englishes corpus. The darker the matrix element, the more highly ranked that phoneme-pair distance.

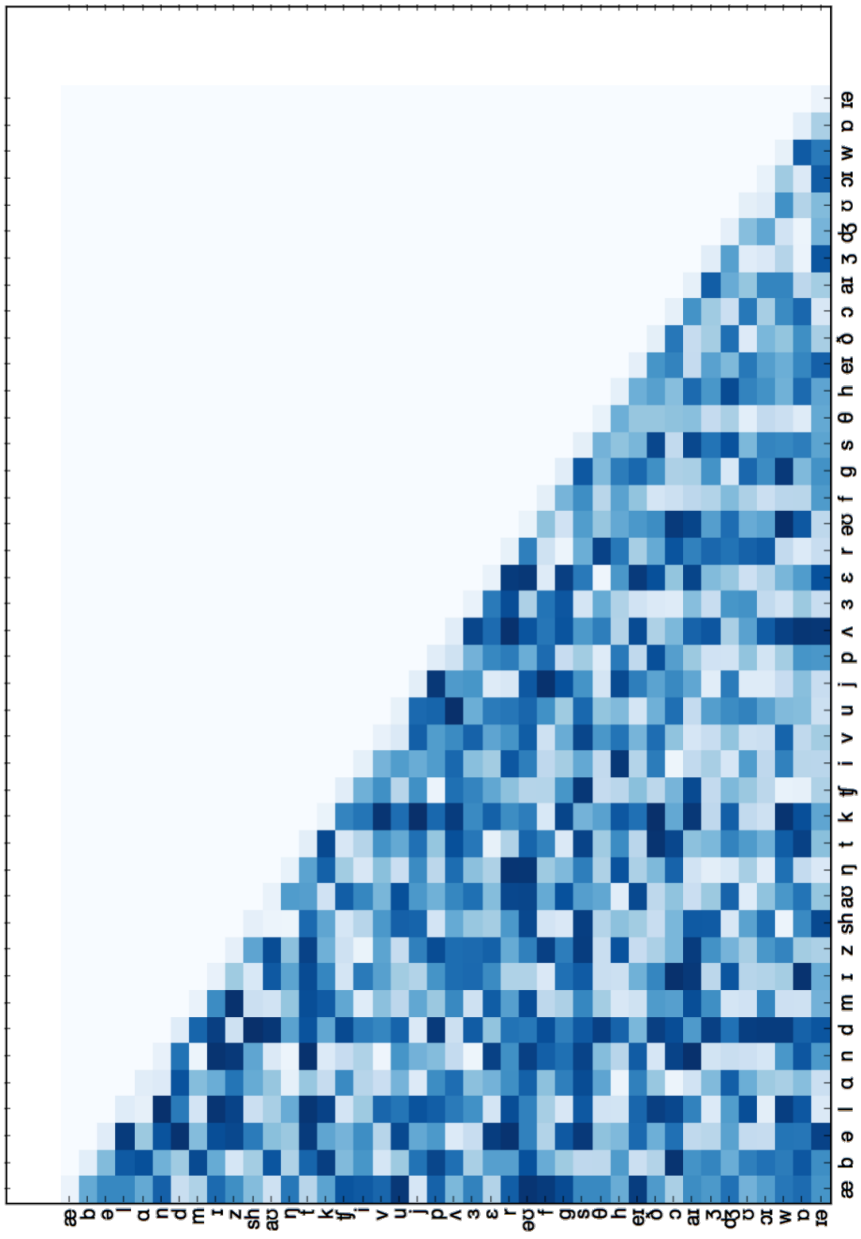

Figure 4.8: Heatmap of the ranking of Y-ACCDIST matrix elements having applied SVM-RFE as the feature selection method to the task of distinguishing between accents in the Northern Englishes corpus. The darker the matrix element, the more highly ranked that phoneme-pair distance.

Again, we see a clearer definition in a smaller number of rows and columns in the ANOVA heatmap, compared with the SVM-RFE heatmap. A reason for this was suggested and explained in the context of the AISEB task above. In contrast to the heatmaps generated for the AISEB task, there appears to be fewer points of corroboration in the case of the Northern Englishes heatmaps. For example, it was noted above that the AISEB heatmaps both computed the same Y-ACCDIST matix element as the highest-ranking element. For the Northern Englishes heatmaps, the highest-ranking element is not the same in each. We witnessed in the performance graph (Figure 4.5) that in the case of the Northern Englishes data SVM-RFE is not such a successful method for feature selection, than what we saw in the case for the AISEB task using more speakers. It is likely to be less effective because of the significant reduction of number of data points (speakers) per accent group in the case of the Northern Englishes corpus. SVM-RFE requires a larger number of data points than what it has been presented with here. What we see in the heatmap showing the performance of the SVM-RFE method is perhaps, therefore, not a meaningful ranking of matrix elements. We should only use SVM-RFE for larger datasets and these outputs seem to confirm this.

When comparing the heatmaps generated by the ANOVA method for each of the corpora (Figures 4.3 and 4.7), we do appear to observe different rankings of elements. To distinguish between the AISEB varieties, we can observe a large number of vowel elements in the higher rankings. For these Northern English varieties, however, this trend does not appear to be repeated. In the case of the AISEB varieties, /i/, /ɛ/, /ae/ and /ɪ/ stand out as some of the more highly ranked segments overall. However, this combination does not seem to be consistent with the Northern Englishes task. In the Northern Englishes task, we might suggest that fewer segments are so clearly highlighted as

potentially valuable phonemes in distinguishing between these varieties. However, /t/ is indicated as the most useful phoneme in distinguishing between these Northern English varieties. /t/ is included in the highest ranked matrix element in this task. For these Northern English varieties, seeing /t/ come through as a key distinctive feature is perhaps unsurprising. This is because Newcastle English typically exhibits glottal reinforcement on its voiceless plosives, distinguishing it from other varieties of English (Docherty and Foulkes, 1999). Other voiceless plosives typically have glottal reinforcement in Newcastle English, but it is perhaps due to the relative frequency of /t/ (i.e. it has a much higher frequency than /p/ and /k/) that has meant that this is shown to be more valuable in this task than the other two voiceless plosives.

As discussed above, a criticism of the Y-ACCDIST-based systems is that they only represent individual phonemes using average midpoint MFCC vectors. By only taking a single midpoint measurement, there is only so much information about that phoneme that is represented. For example, a lot of information that characterises diphthongs is expected to be overlooked. This is also expected to be the case for plosives, like /t/. What characterises plosives are the different stages of plosive production at different temporal points through the sound. It is therefore expected that important distinguishing information would be missed. However, from the ANOVA feature selection heatmap above, it seems that at least some distinguishing information is in fact captured and subsequently taken advantage of within the recognition process.

In addition to /t/, we also seem to observe some value from /ə/ and /ɪ/. Past studies excluded /ə/-based units because it was not expected to be a useful accent discriminator (Huckvale 2004, 1007; Hanani, Russell and Carey, 2013). However, there is reason to believe that /ə/ could indeed carry telling

information about an accent variety. This is more support to use feature selection methods to remove these kinds of assumptions about certain speech segments.

4.3.4 The Effect of Phoneme Frequency on Feature Ranking

The phonemes identified in the section above (/t/ and /ə/) are very frequent in English. In the case of the Northern Englishes task, where we have used natural spontaneous speech, rather than recordings of a controlled reading passage from speakers, we might expect phoneme frequency to play a significant role. The more frequent phonemes are, the more likely that they will provide stable representations in the Y-ACCDIST models. This will subsequently lead to more stable models. As discussed in Chapter 3, natural spontaneous speech means that we find the same phoneme in a number of different contexts which will mean we can expect greater variability in the acoustic representations. We can also speculate that the vowels uncovered as particularly useful in the AISEB task are relatively frequent. It is likely that there is an interaction between frequency and the features highlighted as the most valuable by feature selection. Taking inspiration from the approach seen in Franco-Pedroso and Gonzalez-Rodriguez (2016), this subsection speculates about how the frequency of a phoneme might contribute to how valuable it is in an accent classification task.

The ANOVA feature rankings from this chapter have been used to form a picture of these effects for each corpus. This is because we have concluded that this ranking method seems to expose individual segments more effectively across both corpora. Obviously, because of how Y-ACCDIST modelling works,

we need to find a way of separating out the pairs of phonemes, so we can just observe the performance of individual phonemes. While the following procedure will not provide a high-resolution view of the effect of phoneme frequency, it should offer an approximate indication of whether there is an effect or not. The average ranking is computed for each phoneme, where all the rankings are included, where a single phoneme forms part of the pair. In other words, we calculate the average ranking for a phoneme by taking all the rankings for that phoneme's row and column in the feature ranking heatmaps presented above. We also extract frequency information for each phoneme from the transcriptions of the data. In the case of the AISEB corpus, we count the number of occurrences of each phoneme in the reading passage. In the case of the Northern Englishes corpus, we compute the average number of times a phoneme occurs in a 10-minute speech sample in that corpus. Table 4.1 below presents the Pearson r correlation values that are calculated between the average ranking and frequency variables for each phoneme, and then scatterplots follow to visually support this analysis:

Table 4.1: Pearson r correlation values calculated between phoneme frequency and feature selection ranking.

Corpus	r
AISEB	0.365
Northern Englishes	0.640

The average ranking and frequency counts are plotted against one another in the plots below for each corpus. Figure 4.9 displays the plot for the AISEB corpus and Figure 4.10 displays the plot for the Northern Englishes corpus.

The relevant mappings of the phone symbols to IPA symbols can be found earlier in this thesis in Sections 2.3.3 and 3.2.2. Note that the frequency counts for each corpus (on the y-axis) are working to different scales. The lower the average ranking, the more valuable to the task that segment is estimated to be.

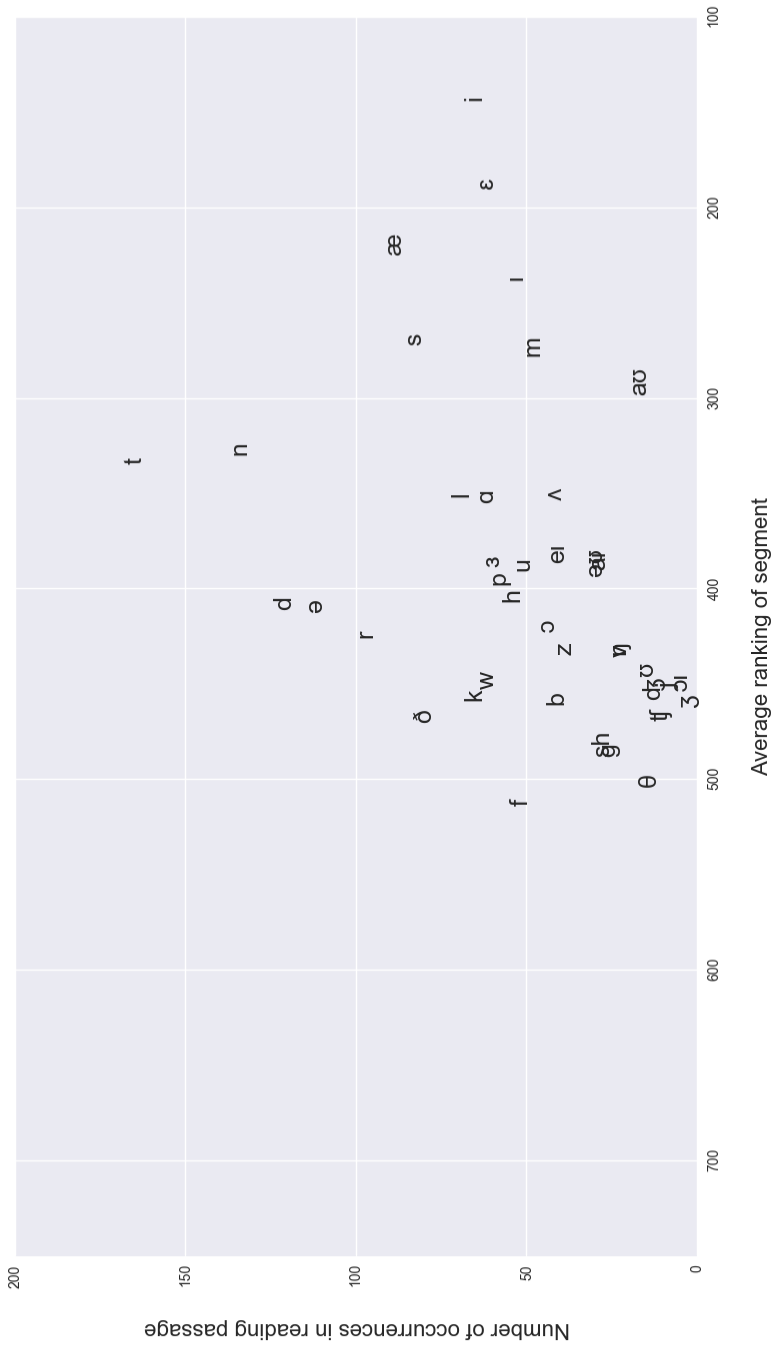

Figure 4.9: Scatterplot showing the effect of phoneme frequency on the ANOVA feature selection ranking of each phoneme for the AISEB data.

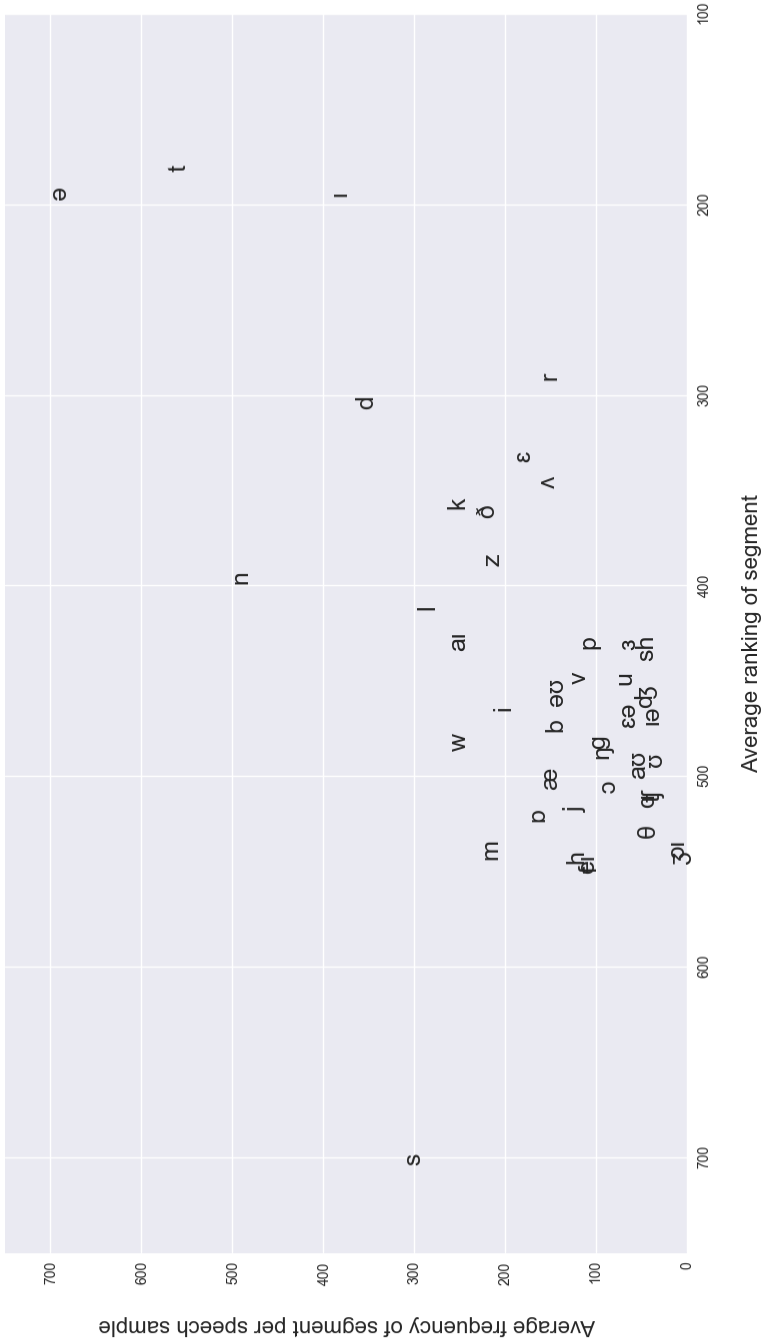

Figure 4.10: Scatterplot showing the effect of phoneme frequency on the ANOVA feature selection ranking of each phoneme for the Northern Englishes data.

In the plots above, we appear to see at least some effect of phoneme frequency for both corpora. However, it seems that a stronger effect (shown by both the correlation values and the scatterplot) is present in the case of the Northern Englishes corpus. This is to be expected because of the reasons given above surrounding the fact that in the Northern Englishes corpus we have spontaneous conversational speech, whereas in the AISEB data we have content-controlled reading passage data, where all speakers are producing the same segments in the same phonological environments (and so reducing the variability in those phoneme representations). In the case of spontaneous speech, the contribution of a single phoneme could therefore more heavily rely on having a higher number of tokens to form a stable phoneme representation.

We can also make some interesting observations about some individual phonemes through these plots. It appears that /s/ is ranked very differently between the two corpora. In the case of the Northern Englishes corpus, it appears to be an anomaly, where it is a reasonably frequent segment, but is ranked as particularly valueless in this task. For the AISEB corpus, however, it appears that it ranks quite highly. It could be that /s/ is particularly variable in different contexts, and so in the case of the Northern Englishes corpus of spontaneous speech, it could actually be detrimental to recognition performance. We also see a different placement of the TRAP vowel (ae) between the two plots. Of course, it is expected that we will see relevant differences between these two plots according to the specific differences among the sets of accents within each corpus. Phoneme frequency is just another factor to keep in mind. Chapter 5 of this thesis deals with the effects of phoneme frequency further from a different perspective.

4.4 Discussion

In the experiments throughout this chapter, we have seen a contrast in the performance of the feature selection methods between the different datasets. We already listed above some points to consider when comparing the outputs for the different datasets in these experiments, such as the lower number of speakers per accent group and the lower number of accent groups in the Northern Englishes corpus. Additionally, the nature of the data itself is likely to have an effect, in that the Northern Englishes data consisted of spontaneous conversational recordings, whereas the AISEB corpus consisted of reading passage recordings (the same one read by all the speakers). However, it is also possible that the nature of the sets of accents is also partly responsible for the differences in performance. We might assume that the overall degree of similarity between the accents in the Northern Englishes corpus is lower than between the accents in the AISEB corpus. If this is the case, we can reasonably expect differences in the outputs from the experiments in this chapter. It is quite possible that the AISEB varieties simply have more features in common than the varieties in the Northern Englishes corpus have. If we have a set of accents where there are simply a lot of features that help to distinguish between them, then the benefits of feature selection are likely to be quite limited. In the case of a set of varieties where there are relatively few discriminating features, feature selection can be expected to be very beneficial as the removal of 'noisy' features becomes more important to optimise performance. We should run further feature selection experiments on a variety of different accent corpora to more thoroughly explore this possibility.

The main reason for trialling feature selection in this work is to see whether

we can improve overall system performance of the Y-ACCDIST-SVM system. In the case of AISEB, feature selection had this effect. However, it would be interesting to look into the recognition rates in much more detail. First of all, it would be interesting to look into which speakers are being misclassified and speculate about why that might be. As an extension to this, it would also be interesting to see whether we get the same speakers misclassified with different numbers and sets of features. In other words, are we necessarily correctly classifying the same speakers we do in the baseline setting, plus additional ones, when we use our optimal number of features for processing? Alternatively, do we actually correctly classify speakers in the baseline condition that are misclassified in the task where we use the optimal number of features? We perhaps should not assume that we simply add more correctly classified speakers to those that we have correctly classified in another setting, but perhaps some correctly classified speakers are lost as we approach optimum settings overall. A more detailed investigation would be required to establish what is happening in this respect.

Y-ACCDIST currently works using features extracted from midpoints of segments. As previously pointed out in this chapter, this could be a criticism of the system because we know that there is dynamic information through sounds that can be indicative of accent (monophthongs vs. diphthongs, for example). In Brown (2014), experiments were presented where more points throughout the segments were included, in an attempt to take advantage of dynamic information. However, this seemed to have a detrimental effect on performance and it was put down to the fact that these additional features brought 'noise' into the models. Future work could look at whether feature selection could overcome this problem and revisit whether dynamic features could in fact contribute to an ACCDIST-based system.

The experiments in this chapter obviously rely on the fact that speakers will exhibit a specific set of features that are produced by typical speakers from the geographical locations. In the making of the two corpora, speakers were selected based on their background, which would have a strong likelihood of them speaking a typical variety of the area. This is the usual nature of corpora that are intended for sociolinguistic research purposes. However, in reality it is common to find speakers that produce a combination of features that typically belong to speakers from different areas. It would be interesting to extend this research to investigate whether we could identify if some features of a speaker's speech would be typical from one variety, whereas other features are more typical of another variety. This could be of value to some strands of sociophonetic research.

4.5 Summary

This chapter has firstly shown the effects that two different feature selection methods, ANOVA and SVM-RFE, have when integrated into the Y-ACCDIST-SVM system. We cannot necessarily conclude that these feature selection methods improve the performance of the Y-ACCDIST-SVM system in every instance. This chapter has demonstrated that the nature of the datasets that are being used appears to have an effect on whether feature selection methods are useful to the task at hand. This is important to keep in mind when considering these kinds of technologies for forensic applications, where different datasets are relevant to different tasks. This chapter has also speculated about individual phonemes and their individual contribution to a given accent classification task, while also observing that there is likely to be a frequency effect involved.

CHAPTER 5

Effects of Segmental Content on Accent Recognition Performance

5.1 Introduction

The purpose of working with spontaneous speech data is to simulate more closely the type of data we might encounter in forensic casework. So far shown in this thesis, the Y-ACCDIST system has been challenged by testing it on geographically-proximal accents, spontaneous speech data and degraded speech data. Within the spontaneous speech data and degraded data conditions, the duration of the speech samples being tested has also been observed as an experimental variable. Of course, it is of interest to determine approximately how much speech is required in an unknown sample for a reliable analysis to take place. Unsurprisingly, the experiments in previous chapters have shown that shorter speech samples are less likely to be correctly classified than longer ones. However, it is hypothesised here that the segmental contents (i.e. the specific vowels and consonants) of the unknown speech sample also

have some bearing on the chances of the sample being classified correctly. In other words, for a given accent classification task to take place, which specific phonemes within the unknown sample will assist with the particular analysis, and how many of them are required? The focus of the present chapter is the content of the test samples, rather than the training data, because we are interested in stepping towards the criteria a speech sample needs to meet to successfully undergo an analysis. Already in this thesis, we have seen that some segments are more valuable than others in distinguishing between accents in a given dataset, through the feature selection results presented in Chapter 4. Feature selection was conducted on the training data to determine which Y-ACCDIST matrix elements (the phoneme-pair distances) are likely to be most valuable. By running a feature selection phase during system training, and only including the elements which have been ranked highly, we can in principle improve system performance. It is therefore not unreasonable to assume that some test samples are more suitable for an analysis, depending on their segmental contents.

Similarly, in automatic speaker recognition research, there have been a number of studies which only include linguistically-defined units in the analysis, rather than including what the whole speech sample has to offer. This has been enabled by the advances in automatic speech recognition technology, so the speech content can be recognised or estimated by a speech recognition system, in order for particular linguistic units to be selected or deselected. A more linguistically-constrained speaker recognition analysis can then take place. Shriberg (2007) talks about this selective, linguistically-constrained, approach as a 'conditioning' process of the speaker recognition models. One speaker recognition study which takes advantage of this is Bocklet and Shriberg (2009), which trials a number of syllable-based constraints on a GMM-UBM

speaker recognition system using NIST Speaker Recognition Evaluation (SRE) 2008 data. They discovered that by only using monosyllabic words they can achieve comparable recognition rates to using all the data available (an EER of 4.4% with telephonic data), while also significantly reducing the amount of data that the system has to process. This could be beneficial in terms of computational cost. Franco-Pedroso and Gonzalez-Rodriguez (2016) provide another study, which makes use of linguistic constraints in speaker recognition technology, where they test many i-vector-based systems that have been conditioned using only tokens of a single phoneme. They did this for each phoneme in a phoneme inventory for American English speech (also using NIST SRE datasets). They took this a step further and conducted the same tests on systems constrained to specific diphones (unique speech units consisting of two phones in sequence). They discovered a very general pattern of systems constrained to more frequent phonemes generating lower equal error rates. They also propose that perhaps the particular linguistic constraints placed on a system are speaker-dependent (i.e. some speakers' samples will be more suited to some linguistic constraints that other speakers' samples for an analysis), but this requires further investigation. Overall, these studies demonstrate that some linguistic units are more useful to a recognition task than others.

While linguistically constraining a system at the training stage is a move to improve overall performance (in a similar way to the feature selection experiments in Chapter 4 of this thesis), it may also be worth considering the linguistic constraints that the unknown (test) speech sample poses for an already-trained and optimised system, and whether this affects the likelihood of a reliable analysis taking place. The effects of the segmental content of test samples is touched upon in automatic speaker recognition research by Hasan, Saeidi, Hansen and van Leeuwen (2013). Their study is largely motivated by

the problem of duration mismatch in speaker recognition, whereby system performance is challenged by short test utterances. They look at this in relation to i-vector speaker recognition systems. One key factor they hold responsible is the phoneme distributions of shorter test utterances. They suggest that it is the coverage of the phoneme inventory represented in a test utterance that affects the quality of the speaker's representation. They show the exponential reduction of the number of unique phonemes as a speech sample's duration decreases. While Hasan *et al* make this important link between the phoneme content of a test sample and its duration, and subsequently speculate about its effect on system performance, their experiments still only concern different durations of speech sample. Their experiments did not explore the direct effect of a sample's phoneme content on performance.

Other work on automatic speaker recognition has looked into the link between the segmental content of test samples and the outcome of its trial. Kahn, Audibert, Rossato and Bonastre (2010) ran experiments on French data to see if the phonetic content of test samples affected speaker recognition performance. They report that they did not find a significant result in relation to this, but they acknowledge that this does not mean that it does not have an effect at all. One criticism of their experiments is that they used a corpus of phonetically balanced read prompts. Controlling data like this does not allow us to see the effects of the natural distribution of speech segments that a language offers. It is desirable to run these kinds of experiments on spontaneous speech, but of course finding or acquiring enough transcribed spontaneous speech data can prevent this kind of valid research from taking place.

It is hypothesised in this chapter that the segmental composition of the test sample is likely to have an effect on its chances of being correctly classified in the context of automatic accent recognition. We will continue to use the

Y-ACCDIST-SVM system to explore this. Since the Y-ACCDIST system takes a text-dependent approach to accent recognition, we can expect that the segmental content of a test sample is particularly important in predicting whether it will be correctly classified. This chapter explores this hypothesis in the context of spontaneous speech. Unlike similar experiments run by Kahn, Audibert, Rossato and Bonastre (2010), this chapter will focus on the natural segmental distributions that spontaneous speech presents. To do this, the Northern Englishes corpus (used for experiments reported in Chapters 3 and 4 of this thesis) has been used as it provides a substantial amount of transcribed spontaneous speech per speaker. However, we will be using shorter samples for the experiments in this chapter, than the samples used in the experiments in previous chapters. As Chapter 3 acknowledged, 10 minutes of net speech per speaker is an unrealistic sample length to expect in forensic casework. Also, two short samples of the same duration are much more likely to have different segmental contents, and these differences might influence the outcome of an analysis. It is these shorter samples that the present chapter is concerned with.

5.1.1 Outline

Section 5.2 will further explain the idea of *segmental content* before this chapter moves on to describe the methodology employed to address the above hypothesis in Section 5.3. Section 5.4 will first analyse the effects of the segmental content of 30-second test samples on accent recognition performance using the Y-ACCDIST system. Section 5.4 will then move on to analyse these same effects on different sample lengths. Section 5.5 will evaluate these experiments.

5.2 Segmental Content of a Speech Sample

The purpose of this section is to further expand and clarify the idea of a speech sample's *segmental content*, which was given and touched on in the title of this chapter and in the section above. Using 30-second test sample durations as an example, we observed earlier in Chapter 3 that speech samples of this shortened length achieve an overall automatic accent recognition performance, with the Y-ACCDIST-SVM system, of 53.3% correct. To further illustrate the idea of *segmental content*, two of the 30-second samples from one of the speakers in the Northern Englishes corpus were selected at random, and number of tokens of each phoneme was counted in each sample. These phoneme frequency counts for each sample form the segmental content, and we can observe the distributions very simply in Figure 5.1 below. For reference, the corresponding IPA symbols to accompany the segmental symbols used can be found earlier in this thesis in Section 3.2.2:

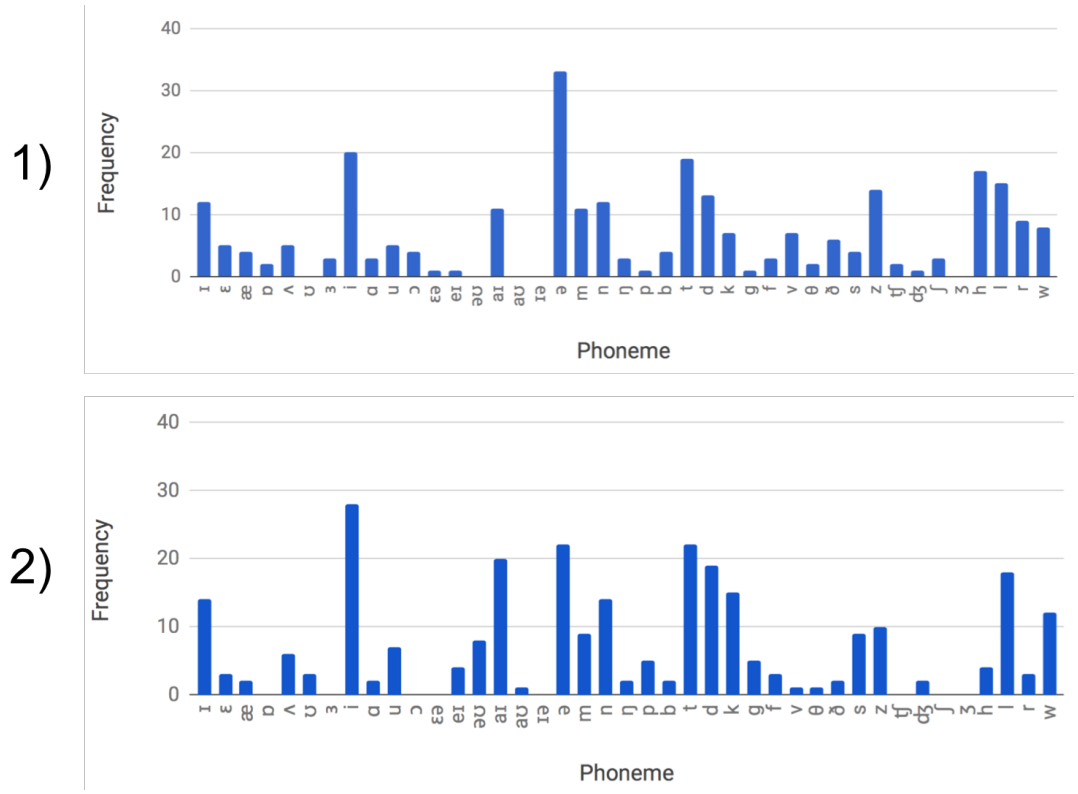

Figure 5.1: The segmental distributions of two randomly selected 30-second speech samples.

The general shapes of these distributions at a first glance look fairly similar, reflecting the natural phonemic distribution of English. However, some differences exist between the two distributions. For example, Sample 1 appears to contain roughly 50% more schwas (indicated by the 'ax' phoneset symbol) than Sample 2. Also, Sample 1 appears to contain three instances of the /ɜː/ vowel (indicated by the 'er' symbol), whereas Sample 2 contains none. We are interested in whether these kinds of differences matter when classifying these samples. Suh and Hansen (2012: 1515), in relation to speaker recognition technology, refer to these gaps of phone coverage in test or enrollment

datasets as 'acoustic holes'. They investigate a way of addressing these acoustic holes when dealing with particularly short speech samples that do not cover the whole of the language's phoneme inventory. We can gather that this is a problem that spans speech technology when short utterances are being used, and is not a problem restricted to accent recognition. It is expected to be of particular relevance to accent recognition, however, because segmental cues can play a very large part in accent diagnosis.

The key question posed by this chapter is, do these segmental differences between speech samples greatly affect the likelihood of the sample's successful accent classification? It is possible that the natural phonemic distribution of English will largely allow for a reliable analysis to take place. The next section outlines the methodology used to try to establish whether the segmental content of a speech sample affects the likelihood of a successful analysis taking place.

5.3 Methodology

The *Language Change in Northern Englishes* corpus (Haddican, Foulkes, Hughes and Richards 2013), which has been used for the experiments reported in previous chapters of this thesis, provides enough transcribed data to allow us to obtain a number of different speech samples per speaker with different natural segmental distributions. The subset of speakers taken from this corpus provides 15 speakers per accent group. The accent groups being used are Manchester, Newcastle and York English. In the same way seen previously in this thesis, 10 minutes of orthographically transcribed net speech is available. For these experiments, the 10 minute stretches for each speaker allow us to generate a number of same-session short samples per speaker. As well as

producing numerous speech samples with different natural segmental distributions, having multiple same-session samples per speaker enables us to observe whether there are speaker identity effects (i.e. are some speakers more classifiable than others?). Speaker identity effects might be down to a number of factors. One might be that some speakers are simply more typical of an accent group than others. Another might be that some speakers are more suitable for processing by Y-ACCDIST. For example, the first step in Y-ACCDIST is forced alignment and some speakers may yield a better segmentation than others. It is possible that these factors mean that performance is largely down to the specific speaker. The setup of this experiment will allow us to speculate about the effect of speaker identity on performance.

Given 10 minutes of transcribed net speech per speaker, we can obtain 20 30-second speech samples per speaker. These were smaller samples were divided by simply concatenating the full 10 minutes available and cutting at 30-second intervals. This did mean that samples were cut mid-utterance. For the 45 speakers, this means that there is a total of 900 30-second speech samples available for testing. The frequency of each phoneme was logged for each sample, capturing the segmental distribution of each sample, before being passed through Y-ACCDIST as a test sample. When each sample was tested, no speech from the same speaker was used to train Y-ACCDIST. The full 10-minute stretches of the rest of the speakers in the dataset were used to train Y-ACCDIST. Whether the test sample was correctly classified or incorrectly classified was logged for that sample, in a binary fashion. This process was conducted for each of the total 900 30-second speech samples. Logging the phoneme frequencies and the success or failure of each sample in this way prepares a results dataset that can be analysed by a mixed-effects logistic regression model.

5.4 Analysis

This section initially looks at the effects of segmental content on the classification of 30-second speech samples. We also consider the effects of speaker identity on classification on these samples. Section 5.4.2 then progresses on to look at these factors on other durations of speech sample.

5.4.1 30-second speech samples

We can observe the successfully and unsuccessfully classified speech samples for each speaker in Figure 5.2 below:

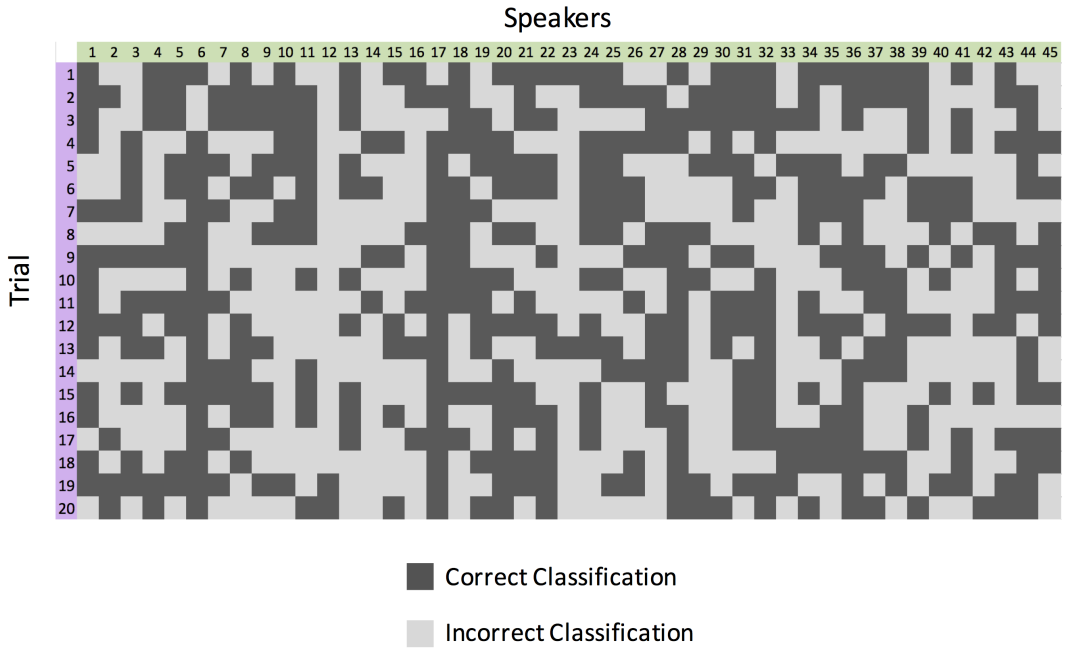

Figure 5.2: Successful and unsuccessful classifications of the 30-second trials for each speaker.

The figure above allows us to rule out that the successful classifications are solely down to the specific speaker being classified. Taking an initial glimpse at the chart shows that there does not seem to be a huge weighting of correct classification assigned to specific speakers (i.e. we do not see entire columns of only correct classifications or only incorrect classifications). They seem to be much more randomly distributed among the different speakers, indicating that other factors play some part in correct and incorrect classifications of 30-second speech samples. However, a closer inspection of the correct classifications does seem to suggest that speaker identity accounts for the outcome of an analysis to some extent. Focussing on speakers 6 and 17, for example, we see that many more of these speakers' trials are successfully classified compared with other speakers. In contrast, speaker 12 seems to be an example of the opposite situation, where the majority of this speaker's trials are incorrectly classified. These observations seem to reinforce our expectations that some speakers are more suitable for this kind of analysis, while others (like speaker 12) seem to be less so.

Even though we can observe some suggestion of individual speaker effects, it seems that there are other factors which separate the successfully classified samples from the unsuccessfully classified ones. As hypothesised in this chapter, it is possible that this is partly down to the segmental contents of the samples. To investigate this, the phoneme frequencies of each sample and its classification outcome (successful or unsuccessful) were analysed by a mixed-effects logistic regression model to assess whether there are certain phonemes (if any) that seem to occur more frequently in the successfully classified samples than they do in the unsuccessfully classified samples.

For the mixed-effects logistic regression model, the frequency of each phoneme was coded as a fixed effect, as well as the true accent group of the speaker.

This is because some accent groups might be more 'distinct' than others (as was discussed in Chapter 2 of this thesis). Further details of the AISEB corpus were outlined above in Chapter 2. One of the four accent groups, the speakers from Gretna, was collectively correctly classified on fewer occasions than the other accent groups in the corpus. This was put down to the fact that Gretna is perhaps expected to be the lowest performing group due to the town's history. Watt, Llamas and Johnson (2014) explain that Gretna is a relatively new town, having been formed in the First World War quite suddenly when settlers came to the area to work at a munitions factory. This sudden and relatively recent formation of Gretna could lead to the possibility that the spoken variety might still contain more variation among the speakers than among speakers of the other varieties, leading to an overall lower accent classification rate as a group. The true accent group of the speaker is therefore also expected to contribute to the outcome of the trial.

Results

From the mixed-effects model (making use of the *lme4* R package), three phonemes were revealed to be significant. These are displayed in Table 5.1:

Table 5.1: The phonemes identified as significant by the mixed-effects logistic regression model.

Phoneme	**Coefficient**	**Standard Error**	**Significance**
ɛ	0.0698018	0.0291341	0.0166
u	0.0944547	0.0429213	0.0278
ə	0.0369923	0.0150073	0.0137

These results indicate that the more of each of these phonemes we find in a 30-second speech sample, the more likely it is that the speech sample will be correctly classified. The fact that any phonemes were flagged up at significant at all suggests that the segmental content does have an effect on the likelihood of the speech sample being successfully analysed by Y-ACCDIST. This is worth bearing in mind when considering automatic accent recognition for forensic applications. Especially for shorter recordings, we might need to think about whether they contain the speech segments that would assist with the analysis. On the basis of these results, it is suggested that some samples are more segmentally suitable for an analysis than others. When we consider the specific phonemes that have been identified as significant in the analysis, we can draw sociolinguistic links with the particular accent varieties we are distinguishing between.

Beginning with the phoneme which was revealed to be most significant, schwa is perhaps expected to be a key distinguisher for these particular varieties. Watt, Llamas, French, Braun and Robertson (2016) show that schwa is a distinguishing feature of varieties of northeastern English English, of which Newcastle is one. Watt *et al* report that Newcastle speakers typically produce

[ɐ], rather than [ə]. Recall also from the description of the Northern Englishes corpus in Chapter 3 (Section 3.2.1) that Hughes, Trudgill and Watt (2012) point out that in Manchester English, the LETTER vowel is produced more like [ʌ] or [ɒ], also deviating from the more standard production of [ə]. From the literature, it is not clear what we might expect of York speakers in the production of these segments, but it is likely that York speakers typically produce [ə]. If these are characteristic features of these varieties as suggested, it is no surprise that schwa, as a phoneme, arises as significant in the results above. Watt *et al*'s results also align with those of Franco-Pedroso and Gonzalez-Rodriguez (2016), who found that, in their i-vector formant-based speaker recognition experiments, schwa can also function as a good individual speaker discriminant. It is expected that schwa's frequency as a segment contributes to these results, as it provides more data to strengthen models and representations, which naturally lead to improved results. We can also refer to the feature selection outputs generated for these Northern Englishes data in Chapter 4 to uncover any signs of corroboration between the two analyses. Recall from Figure 4.10 that schwa was shown to be a very high-ranking segment (indicating its distinctiveness among these specific Northern English accents).

The /u/ vowel can also be explained in this context. GOOSE-fronting is a phenomenon found in accent varieties across the UK. GOOSE is the keyword taken from Wells (1982) to describe the /u/ phoneme in English. It is typically thought of as a close back rounded vowel, but the GOOSE-fronting phenomenon describes its more front realisation by some speakers. Baranowski and Turton (2015: 295) claim that the /u/ vowel has fronted significantly for all social groups in the Manchester area. In contrast, Watt (2000) claims that in Tyneside English, GOOSE-fronting is not evident. Given the difference in the

reported quality of the /u/ vowel in these two varieties, it might be expected that /u/ is a variable which can successfully assign the speakers the correct accent label, particularly for the Manchester and Newcastle speakers.

The /ɛ/ vowel, however, is not necessarily expected to appear among features which distinguish between these varieties. It has not been proposed as a distincive feature of these accents by the sociophonetic literature. One possible explanation for it appearing among the very few significant effects is that it is a good discriminant, but it has not been sufficiently researched by sociophoneticians. An alternative explanation is that it is more to do with the inner workings of Y-ACCDIST, and what it requires to express realisational differences between varieties. In the modelling of speakers' accents, Y-ACCDIST calculates distances between pairs of sounds, rather than treating each phoneme segment individually. While we intend to express individual segmental realisations in this way, we must remember that it is pairs of phonemes which provide the basis for the expression. To be able to express realisational differences, a phoneme must be able to create a reflective distance with another phoneme's representation, treating it like a reference point. We could therefore accept that there might be particularly stable phonemes found among the significant effects in the results above. This is because an analysis might require at least one phoneme which provides stability across all the accent varieties in our corpus for the realisational variation to be sufficiently expressed. In support of this finding, we also see this segment ranked reasonably highly by the feature selection analysis presented in Figure 4.10.

The three phonemes revealed in the results do not, of course, exhaust the list of phonemes which might be expected to assist in an accent recognition task between these three accent varieties. Based on description of the varieties in this dataset given in Chapter 3 (Section 3.2.1), we might also expect to see

the FACE and GOAT vowels, since it was reported that in Manchester English these are diphthongal, while they are mostly monophthongal in Newcastle English. Likewise, we might also expect to see the plosives /p, t, k/ as these are typically glottalised in Newcastle English (again, this was discussed in Section 3.2.1). It should be kept in mind, however, that these are the phonemes which are highlighted when it is 30-second speech samples being tested. If longer speech samples were used, other phonemes might emerge as significant components. We can expect that phoneme frequency plays a large part in these results, and a 30-second stretch of speech might not allow for other expected phonemes to form strong enough representations, because there are simply not enough of them. The number of phone tokens it takes to form a reliable representation of a phoneme's realisation for a speaker may also be phoneme-dependent (i.e. some phonemes might require fewer tokens to produce a reliable average representation than others). One reason for this might be to do with a phoneme occurring in a greater variety of contexts than others, and so a wider range of coarticulatory effects might vary a phoneme's range of realisations.

The Effect of Speaker Identity

In addition to the fixed effects, the model outputted a variance for the random effect of speaker identity (the kind of effect we discussed above in relation to Figure 5.2). The model outputted a variance of 0.398 attributed to speaker identity. If the variance were 0, this would indicate that the specific speaker identity does not contribute to the outcome of a speech sample's analysis. However, the variance outputted suggests that speaker identity does indeed contribute to the outcome, reinforcing initial expectations.

To explore whether these effects are still significant at other durations of speech sample, the section below looks at different sample lengths, varying sample duration to lengths less than 30 seconds and more than 30 seconds.

5.4.2 Varying sample length

This section compares the results of running the same analysis on sample lengths of different durations. This will uncover whether the phonemes revealed as significant in the analysis above are still significant in other sample durations. As discussed above, we can expect phoneme frequency to interact with these results, and sample length obviously affects the number of tokens of a phoneme we find in a sample. The extent to which this affects results is explored in this section.

In addition to the experiments using the 900 30-second samples above, a mixed-effects logistic regression model was run for each of the following sample durations (we also give the total number of speech samples the dataset will allow for the longer durations, considering that there are 10 minutes of speech per speaker):

- 20 seconds (900 samples)
- 25 seconds (900 samples)
- 35 seconds (765 samples)
- 40 seconds (675 samples)

The same model setup was used for each of these sample durations, where the frequencies for each phoneme in the sample are coded as fixed effects, along with the accent category, and speaker identity is coded as a random effect. For

each model (for each sample duration), the significant effects are given in Table 5.2 below. For the sake of easier comparison, the 30-second sample results, which were shown in the section above, have also been included in the table.

Table 5.2: The phonemes identified as significant by the mixed-effects logistic regression model.

Duration Model	**Phoneme**	**Coefficient**	**Std. Error**	**Significance**
20 secs	ɜ	0.155557	0.050521	0.00208
25 secs	n	-0.038775	0.019775	0.0499
	tʃ	-0.128558	0.064732	0.0470
30 secs	ɛ	0.0698018	0.0291341	0.0166
	u	0.0944547	0.0429213	0.0278
	ə	0.0369923	0.0150073	0.0137
35 secs	ɛ	0.055944	0.028402	0.04887
	ɒ	0.083376	0.028982	0.00402
	ɪə	0.177996	0.057549	0.00198
	d	0.037795	0.019263	0.04976
	f	-0.079694	0.031727	0.01201
40 secs	ɛ	0.057348	0.028816	0.0466
	ɜ	0.079076	0.039434	0.0449
	ɛə	0.091353	0.040311	0.0234

At a first look, inconsistencies seem to show up across the different sample durations. Generally speaking, different segments are flagged as significant for different sample lengths. We should view these results with caution. A larger dataset that allows for more samples to be included in the analysis would help to overcome the volatility. It could also be due to focussing on such short sample lengths, where the phoneme distributions might change considerably between different durations. Despite that, it appears that among these identified phonemes, there are some patterns and alignments with what we might expect sociophonetically. Not only can we refer to the sociophonetic literature, but we can also refer to outputs from analysis run in Chapter 4 from the feature selection experiments. Section 4.3.4 of Chapter 4 presents visual outputs that indicate which phonemes were most valuable to two different accent classification tasks through *feature selection*. These outputs plotted each phoneme's overall ranking against its frequency. One of these plots (Figure 4.10) shows this for the Northern Englishes corpus used in the experiments in this chapter. We can therefore check for points of corroboration between the segments that are suggested to be particularly distinctive in Figure 4.10 and those which are identified as significant in the mixed-effects analysis presented in Table 5.2 above. The key difference between them, however, is that Figure 4.10 was produced using all 10 minutes of speech per speaker, while the results presented in Table 5.2 were produced using much shorter speech samples.

One key observation is that across these durations, it is mostly vowel segments that have been identified, and these all have positive coefficients (suggesting that the more of these segments there are, the more likely one of these speech samples will be classified correctly). There are also consonants which have been identified as significant, but with a negative coefficient. This means that having more of these tokens in a speech sample is more likely to lead to an

incorrect classification. Such segments are /n/ and /tʃ/ in the 25-second duration model and /f/ in the 35-second model. The negative coefficient implies that perhaps some segments have a detrimental effect in the accent recognition task. It is inevitable that some segments will not play a role in accent diagnosis, but there appears to be some that go beyond this and mislead the system. It could be that these are segments which are good individual speaker discriminators, rather than accent discriminators. Particularly in the case of /n/, Scheffer *et al* (2011) demonstrated that nasal segments seem to be advantageous in speaker recognition technology, especially under high vocal effort conditions. Seeing that vowels overwhelmingly appear to be significant variables with positive coefficients indicates that the outputs of these analyses are not entirely random. The sociophonetic literature usually focusses on vowels because they are expected to be particularly good discriminating features among accents of English.

Only one segment has been revealed as significant for the 20-second samples, and this is the /ɜː/ vowel (or the NURSE vowel when we refer to Wells' (1982) lexical sets). As with /u/ and schwa in the 30-second samples, we can expect that /ɜː/ would be a valuable segment to an accent recognition analysis. Hughes, Trudgill and Watt (2012: 117) note that /ɜː/ is fronted in Manchester English, which might separate Manchester speakers from Newcastle and York speakers. Interestingly, /ɜː/ does not appear as significant in the analyses for other durations, but then reappears for the longest duration, 40 seconds. This could be linked to the point previously made in the section above with regard to some phonemes requiring more tokens to form a strong enough representation in the model than others. It could be that /ɜː/ does not require many tokens to form a reliable representation in a sample. However, in the longer sample durations, as more phonemes gather more tokens, these might provide

more distinctive power than /ɜː/ in longer durations. This does not mean to say that /ɜː/ does not have anything to offer at all in the longer durations, but has less of an effect when accompanied by other phonemes' stronger representations. Interestingly, when we look at where /ɜː/ is positioned within the feature selection output in Figure 4.10, we can see that there are also signs of this phoneme, showing distinguishing potential as a relatively infrequent segment. /ɜː/ is the highest-ranking segment among the least frequent phonemes (i.e. it is the highest-ranking segment among those phonemes which have fewer than 100 occurrences within a 10-minute speech sample, on average).

Turning our attention to the segments /ɪə/ and /ɛə/, we can assume that similar factors for each of these are at play. Both of these phonemes have schwa as a component, and so a similar effect to the one discussed above in relation to Newcastle speakers' schwa are also likely to apply with these phonemes. In contrast, Hughes, Trudgill and Watt (2012) note that the /ɪə/ and /ɛə/ phonemes are 'smoothed' in Manchester English, and so are realised more as [ɪː] and [ɛː]. It seems that these expected realisational differences might perhaps be influential in distinguishing between these particular varieties when using shorter speech samples.

Although schwa was identified as significant and was justified in the context of 30-second speech samples, it has not emerged as significant in other durations. We can perhaps expect schwa's representation to be sensitive to the addition or removal of tokens in a sample. Even though schwa is a very frequent segment, it appears in many different contexts. We can expect that some contexts help with indicating the speaker's accent, whereas others do not. The tokens of schwa which seemingly do not contribute to correctly classifying a speaker's accent may therefore introduce 'noise' to the representation and so would lose distinctive value as a result. This might explain schwa's

inconsistency as a significant effect, despite being a highly frequent phoneme. This is explored further in Section 5.5 below.

There are of course phonemes which have not been highlighted by the mixed-effects model that we might expect to assist with successful classification. There are diphthongs that are known to vary among these accent varieties. For example, making reference to Wells' (1982) lexical sets, we might predict the vowel in FACE to be significant, but it has not been revealed as particuarly distinctive in this analysis. Likewise, in the feature selection outputs in Chapter 4, the FACE vowel ('ey') was not shown to be ranked highly overall in the case of the Northern Englishes accents (in Figure 4.10). Watt and Milroy (1999: 29) describe the typical variants of this vowel in Newcastle English, stating that it tends to be raised to [ɛi] when preceding voiceless stops or fricatives, and is realised as [ai] elsewhere. Other diphthongs might provide other kinds of realisations which might contribute to distinctive accent models in Y-ACCDIST. This analysis may not be reliable enough to reflect all our expectations that have been fuelled by previous sociophonetic research. An alternative reason is that the frequency of FACE might not be sufficient to establish a stable representation of it in these shorter speech samples. As discussed above, it is quite possible that this is a phoneme that requires a large quantity of tokens to form a reliable representation in a Y-ACCDIST model.

Some consonants are of course also expected to be good indicators in this classification task among these particular accents. For example, Watt and Milroy (1999) describe the /t/ variants that are characteristic of Newcastle English, one being that in pre-pausal position, we tend to get /t/ spirantisation. Another example of typical Newcastle English features they give is the glottal reinforcement of consonants. In some phonological contexts, /p/, /k/ and /t/ are often reinforced by [ʔ] in Newcastle English. /t/ has not sur-

faced as a significant variable in these analysis however. It seems that in the case of these short durations of speech, having these segments included in the unknown speaker's sample does not significantly add to the chances of a successful classification. This could be down to a number of factors, like the frequency of the particular contexts these characteristic segments are found in, or perhaps these segments require a more dynamic representation in the model to be able to sufficiently reflect the speaker's accent. Because the absence of this segment appears to contradict the outputs from feature selection for this data (Figure 4.10), it could well be the factor of frequency at play here, or indeed, it could be a reflection of the reliability of the logistic regression analysis. According to the feature selection outputs, /t/ is the highest-ranked segment that discriminates the Northern Englishes accents. This of course was computed using 10-minute speech samples, rather than shorter samples. It is therefore quite possible that it requires a large number of tokens to form a stable and distinctive representation.

The Effect of Speaker Identity

Individual speaker identity was coded as a random effect in the model, because we can reasonably expect that some speakers are more likely to be successfully classified than others. As discussed above in relation to the 30-second speech samples, speaker identity does indeed account for some of the variance in the model, therefore indicating that speaker identity does contribute to the likelihood of a successful classification ($\sigma^2 = 0.398$).

We can observe this variance across the models for each of the sample durations. In Figure 5.3 below, the variance of speaker identity for each of the duration models is given.

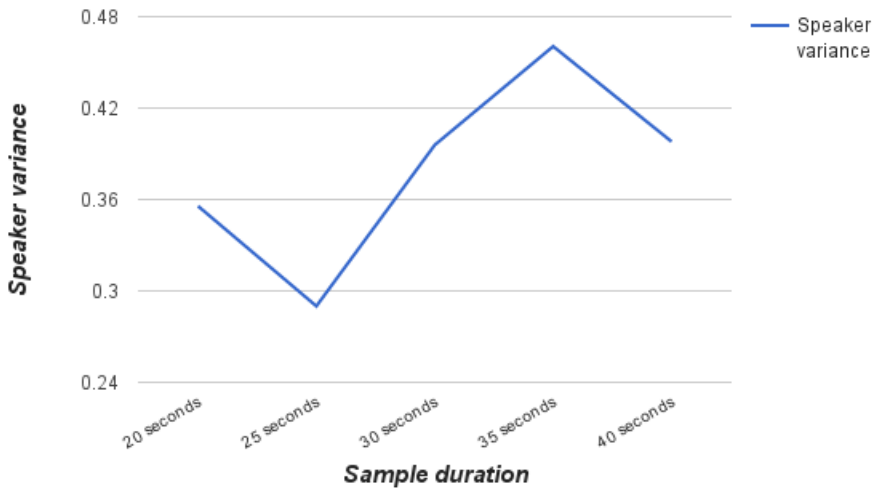

Figure 5.3: Speaker identity variance in each mixed-effects logistic regression model for each speech sample duration condition.

Given that we only have five duration models here, the conclusions we can gather can only be very speculative, but there still seems to be something to consider. It appears that there might be a general increase in the variance in the model assigned to speaker identity. This suggests that perhaps, as the duration of the speech sample increases, specific speaker identity contributes more to the likelihood of a successful classification. More data would be required to test whether this trend continues for longer durations of speech. It is possible that the contribution the specific speaker identity makes to a sample's likelihood of being successfully classified plateaus and stabilises at longer durations, once the individual phoneme segments' representations have stabilised with more tokens.

Similar effects of sample duration on the contribution that specific speaker identity makes to performance also apply to automatic speaker recognition.

Doddington, Liggett, Martin, Przybocki and Reynolds (1998) made steps towards accounting for this by placing labels on different speakers, using statistical analysis, to characterise the "recognizabilty" of speakers in speaker recognition tasks. It might be of interest to investigate how segmental distributions interact with the recognizability of speakers in speaker recognition, and observe whether the label a speaker is assigned changes if the segmental distribution of a speech sample changes. This is another possible direction for further research.

5.5 Discussion

One thing which has been highlighted above, and which is likely to be key to the significance of a phone segment to an analysis, is the context in which the phone occurs. This was suggested earlier in the chapter in Section 5.4.2 in the case of schwa. Watt, Llamas, French, Braun and Robertson (2016) only focussed on word-final schwa tokens in their analysis. We can expect that tokens of a phoneme in particular contexts bear more informative value to the task than others. For example, Watt *et al* (2016) report that word-final schwa has a higher intensity in the Newcastle variety than the preceding stressed vowel in the word. It is context-dependent characteristics like these which might be carrying all of the distinctive value to the task that is encoded in the overall average representation of a single phoneme. Other contexts might present 'noise' to the representation and do nothing for the accent classification task. It might therefore be of interest to conduct further research into the details of the contexts of the phoneme's tokens, particularly for very frequent phonemes like /ə/. It is plausible to run a similar analysis that uses more specific categories of speech segments, rather than the broader phoneme

categories enforced here.

Another factor that could play a part, which has been accounted for indirectly in the analysis, but has not been separated out on its own, is articulation rate. It could be that the articulation rate of individual speakers affects the likelihood of a speech sample being correctly classified by the Y-ACCDIST system. Articulation rate of a speaker has been indirectly encoded in the model because speaker ID was coded as a random effect, and a speaker's articulation rate is embedded within this factor. One hypothesis could be that a higher articulation rate leads to a higher likelihood of that speaker's samples being correctly classified. This is because a 30-second speech sample would include more phone segments to produce more stable representations in the model. A deeper investigation into the data would uncover whether this is the case.

To its practical disadvantage, Y-ACCDIST is a text-dependent system, which enables it to take a very segmental approach to the task of accent recognition. We can therefore expect that the phone segments that are present in the unknown speaker's speech sample are especially important to the analysis. It would also be interesting to run these experiments on a different type of accent recognition system (a text-independent i-vector classification system, for example). These kinds of system take a more global approach to a classification task, not making initial phone segmentations, but take the sample as a whole. It would be interesting to run the same experiments described in this chapter using an i-vector accent recognition system, rather than Y-ACCDIST. It might be the case that the segmental content of a sample also has an effect on the performance of an i-vector system.

In a similar way to different types of accent recognition system, it would also be of interest to run these sorts of experiments, testing the effects of a speech sample's segmental content, on speaker recognition systems. For

forensic applications, it could be useful to determine whether a speech sample is suitable for an analysis by a particular type of system, or to at least bear in mind the factors contributing to a reliable or successful analysis. In the context of security and commercial automatic speaker recognition systems, it is also expected that certain passphrases yield higher success rates than others, and this is partly down to the phonemic contents of these passphrases. Across applications, it seems it would be valuable to move towards establishing the segmental criteria a sample needs to meet for a reliable analysis to take place.

5.6 Summary

This chapter has more directly investigated the relationship between the phoneme frequencies of a speech sample and the likelihood of a successful classification by the Y-ACCDIST-SVM system. Although we should cautiously interpret the outputs of these analyses, it appears that the segmental content of a speech sample influences a speech sample's probability of being correctly classified. Of most of those phonemes that were identified as significant, we can generally offer a sociophonetic explanation for why they are significant when distinguishing between speakers of Manchester, Newcastle and York varieties of English. Despite that, other groups of phonemes which would be expected to contribute to a sample's successful classification were not identified as significant by the model. Possible reasons surrounding Y-ACCDIST's inner workings and the frequencies of these phonemes were offered for this. Importantly, however, there are plenty of other factors that of course contribute to the success of a sample's classification. One of these that has been discussed is individual speaker identity (i.e. some speakers are more classifiable than others), a factor which was accounted for in the logistic regression analysis. This chapter has

not, however, been able to provide specific and fine-tuned criteria that a test sample needs to meet to increase the probability of its successful classification. Instead, this chapter has shown that the segmental content of an unknown speaker's speech sample should be factored in when it is being analysed in an accent recognition task.

CHAPTER 6

Using Y-ACCDIST to classify non-native varieties of English

6.1 Introduction

All the experiments so far in this thesis make use of accent corpora of native varieties of English, starting with native varieties which are expected to be fairly similar to one another (the AISEB corpus). We then moved on to the Northern Englishes corpus, another corpus of native accents that allowed us to test Y-ACCDIST on spontaneous conversational speech. This chapter extends this research by testing Y-ACCDIST on spontaneous conversational speech produced by non-native speakers of English. The objective of the system is to classify speakers according to their native language We can expect this to be a different kind of problem because of a number of additional factors that come into play in non-native speech. These kinds of factors will be described in Section 6.2. Moving on to a different database in this chapter continues with one of the key objectives of this thesis of transferring an

analytical methodology to different types of task. This is in the interest of gaining a thorough understanding of the methodology's capabilities. As has been repeated throughout this thesis, past ACCDIST-based systems have all only been tested on a single corpus, the Accents of the British Isles (ABI) corpus, which consists of good-quality recordings of native speakers of accents of the British Isles reading prepared prompts. This chapter will again observe an ACCDIST-based system's performance on a corpus which presents yet more kinds of challenges.

The data used to test the performance of Y-ACCDIST on non-native accent varieties are from the *National Institute of Standards and Technology Speaker Recognition Evalulation* (NIST SRE) 2004, 2005 and 2006 datasets (Przybocki and Martin, 2004). These datasets are primarily intended for automatic speaker recognition experiments, but metadata are available that make it possible to conduct other kinds of task. These databases are largely made up of telephonic speech, and the subset used for the experiments in this chapter is just made up of telephonic speech, and so constitute a more forensically realistic scenario, compared to experiments in previous chapters. More details about the subset of the NIST SRE data used for these experiments will be given below, but one important feature of this task to note is that we are only using phone labels of the data that have been estimated by an automatic speech recognition system. In a sense, then, this chapter not only observes Y-ACCDIST's performance on non-native speech, but also observes whether it is capable of working without human-generated transcriptions (which are the kind of transcription that have been used for all the experiments in the chapters above). In effect, we are therefore also testing whether it could be used as a sort of text-independent tool (for a faster less labour-intensive accent recognition analysis). This could have repercussions for the Y-ACCDIST

system's overall usability if it can indeed cope with estimated phone labels, rather than accurate human-generated phone labels. This difference in the treatment of data will be considered when analysing the results.

One significant advantage of the NIST SRE subset being used for the experiments in this chapter is that it is much larger than any of the corpora used in previous chapters. While the subsets of the AISEB corpus and Northern Englishes corpus contained 120 and 45 speakers, respectively, the NIST SRE subset contains 700 speakers (100 per accent group). This volume of data means that we can much more reliably assess the effects that small changes to the system have on performance. Consequently, this chapter also presents results where modifications have been made to the engineering of the system.

6.1.1 Outline

This chapter first reviews past research on non-native speech in Section 6.2 before discussing past research that has incorporated estimated phone labels from an automatic speech recognition system in Section 6.3. The purpose of including sections on these two topics is to consider the expected effects of these specific data properties on the Y-ACCDIST-SVM system. Section 6.4 will then outline the details of the experiments run, including a description of the speech data, as well as further information on the estimated phone labels from the speech recognition system. The specific experiments, along with their results, will then be presented in two parts. The first part (Section 6.4.4) deals with the baseline experiments, along with some segmental alternations, varying which speech segments are included in the accent modelling. The second part (Section 6.4.5) presents the effects of making modifications to the engineering of the current default configuration of the Y-ACCDIST-SVM

system. An discussion this chapter's findings is then given in Section 6.5.

6.2 Non-native accents

Numerous factors contribute to how exactly a speaker's non-native accent is produced. In their review of the literature on the factors affecting a speaker's foreign accent, Piske, MacKay and Flege (2001) list the following (among others):

1) Age of the speaker when they learn the language

On the basis of a good deal of linguistic research (as well as everyday observation), we can assume that the younger a language learner is, the better the command of a language the speaker will go on to have (e.g. Asher and García, 1969). This links with the idea of the so-called 'critical period' in language learning in relation to first language acquisition whereby we can only acquire our first language to native level if we do so as children (Lenneberg, 1967), but this idea also seems to extend, to some degree, to second language acquisition where we do see younger learners acquiring a higher standard of a second language than older learners (Johnson and Newport, 1989). Focussing on just the second language speaker's pronunciation, we could generalise that a speaker who started learning at an earlier stage in life is more likely to have an accent that is closer to a native accent of the second language.

2) The length of time a speaker has resided in a place where the L2 is spoken

It is expected that the amount of time that a speaker has lived in a place where the language they are learning is spoken contributes to how speech is

produced. Flege, Yeni-Komshian and Liu(1999) conducted a study of Korean learners of English, using 'age of arrival' (to the United States) as a key variable of interest. They used native listeners to rate the accents of these Korean speakers speaking English, finding an effect of 'age of arrival' on the rated 'strength' of the foreign accent. This relationship also links to the point above in that, as well as the learner's amount of exposure to the language, simply the age at which a speaker starts learning also has an impact.

3) The motivation of the speaker

Of course, the motivation of the speaker to reach a high standard in a language is also expected to affect the overall production of speech. Gardner (2007) discusses different types of motivation in learning a language and what roles these might play in acquiring a language. Gardner concludes that it is generally the intensity of the motivation of a speaker that is the key contributing factor to achievement in second language acquisition.

4) The aptitude of the speaker to learn a language

There is individual variation among members of the language learning community. Some individuals are naturally more suited to learning languages than others, a fact which is regularly acknowledged in the second language learning research literature (e.g. Rubin, 1975). However, it is of course difficult to determine exactly what it is that some of the more successful language learners have that others do not.

The list offered by Piske *et al.* does not, by any means, account for all the complexities that non-native varieties of a language may exhibit. Aspects to do with whether a speaker was formally educated in a language, and how far

through the education system the speaker took this language learning. There are other issues associated with a speaker's identity that could also come into play. In contrast to how speakers might be motivated to reach a high standard of a particular language, Drummond (2012: 110) points out that some speakers may intentionally avoid adopting pronunciation features of native speakers so as to express their own native identity. Connected to this, the statistical analysis run in Drummond (2012) revealed a significant effect regarding non-native speakers' future plans. In analysing the ING variable in English, produced by Polish migrants living in Manchester, UK, Drummond found that if the Polish individuals were planning to return to Poland in the future, it was more likely that they would produce [ɪŋk] for this word ending, rather than other native forms, principally [ɪn] and [ɪŋg], in Manchester English. Through this one variable, there seems to be an indication that even a speaker's future intentions could also impact on the pronunciation patterns in a second language.

Finally, and what the performance of a system in this task relies on, is the effect of the first language (L1) of a speaker on the acquisition of a second language (L2). For a good recognition rate, we would hope that features from the first language are distinctive and consistent enough to be modelled by the accent recognition system. However, within the second-language acquisition research literature, there has been some attention paid to how the L1 of a speaker affects how well the speaker adopts the native features of the L2. For example, McAllister, Flege and Piske (2002) monitor the production of native Estonian, English and Spanish speakers producing Swedish words. In Swedish, there are 'quantity' distinctions that distinguish between different phonemes. For example, the phone sequence [vɛːg] in Swedish means road, whereas the sequence [vɛgː] means wall, with only changes in the durations of different sounds, rather than finding quality differences in the sounds (Helgason, Ringen

and Suomi, 2013). McAllister, Flege and Piske (2002) found that the Estonian speakers acquired this distinction much more readily than the English and Spanish speakers. Estonian has durational distinctions like Swedish, whereas English and Spanish do not. These kinds of features of a speaker's L1 (which increase the similarity of a speaker's pronunciation of the L2 with that of native speakers of the L2) could increase the likelihood of confusion between an Estonian speaker speaking Swedish and a native speaker of Swedish. These kinds of relationships between different L1s could help us to predict confusions by an automatic accent recognition system.

The factors listed and discussed above demonstrate the extent of the complexity among the non-native speaker population. We could assume that these additional factors that come into play for non-native accents might make modelling these accents in a classification system more difficult. We have reason to expect that there is greater variation within these groups (and therefore greater variation among the accent models), which could lead to more confusions, and we might also expect greater within-speaker variation as well. It is reasonable to expect that for a non-native task, we would need more speakers per group to sufficiently train an accent recognition system than we would for a task using native accents.

This section has touched on just some of the considerations we ought to take into account when observing system performance across native accent recognition tasks and non-native accent recognition tasks. There are different kinds of factors that might come into play when we are training and testing. Specifically, we might expect much more variation in the pronunciation systems of the non-native speakers, which might lead to weaker Y-ACCDIST models. However, larger sets of speakers per accent to train the system might help to counteract this expectation.

6.3 Phone estimation through speech recognition

Chapter 2 discussed some of the earlier studies in automatic language recognition which took the Phone Recognition followed by Language Modelling (PRLM) approach to the task (Zissman, 1996). Such an approach was also tried in automatic accent recognition (e.g. Biadsy, Soltau, Mangu, Navratil and Hirschberg, 2010). Using the outputs of an automatic speech recognition system, in the form of phone labels, the system's following processes rely on these outputs. It is important to remember that these labels are simply estimations and very much depend on the quality of the specific speech recognition system being used. Most of the labels are correct, but it is of course important to remember that there are errors among them. The errors, however, are likely to be meaningful in some way. For example, we can expect that an /n/ is more likely to be mistaken for another nasal segment, rather than for a voiceless plosive. A segment that one segment is confused with by a system, is likely to share acoustic properties. These kinds of confusions might therefore not cause so much 'noise' in the system, as we may end up with a category of nasals which might actually function as a label class that can sufficiently serve the system's subsequent processes. The exact details of the automatic speech recognition system used for the experiments in this chapter are given in Section 6.4.2 below.

For these past PRLM systems, it is just the sequence of labels that was important for the system's subsequent processes. In the context of Y-ACCDIST, the system will rely on these estimated labels, but also the accompanying time alignments. Unlike the original PRLM language identification systems

(Zissman, 1996), Y-ACCDIST extracts acoustic values from the labelled segments to represent them, and so not only is the sequence of phone labels for each sample important, but also where each one begins and ends (i.e. the segmentation error). This segmentation consideration also applies to the text-dependent experiments in the chapters above in relation to the outputs of the forced aligner. While the phone labels are expected to be accurate, the time alignments to accompany these labels are not necessarily precise.

6.4 Experiments

This section lays down the components of the non-native accent recognition experiments. The dataset is first described in detail, followed by the nature of the estimated transcriptions. These are aspects of the experiments that set them apart from the experiments in previous chapters.

6.4.1 The Data

Only a subset of the NIST databases from the 2004, 2005 and 2006 speaker recognition evaluations (Przybocki and Martin, 2004) has been used. Together, these databases are comprised of thousands of speakers, but we do not have estimated phone labels for all of them. Only a subset were passed through an automatic speech recogniser. What has reduced the potential pool of speakers further is the number of speakers within each accent class, who are speaking English. Not all speech samples in NIST are in English. Other languages are included, so these recordings are eliminated for the purpose of this study. Additionally, of the recordings where speakers are speaking English, there are great imbalances between the number of speakers within different accent

groups. By far, the group with the largest number of speakers is made up of native English speakers, and there are large numbers of native speakers of Spanish and Russian, but then very few speakers of Vietnamese, for example. From the different accent groups available, seven groups were selected, based on the fact that these categories allowed for groups of 100 speakers to be used for these experiments. The resulting accent groups for these classification experiments were therefore:

- Arabic
- Bengali
- Mandarin Chinese
- native English
- Farsi
- Russian
- Spanish

We do not have any more detailed information about the speakers' language background. For example, the Arabic speakers could be native speakers of any Arabic dialect. Also, having listened to a small number of the recordings in the native English category, it seems that most of the speakers are speaking a North American variety of English (as expected), but there are also recordings of British English speech, too. In experiments using the AISEB corpus and Northern Englishes corpus in previous chapters, the speaker groups were much more controlled. We can expect this lack of control in the collection of the NIST data to also contribute to the weakening of the accent models, alongside

the factors that come from using second-language data (discussed in the section above).

The NIST speech samples are telephone recordings. Each speaker's sample is part of a two-way casual conversation with another speaker which lasts for approximately 5 minutes in total. We can therefore expect that each individual speaker's speech sample contains around 2-2.5 minutes of speech for that speaker (although we should acknowledge that there could well be great variation in duration per speaker). In Chapter 3, we saw some results of experiments where we tested Y-ACCDIST-SVM on different durations of conversational speech recordings. Two minutes of net speech per speaker did not secure the best performance, but it still generated a recognition rate of over 82.2% correct on the three-way classification task of Northern English English accents. This result, however, was produced using good-quality recordings. In the context of the NIST telephone recordings used in this chapter, we might expect some reduction in performance.

Having listened to a small number of these recordings, it seems that it is common for recordings in this database to come with some background noise (for example, other people talking in the background). This sort of feature of course resembles much more realistic recordings, but it is another feature of the data we should consider when comparing these results with those produced in previous chapters.

6.4.2 Phone Recognition

The estimated labels are those that were used for the experiments in Franco-Pedroso and Gonzalez-Rodriguez (2016). These labels were produced by the *Decipher* automatic speech recognition system by researchers at SRI (Ka-

jarekar *et al*, 2009). It is reported that this automatic speech recognition system's Word Error Rate (WER) is comparable to other state-of-the-art speech recognition systems, achieving a WER of 23.0% for native English and 36.1% for non-native English. At first, these WERs suggest that the resulting transcriptions we are using contain numerous errors. However, WER is not necessarily reflective of a "phone error rate", where a word in the transcription can be logged as incorrect where only one phone in the whole word might be incorrect. We can assume that the "phone error rate" is better than the WER and this is the rate that is important for these Y-ACCDIST experiments. However, as a thorough analysis of the transcriptions in relation to this has not been conducted, the precise accuracy of the phone labels remains unknown.

When testing Y-ACCDIST, it is also important to point out that the set of phone symbols used for these NIST experiments is different from those used in the experiments with the AISEB corpus and the Northern Englishes corpus. While there is a lot of overlap with the phonesets that have already been implemented in this thesis, there are also some differences. The phoneset used for these NIST experiments is given and interpreted in Tables 6.1 and 6.2 below:

Table 6.1: The vowel phoneset symbols used for the NIST experiments alongside their corresponding IPA symbols.

Phoneset Vowel Symbol	AE	AA	AO	IY	UW	EH	IH	UH
IPA Symbol	æ	ɑ	ɔ	i	u	ɛ	ɪ	ʊ
Phoneset Vowel Symbol	AX	EY	AY	OW	AW	ER	AH	OY
IPA Symbol	ə	eɪ	aɪ	əʊ	aʊ	ɜ	ʌ	ɔɪ

Table 6.2: The consonant phoneset symbols used for the NIST experiments alongside their corresponding IPA symbols.

Phoneset Consonant Symbol	P	T	K	B	D	G	CH	JH
IPA Symbol	p	t	k	b	d	g	tʃ	dʒ
Phoneset Consonant Symbol	F	V	S	Z	TH	DH	SH	HH
IPA Symbol	f	v	s	z	θ	ð	ʃ	h
Phoneset Consonant Symbol	L	R	W	Y	M	N	NG	DX
IPA Symbol	l	ɹ	w	j	m	n	ŋ	ɾ

This phoneset is almost identical to the phoneset used for the AISEB experiments on English and Scottish accents. However, the key difference here is the 'DX' symbol which maps on to the alveolar tap, [ɾ], which is one feature of North American English, where it is an allophonic variant of /t/ and /d/ that occurs in intervocalic contexts between a stressed and an unstressed syllable (Herd, Jongman and Sereno, 2010). Exemplar words are *writer* and *rider* (Zue and Laferriere, 1979). Because we are dealing with North American English as the native accent variety among this list, 'DX' has been included in the

phoneset.

6.4.3 Experimental Setup

A *leave-one-out* cross-validation setup has been implemented for these experiments, like the experiments presented in Chapter 3 on the Northern Englishes corpus. This setup allows us to maximise the amount of data available to train the system.

We can divide the experiments into two streams. The first stream (Section 6.4.4) is concerned with the the segmental configurations of the system (i.e. which phones are included in the construction of the Y-ACCDIST matrices). The second stream (Section 6.4.5) takes advantage of the the considerably larger size of the dataset we are using for these experiments. This means we can make changes to aspects of the system's engineering, and that changes to the recognition rate are more likely to be an effect of engineering changes, rather than the consequence of smaller datasets naturally prompting larger changes to the recognition rate by just a single speaker being misclassified.

6.4.4 Segmental Experiments

From a comparison of the experiments run on the AISEB corpus and the Northern Englishes corpus earlier in this thesis, we saw a difference in what kinds of phoneme segments generated the highest recognition rate. We noted in previous chapters that in the case of the AISEB corpus, including only vowel segments is more effective than including consonants as well. For the Northern Englishes accents on the other hand, including consonants seems to generate a higher recognition rate than by just including vowels. Because of this difference between the corpora, the initial experiments on the NIST data

will first compare Y-ACCDIST-SVM's performance using these two settings: the vowels-only setting and the all-phonemes setting. For each of these settings in the seven-way accent classification task our NIST dataset allows, the recognition rate is presented in the table below (we would expect a rate of 14.3% correct if the system was working by chance):

Table 6.3: Accent recognition results on the NIST SRE dataset of non-native accents, where the vowels-only and all-phonemes settings have been implemented.

Segmental setting	**% Correct**
Vowels-only	40.3
All-phonemes	52.9

There seems to be a great difference between the two segmental settings. This suggests that there is a lot of distinguishing information across the whole phoneme inventory, which is not constrained to just one group of segments. For the higher performing setting, the all-phonemes setting, the confusion matrix is given below in Table 6.4 for that task.

Table 6.4: Confusion matrix of the NIST SRE non-native accent classification task, where all phoneme segments were included in the analysis.

Accent	Ara.	Ben.	Chn.	Eng.	Far.	Spa.	Rus.
Ara.	**48**	12	7	7	8	6	12
Ben.	11	**65**	12	1	4	4	3
Chn.	6	4	**59**	8	9	8	6
Eng.	7	3	6	**50**	8	12	14
Far.	10	8	10	4	**55**	4	9
Spa.	8	2	7	13	8	**52**	10
Rus.	13	7	3	17	9	10	**41**

The highest-performing group appears to be the Bengali-accented speakers, whereas Russian speakers make up the lowest-performing group. This suggests that the Bengali English accent is the most distinctive out of all the accents. Interestingly, the native English accent does not seem to be performing particularly well as a group. The native English group is ranked fourth in terms of how many speakers are correctly classified. We might expect a higher group recognition rate for the native English speakers because, for reasons discussed further above in this chapter, we might expect less variation among a group of native speakers of a language, and so the system should have formed a stronger representation of this variety. However, this reasoning does not seem to be reflected in these results. It could be that the native accent variation represented (i.e. a range of North American and British varieties, and possibly others) in this model creates just as much instability among the models as non-native groups.

Including Filled Pauses

The speech recognition system that produces phone labels for the NIST data also outputs separate labels for filled pauses ('PUM' and 'PUH'). In the experiments in Chapter 3, on the Northern Englishes data, there were of course filled pauses in the data, because it was spontaneous conversational speech. However, the segments that made up these filled pauses were not separated from the rest of the phoneset. Instead, their component sounds were collapsed into existing phoneme segments in the phoneset (e.g. a filled pause would be represented by 'ax' + 'm' for an 'um'). In the transcriptions provided for the NIST data by the speech recognition system, 'PUM' and 'PUH' map onto two possible realisations for a filled pause: 'PUM' is composed of a vowel and a nasal, and 'PUH' is simply just a vowel sound. In some forensic phonetic studies, filled pauses have been found to be relatively strong speaker discriminators (Wood, Hughes and Foulkes, 2014). Reasons for this include the fact that they can be fairly frequent segments in spontaneous speech and we can also view them as unconscious events, and so they are less likely to be disguised or affected as much by other factors.

Since these additional segments are available for the NIST dataset, it could be of interest to see whether filled pauses can be of value to an accent recognition task, given their suspected value in individual speaker comparison tasks. To investigate, the two filled pause variables have been added to the all-phonemes setting that is used to form the Y-ACCDIST matrices, and the experiments were repeated.

In this segmental setting, the Y-ACCDIST-SVM system generated a recognition rate of 54.4% correct, which is a slight improvement on the all-phonemes setting without filled pauses (52.9% correct), suggesting that filled pauses do

indeed carry some distinctive power for an accent classification task like this one. To take a closer look at the results of this task, the confusion matrix is presented in the table below:

Table 6.5: Confusion matrix of the NIST SRE non-native accent classification task, where all phoneme segments and filled pauses were included in the Y-ACCDIST matrices (54.4% correct).

Accent	Ara.	Ben.	Chn.	Eng.	Far.	Spa.	Rus.
Ara.	**46**	12	8	7	8	7	12
Ben.	8	**69**	12	1	3	4	3
Chn.	5	3	**67**	5	8	7	5
Eng.	7	2	3	**48**	6	20	14
Far.	8	5	10	8	**55**	4	10
Spa.	8	3	3	15	8	**51**	12
Rus.	11	4	5	19	6	10	**45**

In this confusion matrix, we seem to see the greatest improvement is in the group of Mandarin Chinese speakers. Without filled pauses, as a group, they attain 59% correct. When filled pauses are included, this increases to 67% correct. It could be that Chinese-accented English has particularly distinctive filled pauses, over the other varieties. A phonetic analysis would be able to uncover this.

In contrast to the improvement we see in the number of correctly classified Chinese speakers, we see an increase in the number of English speakers being confused for Spanish speakers (an increase from 12 to 20) by including the filled pause segments. Additionally, we see a slight increase in the number of

Spanish speakers being confused for English speakers (an increase from 13 to 15). This may suggest that the filled pauses of English and Spanish speakers are similar to one another, and therefore causing these increases in confusions. Again, a more detailed phonetic analysis should be conducted to confirm this hypothesis.

6.4.5 Engineering Modifications

Until this chapter, the datasets that have been used to test the Y-ACCDIST-SVM system have been relatively small by the standards of much speech technology research (but quite large by the usual standards of sociophonetic research). The reasons for using the AISEB corpus and Northern Englishes corpus were given above, and were more to do with the nature of the accents and the different ways in which these data could challenge the system. Additionally, it is of interest to explore systems' capabilities on smaller datasets for the sake of forensic applications, because it is unlikely that for a given case, forensic speech analysts will have access to large sets of relevant data to work with. However, a larger dataset, like the one used in this chapter, allows us to more reliably test the effects of smaller changes made to the system. By testing the system on more speech samples, improvements and degradations in performance can be properly monitored, rather than only being very speculative about the improvements in performance when only 45 trials have been conducted (in the case of the Northern Englishes corpus, for example). Here, we have 700 test trials, and so we can more reliably test the effects that changes to the system can have on performance.

Two different aspects of the system will be adapted to test different alternations:

1. The distance metric used to construct the Y-ACCDIST matrices
2. Incorporating variance into the acoustic feature vectors to try to account for the dynamic nature of speech.

Both of these aspects will be addressed in turn below.

Testing Distance Metrics

The first way we shall make changes to the system is to modify the distance metric used to construct the Y-ACCDIST matrices. We can refer to Chapter 2 where details of the Y-ACCDIST-SVM system are given in Section 2.3.1. This comes at the point in the process after forced alignment, and once we have computed our average MFCCs for each phoneme in the inventory. In the system's current setup, the Y-ACCDIST matrices are formed by calculating the Euclidean distance between all of the possible phoneme-pair combinations. This was simply to follow the architecture of past ACCDIST-based systems (e.g. Huckvale, 2004; Ferragne and Pellegrino, 2007). However, there are other distance metrics we can test to determine whether Euclidean distance is the most suitable metric. The following distance metrics will be compared:

- **Euclidean distance**

  The Euclidean distance can be used to calculate the distance between two vectors. We can imagine each vector representing co-ordinates in n-dimensional Euclidean space and then simply measuring the length of a straight line between them. The formula for Euclidean distance is given below:

  $$Euclid.dist. = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2},$$

  where a and b are the vectors we are calculating the Euclidean distance between.

- **Manhattan distance**

  The Euclidean distance above was described as the straight line between two vectors in Euclidean space. Manhattan distance takes a grid-based approach, where the distance is computed based on vertical and horizontal lines, where diagonal routes cannot be taken. It is this vertical-horizontal route that is used to compute the Manhattan distance. The formula is given below:

  $$Manhat.dist. = \sum_{i=1}^{n}|a_i - b_i|$$

- **Cosine distance**

  While Euclidean distance is concerned with the straight line between two vectors in n-dimensional space, cosine distance is concerned with the angle between two vectors. Again, we assess the two vectors in space, but instead take the cosine of the angle between them, rather than the traditional idea of 'distance', as such. The formula is given below:

$$Cos.dist. = 1 - \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}}$$

These distances were compared using the segmental setting that has so far yielded the highest recognition rate of 54.4 % correct in the experiments presented in Section 6.4.4 above (all phonemes, plus filled pauses). Again, a leave-one-out cross-validation setup was used to test the system on this seven-way classification task for each distance metric. The results are displayed in Table 6.6 below:

Table 6.6: Accent recognition results on the NIST SRE dataset of non-native accents, when varying the distance metric used to construct the Y-ACCDIST matrices.

Distance metric	**% Correct**
Euclidean distance	54.4
Manhattan distance	50.1
Cosine distance	63.1

Manhattan distance yields a lower recognition rate than Euclidean distance, whereas cosine distance seems to increase recognition rate by a large margin to 63.1% correct. It is difficult to explain why a cosine distance might outperform Euclidean distance on this task.

In the early development stages of the system, different alternatives were trialled, but the amount of data did not allow for these kinds of differences to be expressed, and so the original alternation taken from past studies (using Euclidean distance) was retained. For the remaining experiments using the NIST dataset, cosine distance will be used in the system, rather than Euclidean distance.

6.5 Discussion

We see that, overall, the Y-ACCDIST-SVM system does seem to classify, to some degree, the non-native accent groups that the NIST SRE dataset provides. Section 6.2 listed a number of factors which could lead to weaker Y-ACCDIST models for each speaker. In addition to the fact that non-native accents have been used here, we also only used 'estimated labels' generated

by a state-of-the-art automatic speech recognition system, rather than (presumably) more accurate transcriptions prepared by human annotators. From just observing the results presented in this chapter, it is difficult to determine just how each of these additional challenges impact on performance without a deeper inspection of the data. For the sake of improving performance, it could be of value to discover the specific strengths and weaknesses of the automatic speech recognition system. If there are certain segments that are regularly confused by the speech recognition system, then including these segments in the Y-ACCDIST matrices could well weaken the accent models. Eliminating such segments could improve recognition rates.

We can loosely compare this chapter with Bahari, Saeidi, van Hamme and van Leeuwen's (2013) accent classification study of non-native varieties from NIST SRE datasets. They tested a number of different text-independent systems on the NIST 2008 dataset, using five accent categories, rather than seven. Also, there were much smaller numbers of speakers, as well as an imbalance in the numbers of speakers, per accent group in Bahari *et al*'s study. However, their study does provide some insight into how text-independent i-vector-based accent recognition systems work on non-native accent data of this kind. They report recognition rates of between 41% correct and 58% correct for various system alternations. It could be the case that in non-native experiments, text-independent systems actually perform better than in native-accent experiments (like those on the AISEB corpus in Chapter 2). It could be that the differences brought about by the speakers' L1 are distinctive and consistent enough across the speech signal to be expressed in the acoustic features, without the need for segmental information supplied in advance. In contrast, a segmental approach like Y-ACCDIST could be expected to suffer on a non-native task because of its reliance on consistency of the production

of individual segments, both within speakers and across speakers within a group. A further research direction, then, is to compare text-dependent and text-independent systems on corpora containing native accents and non-native accents to discover whether there are performance differences in this respect.

Chapter 3 attempted to draw a comparison in performance between good-quality recordings and recordings that had been artificially degraded to resemble a quality close to telephone transmission. The degradation was done artificially in order to make a direct comparison of exactly the same recordings to monitor the effects in a controlled way. However, as was acknowledged in Chapter 3, the artificial degradation does not produce data that exactly replicates telephonic recordings. Among other reasons, one is down to the natural behaviour of speakers when speaking into a telephone, compared to a face-to-face conversation. We can expect some *Lombard effect* taking place in a telephone, whereby a speaker may unconsciously adapt his or her speech to increase the likelihood of being understood. The Lombard effect is usually researched and discussed in the context of noisy speech (e.g. Junqua, Fincke and Field, 1999), but we can also expect similar effects in telephone speech. The NIST database has therefore allowed for us to test the Y-ACCDIST-SVM system on realistic telephonic data, which is of course of forensic interest.

6.6 Summary

This chapter has explored whether the Y-ACCDIST-SVM system, the system shown to outperform other types of accent recognition system on another dataset in Chapter 2, can work on the accent groups provided by NIST SRE datasets. These datasets differ from the datasets used in earlier chapters in two key respects: six out of the seven accent categories were groups of non-native

English speakers, and the phone labels used were estimated by an automatic speech recognition system, rather than produced by human annotators. These factors do not seem to prevent the system from performing accent recognition to some level. The best recognition rate from the experiments above was 63.1%, where all segments in the phoneset (plus filled pauses) were included in the matrices, and cosine distance was used instead of Euclidean distance when computing the Y-ACCDIST matrices. This is well above the chance expectation for a seven-way classification task (14.3%), but there is of course plenty of room for improvement. Further research, involving some phonetic analysis or more controlled data collection, is required to gather an understanding of which specific factors (to do with the nature of non-native accents or the quality of the transcriptions) are most problematic for the system to cope with.

CHAPTER 7

Conclusion Frameworks for Accent Recognition

7.1 Introduction

When we consider any kind of system for forensic applications, it is also important to consider how we present the outputs of the system. This includes "human systems" (where *system* is used as a much broader term for analytical methodology, rather than just for automatic technology). French and Harrison (2007) acknowledged the need for forensic speech scientists to change the way they standardly express conclusions to the court or other interested parties. At the time of publication of French and Harrison's paper, in speaker comparison cases, experts would standardly present their evidence by stating that it is "very likely" (just one example selected from an impressionistic scale) that the questioned speech sample was produced by the suspect speaker. In a sense, this is overstating the weight of this one piece of evidence, because we could arrive at that same conclusion for another speaker in that analysis.

We should therefore not link the evidence so directly to the suspect in our conclusions as that is the role of the trier of fact, having combined all of the evidence available. Motivated by a "serious logical flaw" (French and Harrison, 2007: 139) in this way of expressing conclusions, French and Harrison (2007) is in fact presented in the form of a *Position Statement*, which was signed by a number of practising UK-based forensic phoneticians in support of the position. In the foreword of the Position Statement (p. 138), they point out the subtle, but significant, distinction between the task that forensic phoneticians had been doing and the task that they should be looking to be doing:

> "*In the past forensic speech scientists were often thought of as identifying speakers. Within the new approach they do not make identifications. Rather, their role becomes that of providing an assessment of whether the voice of the questioned recording fits the description of the suspect.*"

We should make this distinction because the speech evidence alone (like DNA and other forms of forensic evidence) cannot make an *identification* like this. The speech evidence could also point towards other speakers in the wider population, and this must be acknowledged. It is up to the trier of fact to combine the different pieces of evidence presented in a case to arrive at an overall conclusion over identity. This is not the role of the forensic speech scientist and so it should be reflected in how the conclusions of speech analyses are expressed.

To try to mitigate the logical flaw to some degree, French and Harrison put forward a two-tiered approach to expressing conclusions in an analysis. The first stage is to express the *consistency* between the questioned recording and the suspect recording. The second stage is to express the *distinctiveness* of

the features involved in the analysis. In addition to this proposed approach, they also acknowledged the advantages of implementing the *likelihood ratio framework* (detailed further in Section 7.2 below), but suggest that it is an unrealistic approach to apply to most cases, as it demands large quantities of unavailable data. This argument was reemphasised in French, Nolan, Foulkes, Harrison and McDougall (2010).

In response to the UK Position Statement, Rose and Morrison (2009) welcomed the intentions behind it and are in favour of the direction that the majority of the UK forensic phonetic community seem to be moving in, but they describe the proposed two-tiered approach as a "compromise" and suggest that it does not adequately overcome the problem. They suggest that more should be done to implement the *likelihood ratio framework*, which they believe is the only correct way to express an analyst's conclusions. They argue that it is much more inline with other forensic disciplines.

Saks and Koehler (2005) describe the change in how the courts view forensic analysis of all kinds, and how we are much more ready to question the methods used by forensic practitioners. They highlight the need to challenge expert testimony by stating that the second most common factor that contributes to an individual being wrongfully convicted is incorrect forensic analysis. Using DNA typing almost as a role model for the domain, they call for sound scientific research on methods used right across the forensic sciences, whereby we use scientific protocols and base our conclusions on realistic data, rather than making assumptions about a case. The likelihood ratio framework could help us to remove at least some of these assumptions.

It is more likely that an accent recognition system like the one presented here would be used for more investigative purposes, or in conjunction with human analysts, rather than the outputs being presented in court. There is still

some motivation to incorporate likelihood ratios into an accent recognition. It is of interest to know the strength of the evidence in investigative scenarios to be able to justify resources and efforts given to a specific direction. Additionally, if the LADO application were to employ this kind of technology[1], a likelihood ratio could be more informative about the plausibility of the applicant's claim.

Because of the framework's strengths and current popularity across the forensic sciences, this chapter integrates the likelihood ratio framework into accent recognition. To do this, we will transfer the concept and methodology used in speaker recognition to the Y-ACCDIST-SVM system when training and testing it on the NIST SRE dataset used in the experiments in Chapter 6. So far, we have only conducted what we might call "closed-set" experiments, where we have demanded an accent label to be outputted for each trial. In a way, this resembles an "identification" approach that French and Harrison (2007) rightly propose the forensic phonetics community should move away from. As well as this, it is not necessarily an entirely useful task in the context of forensic applications, because it drastically limits the questions we can try to answer with a forced-choice system. What is expected to be more useful is an "open-set" task, where we ask for the likelihood of a speaker belonging to a given accent category, over other accent categories. We could also draw this distinction using the terms "hard decision" and "soft decision". Until this chapter, we have been making "hard decisions" regarding the test data, but it is perhaps more appropriate in forensic contexts to make "soft decisions". The likelihood ratio framework can help us to achieve this. This chapter looks at how the Y-ACCDIST-SVM accent recognition system performs when

[1]Language Analysis for the Determination of Origin (LADO) was discussed as a possible application of accent recognition technology in Chapter 1 of this thesis.

likelihood ratios are its output, rather than hard-decision accent labels.

7.1.1 Outline

This chapter will first decompose, at a conceptual level, the likelihood ratio framework in Section 7.2. Section 7.3 will then introduce *Calibration*, which is a step we should take to both measure and improve the accuracy of the likelihood ratios a system outputs. The performance measures we will use to evaluate the system when it produces soft-decision outputs will then be described in Section 7.4. We will then present the experiments and results in Section 7.5. Section 7.6 will finally evaluate this soft-decision approach to automatic accent recognition.

7.2 The Likelihood Ratio Framework

A likelihood ratio (LR) is an indication of the weight of evidence expressed as a single number. It provides a way of transparently presenting the conclusion of an analysis. Rose (2002: 57) describes this framework as "logical" and "commonsense". These are properties of an analysis that the forensic science regulators are encouraging (Tully 2016, 2017). This section conceptually describes the LR and what it represents. It is usually considered in terms of a speaker recognition or speaker comparison task, rather than an accent recognition task like the one conducted in this chapter. Initially, we will explain the likelihood ratio in the context of an individual speaker comparison task, and then extend this explanation to the task of accent recognition.

We can present the components of a likelihood ratio through the simple formula below:

$$LR = \frac{p(E|H_p)}{p(E|H_d)}$$

Let us imagine that we have two speech samples, one "unknown" speaker's speech sample and a "suspect" speaker's speech sample. We can see that the ratio is made up of two probabilities. The numerator is the probability of the unknown speech sample having been produced by the suspect speaker, given the evidence (we might call this the *prosecution hypothesis*, H_p). The denominator is the probability of the evidence being found in a population of relevant speakers (we might call this the *defence hypothesis*, H_d). In effect, we have a ratio that puts the degree of similarity between the unknown and suspect samples against the estimated degree of typicality that the evidence presents. In this format, any LR values that are equal to more than 1 are in support of the prosecution hypothesis. Naturally, the higher the number, the stronger the support for the H_p. Conversely, any LR values that are equal to a value of less than 1 provide support for the defence hypothesis. In turn, the lower the value, the stronger the support for the defence hypothesis. A value of exactly 1 does not provide support for either hypothesis.

7.2.1 Log scaling

While this provides a framework to express the weight of evidence, there is a problem with using an LR value in the form we have explained so far. Currently, values that are in support of the defence hypothesis fall between 0 and 1, whereas values that are in support of the prosecution hypothesis fall between 1 and infinity. A skew like this does not allow us to effectively compare the

degree of support between evidence in support of the prosecution hypothesis and evidence in support of the defence hypothesis. To counteract this distribution we can apply log scaling to the values. This provides us with a more even distribution of scores. It also changes the point at which we do not have support for either the prosecution or defence to 0, meaning that any negative scores are in support of the defence, while any positive scores are in support of that of the prosecution. Using a range of log LRs, Rose (2002: 62) offers some "verbal equivalents" to the log-likelihood ratio value range, labelling log LRs of more than 4 "very strong support for the prosecution" and log LRs of less than -4 very strong support for the defence hypothesis, for example.

7.2.2 Application to accent recognition

In applying the likelihood ratio framework to the accent recognition task, we will swap the suspect speaker sample for the collection of samples that contribute to a model that represents a single accent. For individual speaker recognition a number of speakers would be used to form a model of typicality to contribute to the denominator of the LR. In the case of accent recognition, the rest of the accent models for different accents in the corpus are used to estimate a measure of typicality.

7.3 Calibration

Calibration is a means of measuring the reliability or "confidence" (Gonzalez-Rodriguez *et al*, 2007) of a system's outputs, as well as improving overall performance. Rather than simply outputting a likelihood ratio from a system and assessing whether the value outputs a probability that is in favour of

the truth outcome (based on a test set of data), we can gather a much more proportionate indication of how accurate the outputs are. For example, if a system outputs a score that is in strong support of a given hypothesis and it turns out that the truth value lies with the alternative hypothesis, it is important for us to know the extent to which a system does this. In these kinds of instances, we would want to 'penalise' the system heavily and for this to be reflected in an overall measure of the system's reliability. Calibration can help us to capture this kind of information. Ramos-Castro, Gonzalez-Rodriguez and Ortega-Garcia (2006) demonstrate in their experiments the importance of integrating calibration within an automatic speaker recognition system when we are intending to use that system for forensic applications.

It has been suggested that in a classification task, rather than a recognition task like speaker recognition (a sort of binary setup), we might want to conduct calibration to account for multiple hypotheses (i.e. the likelihoods of a speech sample belonging to each of the classes in our set). Brümmer and van Leeuwen (2006) offer solutions to the automatic language identification research community enabling them to calibrate scores for a classification task like this, and indeed, Bahari, Saeidi, van Hamme and van Leeuwen (2013) implement such a method on an accent classification task similar to the one presented in this chapter. While the experiments in this thesis so far have also been to classify speech samples, and these experiments have largely been inspired by previous classification studies, it is proposed here that we continue to use the binary setup that is associated with speaker recognition problems. This is because we are considering accent recognition systems for forensic applications and it is expected that the questions posed to forensic analysts are more likely to involve a binary hypothesis, than multiple hypotheses. That is to say, we are more ready to expect a problem that requires investigating how

likely it is that a speaker's speech sample belongs to a category relative to the likelihood that the speech sample does not belong to that category. Accordingly, the approach taken by speaker recognition research will be applied to the accent recognition task in this chapter.

These experiments make use of *pool adjacent violators* (PAV) calibration (as outlined by Brümmer and du Preez (2007))[2]. Other methods of calibration exist. A more common option for calibration is the use of logistic regression (e.g. Morrison, 2013), but the scores from the Y-ACCDIST-SVM system are not normally distributed, and so a non-parametric method of calibration (PAV calibration) has been selected to produce the likelihood ratios in these experiments. To conduct calibration, we need an additional partition of data to develop the system in this way. This will be implemented in the experiments run in this chapter, as outlined in Section 7.5.1 further below.

7.4 Performance Measures

So far in this thesis, we have largely used the measure of *% Correct* to monitor changes in system performance under different settings or when testing on different data. To a certain extent, this has served its purpose of allowing us to observe these differences. However, to consider this system in forensic contexts, it has been argued that alternative performance measures should also be explored so as to cater for the more common "open-set" type of question that forensic problems often ask of analysts. This section describes performance measures that are standardly used in automatic speaker recognition research: *Equal Error Rate* and *Log-likelihood-ratio Cost Function*. In the first instance,

[2]PAV calibration was conducted by using a MATLAB script written by Daniel Ramos-Castro.

we will describe these measures in the context of automatic speaker recognition, because this is the more established scenario in which they are used. They will be transferred and applied to the task of accent recognition in the experiments in this chapter.

7.4.1 Equal Error Rate (EER)

For speaker recognition, we can generate the Equal Error Rate (EER) through a number of test trials, where we know what the true identities of the speakers in the test samples are. Using these test trials we can draw up the distributions of *same-speaker scores* (the probability that the two samples were produced by the same speaker) and *different-speaker scores* (the probability that the two samples were produced by different speakers - these might also be called "impostor scores"). In an ideal world, these two sets of scores would consistently produce values that are completely separable. We can visualise this via the two distributions in the Figure 7.1:

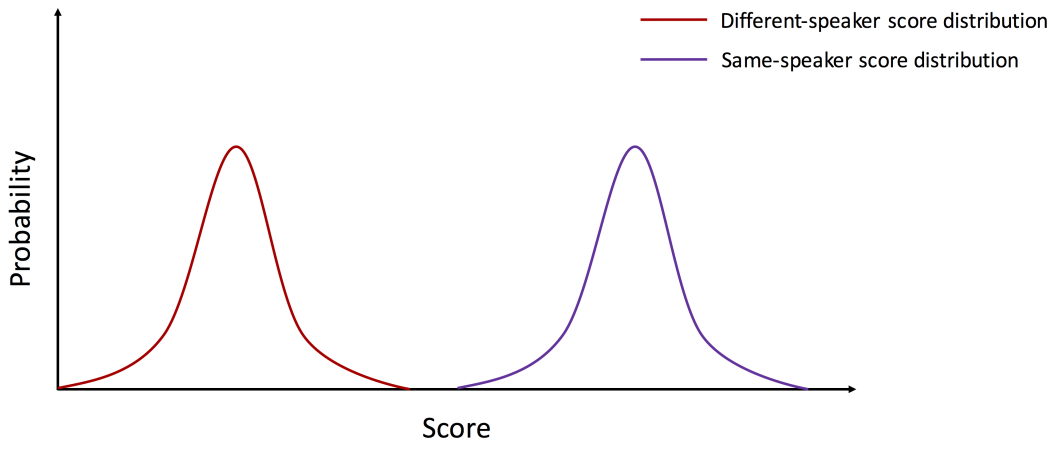

Figure 7.1: Different-speaker and same-speaker scores with no overlap between the two sets of scores.

From this kind of situation, we could take a score from our system and quite confidently determine whether it is a same-speaker or a different-speaker score. However, in reality, we do not tend to get distributions like this. Instead, we encounter distributions of same-speaker and different-speaker scores that overlap (as illustrated in the figure below).

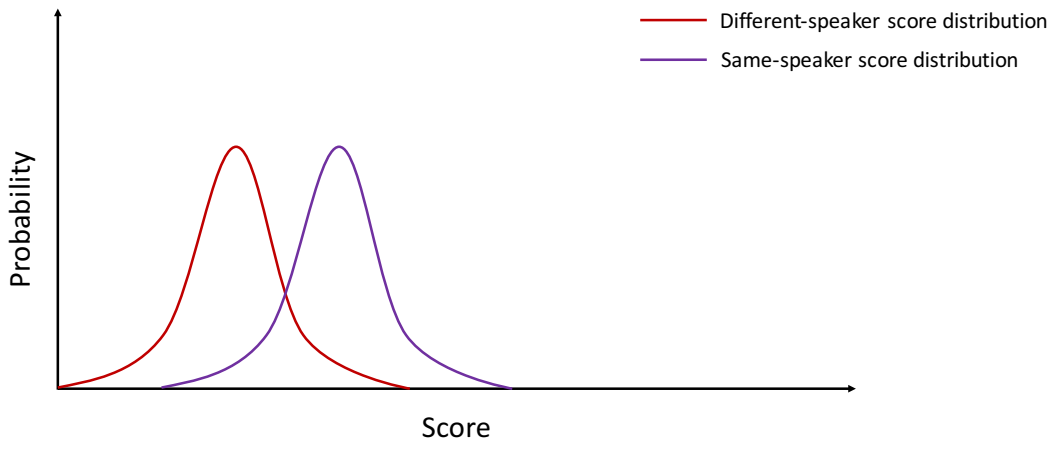

Figure 7.2: Overlapping different-speaker and same-speaker score distributions, this being a more realistic idea of what we can expect in speaker recognition.

The smaller the overlap between these distributions, the better our system is at discriminating between speakers. Based on these two distributions, we can compute a threshold which can determine whether or not the system concludes that a questioned sample was produced by the same speaker who produced the suspect sample.

The overlap between these two distributions inevitably leads to system errors, and we can classify these errors into two types: *False Acceptance Rate* (FAR) and *False Rejection Rate* (FRR). The FAR is the proportion of times a speaker is incorrectly 'matched' with a sample from a different speaker. The FRR is the proportion of times a test speaker is incorrectly concluded to not be the same one who produced the suspect sample. Within the overlap of score distributions, the distributions of these kinds of errors will also overlap, and it is the intersection at which they do that determines the EER (i.e. the point

at which the FAR and the FRR are equal). In doing so, the EER accounts for both kinds of error, and the lower the EER, the better the system. We can imagine this by returning to Figures 7.1 and 7.2 above and what they show. The further apart the two distributions, the lower down it will be where the FAR and FRR intersect, yielding a lower EER.

We can transfer the measure of EER to automatic accent recognition by using *same-accent* scores and *different-accent* scores produced by the system. Distributions and subsequent calculations can then be made in the same way described in earlier paragraphs.

Detection Error Tradeoff (DET) curve

Alongside the EERs achieved by a test set, we can also graphically represent the performance of a system through a DET curve. Martin, Doddington, Kamm, Ordowski and Przybocki (1997) explain what DET curves show, while simultaneously offering reasons why they are preferred over Receiver Operating Characteristic (ROC) curves, which were more commonly used before the introduction of DET curves. One of the features of a DET curve that Martin *et al.* point out as being an advantage over the more traditionally used ROC curve, is that the different systems viewed on a single plot are more separable, allowing us to make more refined comparisons among different systems. The DET curve plots the probability of getting a false acceptance against the probability of getting a false rejection, given a likelihood outputted by the system. The result for a single system is something that looks like Figure 7.3 below, where we have a line that is positioned diagonally from around the top left of the graph to the bottom right:

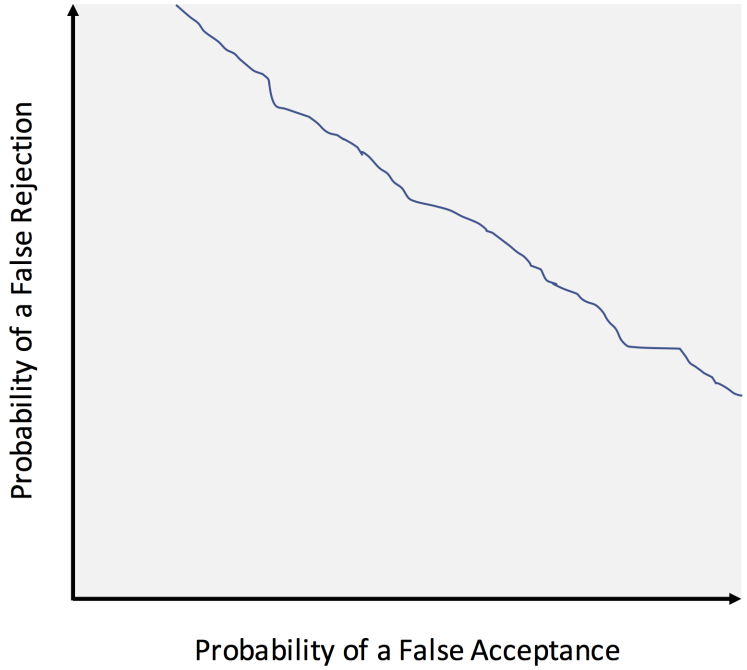

Figure 7.3: An example diagram of a DET curve.

A single line represents the performance of a single system. The further towards the bottom left-hand corner the line is, the better the system (because the probability of getting an error reduces).

The experiments in this chapter (as well as the experiments in Chapter 8) will make use of both the EER and the DET curve to evaluate system performance.

7.4.2 Log-likelihood-ratio Cost Function (Cllr)

The Log-likelihood-ratio Cost Function (Cllr) is a measure of the calibration and discrimination abilities of a system. Van Leeuwen and Brümmer (2007) describe the Cllr as the measure of the quality of log-likelihood ratios. The

following formula (taken from Ramos-Castro (2006)) defines Cllr:

$$\text{Cllr} = \left(\frac{1}{N_{H_p}} \sum_{i for H_p=true} log_2\ 1 + \frac{1}{LR_i}\right) + \left(\frac{1}{N_{H_d}} \sum_{j for H_d=true} log_2\ 1 + LR_j\right)$$

We can see from this equation that the first half takes into account the proportion of likelihood ratios outputted for the prosecution hypothesis (H_p), and the second half does so for the defence hypothesis (H_d). This equation applies a higher penalty to scores that are significantly in favour of one hypothesis, when the alternative hypothesis is actually true. This contributes to a larger Cllr value overall. If the Cllr is higher than 1, the likelihood ratios we are generating are not useful to the task. The closer the Cllr is to 0, the higher the quality of the likelihood ratios.

7.5 Experiments

This section outlines the experimental setup and tools used to output and monitor the performance of the Y-ACCDIST-SVM system when it is adapted to output likelihood ratios, rather than hard-decision accent labels. Concepts discussed in Section 7.4 above have been applied.

7.5.1 Methodology

The same data used in the NIST experiments in the previous chapter have been used to generate the results in this chapter. This is because the NIST subset is the largest corpus used in this thesis which allows us to compute more

reliable outputs. We therefore have seven accent groups, with 100 speakers per group, totalling 700 speakers (more specific details about the data can be found in Chapter 6 in Section 6.4.1).

For calibration to take place, we require an additional partition of data, compared to the experiments we have already run on the NIST dataset. In the previous chapter, we adopted a leave-one-out cross-validation configuration, where each speaker became the test speaker on rotation, leaving the rest of the dataset (of 699 speakers) as training speakers. To be able to integrate calibration, this training and testing setup has been altered. For these experiments, 80 speakers per accent group have been used to train the Y-ACCDIST-SVM system. The remaining 20 speakers per accent (totalling 140 speakers) are used as calibration data. To generate the evaluation trials, the 80 speakers per accent are used in a leave-one-out cross-validation setup to test the system. This is the same way as the experiments in the previous chapter were run, but we are using smaller amounts of data to train the system. This naturally might mean that we reduce the overall recognition rate, as lower numbers of speakers to train the system might lead to weaker accent representations in the SVM. This will be taken into consideration when analysing the results. The EER and accompanying DET curve for this task were computed using the *FoCal* toolkit (Brümmer, 2007).

7.5.2 Results

% Correct

It was mentioned above that the trained system may have been weakened by a reduction in the number of speakers per accent used (when we consider this training and testing setup in relation to the experiments run in Chapter 6).

So as to be able to compare these results with those obtained in the previous chapter, a leave-one-out cross-validation experiment was conducted using 80 speakers per accent group, allowing us to generate a value using the *% Correct* measure of performance. This is to observe whether the reduction of 20 speakers per accent category greatly affects system performance compared with using 100 speakers per accent group. In this setting a recognition rate of 61.4% correct was achieved. We see only a slight reduction in performance relative to when we used 100 speakers per accent to train the system (which yielded 63.1% correct).

EER

This task generated a result of 19% EER. The figure below displays the DET curve for this task.

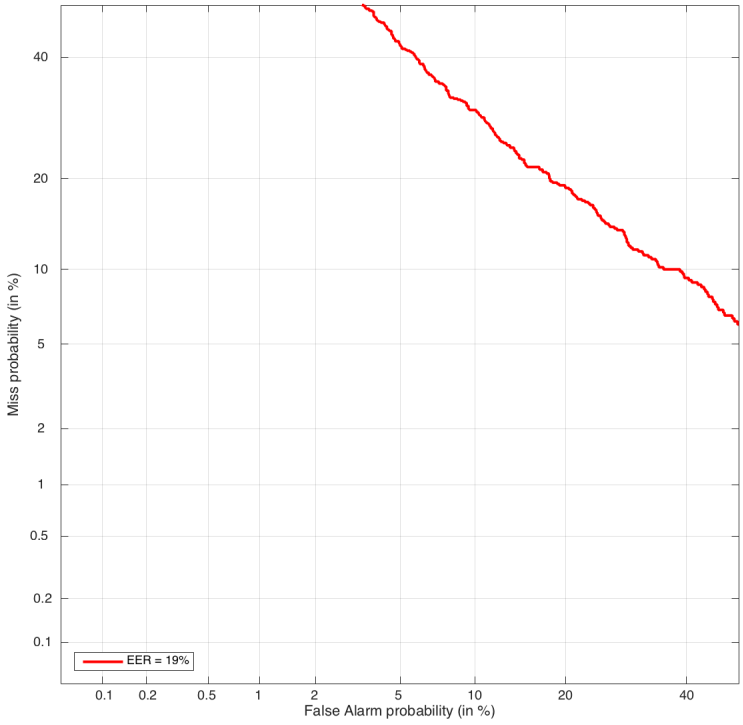

Figure 7.4: DET curve for the NIST accent recognition task

Cllr

Having calibrated the system, we achieve a Cllr of 0.6316. The fact that this value is below 1 suggests that the likelihood ratios the system outputs are reliable, to some extent, but there is plenty of room for improvement, as it is still quite far from 0.

Tippett plot

It can also be useful to observe the likelihood ratios themselves. So far, we have looked at single-value measures (EER and Cllr) that tell us about the likelihood ratios that the system outputs collectively, but we have not inspected the distribution of likelihood ratios. Tippett plots are one way that can allow us to do this. Tippett plots consist of two lines to represent the performance of a single system: one line to represent the same-speaker (or in this case, same-accent) likelihood ratios, and the other line to represent different-speaker (different-accent) likelihood ratios. The same-accent line in the plot represents the cumulative proportion of likelihood ratios that are smaller than the likelihood ratio marked on the x-axis, whereas the different-accent line represents the cumulative proportion of likelihood ratios that are greater than the likelihood ratio marked on the x-axis. The accompanying Tippett plot that illustrates the performance of the system after calibration is given in Figure 7.5 below[3]. The same-accent line is blue, and the different-accent line is red.

[3]This Tippet plot was generated using a MATLAB script developed and made available by Geoffrey Morrison. See: http://geoff-morrison.net/Software/plot$_t$*ippett.m*

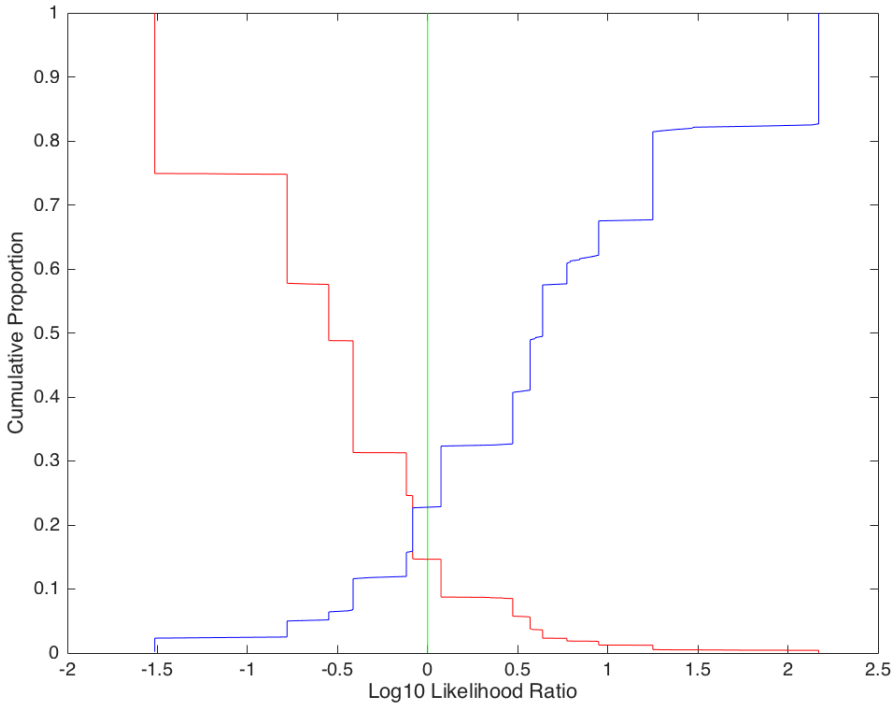

Figure 7.5: Tippett plot of the log-likelihood ratios generated from the Y-ACCDIST-SVM system after PAV calibration.

The stepwise nature of these lines is not typically seen so vividly in Tippett plots. This is a result of the non-parametric PAV calibration method that has been used for this analysis, in combination with a relatively small test set. Typically, logistic regression is used for calibration, but this was not the method selected in this instance, because the scores from the Y-ACCDIST-SVM system were not normally distributed.

The points at which these two lines cross the vertical 0 line are of some interest. This shows us the proportion of these scores that reflect a value that supports the alternative hypothesis (i.e. the errors). Decomposing these, the

point at which the same-accent line crosses 0 into the negative values shows us the proportion of "false misses" the system makes, whereas the point at which the different-accent line crosses 0 indicates the proportion of these scores that we can classify as "false hits". From this Tippett plot, we can see that the Y-ACCDIST-SVM system is slightly more likely to make a "false miss" than a "false hit" using these data.

7.6 Discussion

Compared with the EERs generated by state-of-the-art speaker recognition systems (often using the same or similar data), the results of this accent recognition task do not appear to be so impressive. It seems that Y-ACCDIST models are not as good at distinguishing between accents as i-vector models are at distinguishing between speakers. It could be that the samples we use to train a system on a single accent are simply characterised by too much variation to permit us to confidently assign an unknown speaker.

Within forensic speaker comparison research, there has been some work to look at what is meant by a "relevant population" (Hughes and Foulkes, 2015; Hughes, 2014). They investigated how selecting different datasets that form an estimation of *typicality* (i.e. the denominator of the likelihood ratio formula) can affect the resulting LR from an analysis. Ideally, we need to use reference data that are similar to the speech samples being analysed (for example, to match them in terms of speaker sex, accent, etc.). Among their findings, Hughes and Foulkes (2015) found that when a mismatched dataset is used to make speaker comparisons we tend to produce weaker results in support of the defence hypothesis. There are also questions surrounding the size of the reference database and what number of speakers is sufficient to conduct

a comparison (Hughes, 2017). These kinds of concerns also transfer to the problem of accent recognition when applying the likelihood ratio framework. Perhaps using the other accents in the NIST dataset was not a suitable choice. We should look further into the nature of the "relevant population" we use to compute the LRs for accent recognition (e.g. how many different accents we should include etc.). Throughout this thesis, by running accent recognition experiments on three different corpora, we have seen that the nature of the set of accents is partly responsible for the overall recognition rate. It has been argued that the overall degree of similarity that exists among the accents is likely to play a key role in the likelihood of an unknown speaker being classified. This should be trialled within the likelihood ratio framework to examine the extent to which the degree of similarity among the accents within an accent database affects the likelihood ratios.

7.7 Summary

The likelihood ratio framework was integrated into Y-ACCDIST-SVM accent recognition experiments on the NIST SRE dataset. This was in a move to make "soft decisions" about a speech sample, rather than making "hard decisions", using a framework that is widely accepted across the forensic sciences. In doing so, this chapter has introduced and implemented new performance measures and graphical representations that are typically used in speaker recognition. An EER of 19% was generated for the NIST non-native accent recognition task. This result does not compare well to the EERs we tend to see in automatic speaker recognition research (as we will observe in the following chapter). This thesis has largely observed how various changes (to the accent recognition system and the data) have affected *% Correct*, a performance measure based

on hard decisions. However, it would also be of interest to investigate how similar changes (particularly to the dataset choice) affect these soft-decision measures in accent recognition.

CHAPTER 8

The Y-ACCDIST system as an assistive speaker recognition tool

8.1 Introduction

In previous chapters, some results and system outputs have suggested that there might be scope for using accent recognition technology on an individual speaker comparison basis, not just for categorising speakers. Chapter 2 presented the results from experiments which explored the performance of different accent recognition systems when aiming to distinguish between geographically-proximal accents (i.e. accents that are assumed to share a large number of features with one another due to their geographical proximity). This was thought to be a challenge to a system which is of interest to forensic applications, and the performance of the Y-ACCDIST-based systems on this sort of task showed promise. However, when we looked at the similarity between individual speaker Y-ACCDIST models in Chapter 3, we did not see the patterns we necessarily expected. Individual speaker models did not necessarily fall ex-

clusively by other speakers in the same accent group. There is also reason to believe Y-ACCDIST would not perform very well in this sort of task, because it is thought that the Y-ACCDIST models actually remove voice-quality information that could help to discriminate speakers. Despite this, we did observe some interesting positioning of specific speakers in Chapter 3's visual outputs that indicated how similar the speakers were to one another. There were some individual speakers that were seen to be particularly distant from the rest of the speakers in the dataset, which could be suggesting that these Y-ACCDIST models could discriminate individual speakers. The experiments in Chapters 2 and 3 naturally spark curiosity about whether this kind of approach could offer assistance to forensic speaker comparison tasks (the most dominant type of task in forensic speech science). The main purpose of this chapter is to assess whether the speaker-specific models that a Y-ACCDIST approach forms are fine-grained enough to be able to distinguish between specific speakers. This chapter therefore evaluates the extent to which speaker-specific information remains in the Y-ACCDIST models.

8.1.1 Outline

The present chapter first outlines some of the current approaches used in forensic speaker comparison tasks in Section 8.2. Section 8.3 then turns to approaches to automatic speaker recognition, and where a Y-ACCDIST-based approach might fit among these. Section 8.4 presents the speaker recognition experiments that have been run using our Y-ACCDIST-based approach. Section 8.5 then evaluates Y-ACCDIST's potential in speaker comparison tasks.

8.2 Forensic Speaker Comparison

Forensic speaker comparison takes advantage of the fact that there are collections of speaker-specific features that could, in certain circumstances, discriminate a speaker from others in a population. The aim is to take two or more speech recordings and determine how likely it is that the speech was produced by the same speaker, relative to how likely it was produced by another speaker in the population. This idea was elaborated on in Chapter 7 of this thesis (in Section 7.2) in the context of the likelihood ratio framework.

To form a picture of the kinds of methodologies that are being implemented in current casework, Gold and French (2011) conducted a survey of the methods used by 36 forensic speech analysts who undertake casework around the world. The practitioners had varying casework experience in terms of how many cases they have been involved in and also the countries in which they practise. Gold and French categorised the different types of methodology that might be implemented by these different analysts into five groups:

1. **Auditory Phonetic Analysis Only**, where the analyst listens to the speech samples, taking note of segmental and suprasegmental features.

2. **Acoustic Phonetic Analysis Only**, where the analyst uses software (such as Praat (Boersma and Weenink, 2017)) to extract acoustic phonetic features to make judgements.

3. **Auditory Phonetic cum Acoustic Phonetic Analysis**, where the analyst uses a combination of the methods described in 1) and 2) above.

4. **Analysis by Automatic Speaker Recognition System**, where the analyst passes the speech samples through automatic speaker recognition

software.

5. **Analysis by Automatic Speaker Recognition System with Human Assistance**, where the analyst makes use of an automatic speaker recognition system along with some human analysis (using the methods described in 1), 2) or both).

From the survey, Gold and French reported that a range of these approaches were used. However, no practitioner, at the time of the survey, reported to only use an automatic speaker recognition system (approach 4)). It was found that approach 3) was the most popular approach, followed by approach 5).

If we were to discover that a Y-ACCDIST-based approach does have some value for this kind of task, it would not neatly fall into one of the five categories listed above. Some aspects of it would be automatic, while there is a certain degree of manual pre-processing involved. Fully automatic forensic speaker recognition systems are text-independent, forming a model of the speaker's speech based on acoustic features extracted from across the whole sample. Y-ACCDIST, on the other hand, is a text-dependent system that takes a more segmental approach, modelling individual phonemes separately. We could view a Y-ACCDIST-based approach as a 'hybrid approach'. It analyses and compares phoneme segments (similar to some of the analysis we might see in approaches 1) and 2)), but it also uses aspects that we see in automatic systems (i.e. acoustic feature extraction and recognition through Support Vector Machines).

In Gold and French's (2011) survey, they also asked the participating forensic speech practitioners about their use of population statistics. Population statistics refer to speech features (e.g. formant frequencies, articulation rate,

etc.) that occur among speakers in a population that is relevant or similar to the case. Using these population statistics, analysts can gauge how typical a particular feature value is, which can then assist with the analysis. 70% of Gold and French's respondents claimed that they use these kinds of statistics for casework. Perhaps more importantly, however, 'A number of respondents commented that if more population statistics were available they would use them' (Gold and French, 2011: 299). Collecting and reliably measuring or extracting these kinds of population data can be very laborious. Introducing new methodologies that can extract and model speech more efficiently could help with this problem in the field. If a method like Y-ACCDIST were to work for these kinds of cases, it could help to achieve this aim. Although some pre-processing is still involved in the form of transcription work, extracting features and modelling the pronunciation systems of a database of speakers is comparatively fast.

The automatic properties of Y-ACCDIST could be seen as advantageous to some aspects of forensic science. One part of a forensic speech scientist's work is the analysis itself, but another part is presenting the evidence to the court. The Daubert v Merrell Dow Pharmaceuticals Inc. (1993) ruling highlights the importance of being able to test and retest a methodology and, in effect, communicate the error rate of a methodology to the court. The ability to run masses of test trials is one key advantage of using automatic methods to get a good idea of a system's performance. While not as straightforward as the text-independent systems, where minimal pre-processing is required, a Y-ACCDIST-based approach still offers a relatively good basis upon which to run large numbers of test trials so as to be able to offer the objective performance information that might be viewed as preferable to the court.

Before outlining the experiments carried out to explore this new speaker

comparison objective for Y-ACCDIST, the next subsection discusses text-independent and text-dependent automatic speaker recognition systems.

8.3 Automatic Speaker Recognition

Automatic speaker recognition systems are not just intended for the kinds of forensic speaker comparison tasks that have been referred to throughout this thesis. Often, they are considered for security access applications. Some security applications are exemplified in Furui (1997), such as telephone banking or accessing computer devices through spoken passwords. Another area of research within speech technology that involves speaker recognition research is *speaker diarization* (Tranter and Reynolds, 2006; Anguera *et al*, 2012). Speaker diarization is the task of taking a recording that contains speech produced by multiple speakers, and being able to assign different turns of speech to the correct speaker. This might benefit the resulting transcription from a system. For example, we can imagine a speech recognition system that is being used in a business meeting to produce a transcription of what has been said. Speaker diarization would make these transcriptions much more readable and useful if we can attribute a turn to an individual speaker. For this kind of task, overlapping speech is a particular research problem (Boakye, Trueba-Hornero, Vinyals and Friedland, 2008).

When discussing these speaker recognition technologies in relation to forensic applications, Drygajlo (2007) points out that automatic speaker recognition is a controversial issue, despite the advantages of an automatic analysis mentioned at various points in this thesis. One of the reasons for the controversy is that the area of forensic speech analysis 'lacks worldwide standards and recommendations' (pg. 132). There have been some efforts to resolve

this situation. Recently, Drygajlo *et al* (2016) published some guidelines for 'best-practice' in forensic speaker recognition (for both automatic and semi-automatic methods) on behalf of the European Network of Forensic Science Institutes (ENFSI), an organisation dedicated to improving and standardising practices across forensic science. Drygajlo *et al*'s guidelines list a number of precautions that should be taken, and points we should consider, when we use these technologies for forensic applications. For example, a number of pre-processing guidelines are given (such as removing pauses and separating speakers), as well as encouragement to provide detailed documentation of the training and validating databases used for a given analysis. Developing a set of guidelines like this should assist in standardising practices across analysts and ensuring that the methods used are as transparent as possible.

Another organisation relevant to automatic speaker recognition (and one that has already been introduced and discussed in this thesis) is the National Institute of Standards and Technology (NIST). However, we should keep in mind that NIST considers applications beyond forensic ones. Every year or two, NIST releases a Speaker Recognition Evaluation (SRE) dataset. This dataset is intended as a sort of competition dataset that speaker recognition researchers can use to train and test their text-independent speaker recognition systems, and then they can submit their results by a given deadline. The SRE databases are of a substantial size and so allow for the training and testing of system types that require larger sets (e.g. i-vector-based systems (Dehak, Kenny, Dehak, Dumouchel and Ouellet, 2011)). In recent years, NIST has also started providing a Human Assisted Speaker Recognition (HASR) task (Greenberg *et al*, 2010). The provided data can either be analysed by only a human analyst (or a team of human analysts), or the human analysis can be combined with an automatic system. The purpose of the HASR task is

to provide a task that simulates more closely what happens in forensic tasks, which could allow for an effective comparison of different analysts' approaches to the same problem. Ultimately, the aim of these NIST tasks is to allow us to more directly compare different researchers' and practitioners' approaches to the same data, and to encourage further innovation in the field.

In the same way we can divide automatic accent recognition systems into two categories according to their text-dependency (as we saw in Chapter 2), we can make the same division with automatic speaker recognition systems. As already mentioned, the number of applications that text-dependent systems can be used for is limited compared to the number that text-independent systems can be used for. Again, the Y-ACCDIST-based system is a text-dependent system, and so it is important to consider past work on this category of systems as well. The two subsections immediately below discuss automatic speaker recognition systems with reference to these text-dependency categories.

8.3.1 Text-independent speaker recognition systems

The same kinds of techniques used in text-independent automatic speaker recognition have already been introduced in this thesis in Chapter 2, in relation to the text-independent accent recognition systems tested there. This is because automatic accent recognition research has often followed in the footsteps of automatic speaker recognition research, by adopting the same techniques and applying them to a different kind of task. Until recently, Gaussian Mixture Models (GMMs) were the main modelling technique for text-independent speaker recognition (Reynolds and Rose, 1995). Now, i-vectors are the standard modelling strategy for text-independent speaker recognition (Dehak, Kenny, Dehak, Dumouchel and Ouellet, 2011) as they generally ex-

ceed the performance of GMM-based systems. Details of how these modelling approaches are implemented are given above in Chapter 2. To offer some idea of the level of performance an i-vector-based speaker recognition system can achieve with reasonably good quality telephone data, Garcia-Romero and Espy-Wilson (2011) report their best system's Equal Error Rate (EER) (the lower the EER, the better the system, as explained in more detail in Chapter 7 in Section 7.4) as 1.27%. Of course, researchers run tests under varying conditions to discover the limits of these systems, and so such low EERs are not always achieved where more challenging (e.g. noisy or mismatched) data are involved (Mandasari, McLaren and van Leeuwen, 2012).

8.3.2 Text-dependent speaker recognition systems

Text-dependent speaker recognition systems present more practical limitations than text-independent speaker recognition systems. For most text-dependent systems, the spoken content that comprises the training utterances needs to match the spoken content of the test utterances. The kind of application that could use this setup is where we have security systems that require a spoken password, for example. This is a particularly controlled case of text-dependent speaker recognition. The experiments in this chapter using Y-ACCDIST will still require segmental information about the speech samples, but we do not need the spoken content to match word-for-word.

In the text-independent speaker recognition literature, there are some concerns about the effect of, what has been termed, 'phonetic variability' between training and test utterances that is expected to affect speaker recognition performance (Kenny, Boulianne, Ouellet and Dumouchel, 2007). Phonetic variability between utterances refers here to the difference in phones used between

the utterances, because it is highly likely that they will be composed of different spoken content. The spectral features used to represent the utterances (like MFCCs) are also used for automatic speech recognition, because they are expected to represent the shape of the vocal tract at that time point. While using spectral features for speaker recognition is expected to contribute to the modelling of an individual's vocal tract shape, we can also expect the specific phones used in the utterance to affect that overall model. Text-dependent speaker recognition systems, where the spoken content of the training and test utterances matches, is expected to remove this variability, but this issue could affect overall performance in text-independent tasks.

Due to the added preparatory efforts that text-dependent speaker recognition requires (i.e. collecting very specific spoken recordings or transcribing many recordings), the pool of suitable databases is considerably smaller than it is for text-independent speaker recognition. As well as the number of databases available, the size of a corpus for text-dependent speaker recognition is also likely to be compromised, because of the time and cost required to provide transcriptions of the data, or to control the data. Larcher, Lee, Ma and Li (2014) raise the problem of available and suitable databases, while introducing the *RSR2015* database for text-dependent speaker recognition research. They claim that this database, containing speech from 300 speakers, is one of the largest databases available that would be suitable for text-dependent speaker recognition research. By comparison, the number of speakers made available for the text-independent NIST SRE tasks is something closer to 3000.

There is also some interest in pursuing text-dependent questions in text-independent speaker recognition research. Already described in this thesis, Franco-Pedroso and Gonzalez-Rodriguez (2016) look at the performance of i-vector speaker recognition systems that are, in their terms, 'linguistically-

constrained'. Using the NIST SRE database from 2006, they use an automatic speech recognition system to estimate the phonetic content of the data (i.e. which vowels and consonants are present in the speech samples), and only using specific phonemic segments to compare speakers in an i-vector system, rather than a whole utterance. They found that the systems that were constrained to only the TRAP vowel or only the PRICE vowel (with reference to Wells' (1982) lexical sets) were the highest performing single-phoneme systems. For these systems that were trained and tested on only male speakers, the EERs were 21.21% and 21.38% respectively. While interesting findings were produced by this study, one weakness is the fact that the phonemic content of the data (and therefore the segments that were used for these linguistically-constrained systems) were only estimations by an automatic speech recognition system (in the same way seen in the experiments in Chapters 6 and 7). It would be of forensic interest to confirm Franco-Pedroso and Gonzalez-Rodriguez's findings with hand-corrected data.

8.4 Experiments

To test Y-ACCDIST as a speaker recognition tool, these experiments aim to explore how a Y-ACCDIST modelling approach might be able to capture individual speaker differences based on the speaker's pronunciation system, without voice quality information. These experiments look at the prospect of using this approach with different durations of speech sample (1-minute, 2-minute and 5-minute samples).

8.4.1 Data

The data used for these experiments are from the Northern Englishes corpus (Haddican, Foulkes, Hughes and Richards, 2013), the same corpus used in the experiments in Chapter 3. The motivation behind using this corpus is down to its specific properties. First, the Northern Englishes corpus contains orthographically transcribed conversational data. Spontaneous conversational speech is much more relevant to forensic applications than controlled read speech that some corpora provide. Also, the amount of orthographically transcribed speech per speaker (approximately 10 minutes of net speech) allows us to test this text-dependent methodology while varying the duration, which is also valuable to know in the context of a given methodology.

Ideally, for automatic speaker recognition research we would have a much larger pool of speakers to work with than what we have here. Here, we have a total of 68 speakers to use. As already stated, we have approximately 10 minutes of speech per speaker to use in these experiments. Although it does not provide us with much data, we do have the possibility of trialling 5-minute same-speaker sample comparisons. We have produced results below with this duration, but it is important to keep in mind that the number of trials we could run in this condition is much lower than those we could run using 1-minute and 2-minute speech samples, because we have a fixed total duration of speech available for each speaker. This becomes apparent when observing the DET curves under each of these durational conditions. Nevertheless, we still obtain some idea of the kind of performance we can expect from the Y-ACCDIST system under this condition.

Another criticism of the data setup in these experiments is that our same-speaker pairs come from the same recording (session). Ideally, we would have

same-speaker speech samples from different recording sessions to better match the kind of scenarios we find in forensic casework. It is not expected, however that Y-ACCDIST will be as sensitive to matching samples based on recording session, because of how the modelling stage of the system works. By making intra-recording calculations between segmental representations, it is expected that Y-ACCDIST cancels out the quality information of a recording, and just accounts for the pronunciation information.

8.4.2 Training and Testing

On rotation, each speaker is treated as the test (unknown) speaker, while the remaining 67 speakers from our dataset are treated as impostors. For each round of experiments, two stretches of speech are extracted from the total 10 minutes available (the duration of sample depends on the duration being tested for that experiment). Each sample is used to compute a sample-specific Y-ACCDIST matrix. The matrix from the first stretch of speech per speaker is fed into a SVM with a linear kernel. In a 'one-against-the-rest' configuration, each speaker becomes the target (and therefore the 'one'), while all other speakers become 'the rest'. In turn, each speaker's second sample is used as the test sample and one genuine-speaker probability and 67 impostor probabilities are computed, where each trained speaker becomes the target. This is similar to the accent recognition experiments described in Chapter 2, using the Y-ACCDIST-SVM system, but instead of modelling accents as 'one-against-the-rest', we are modelling individual speakers in this configuration. We can obtain same-speaker and different-speaker probabilities from the SVM to then measure performance.

8.4.3 Performance Evaluation

Until Chapter 7, this thesis has largely monitored the performance of different systems through a simple percentage indicating the proportion of correct accent classifications. Because this chapter is about transferring a system from one type of task (accent recognition) to another (speaker recognition), the percentages of the system arriving at the correct result will be given for the experiments below, so we can make a very loose comparison of performance across the two task types. We can think of this as a *speaker identification* task, rather than a *speaker verification* task. Reynolds and Rose (1995) make the distinction between these two types of automatic speaker recognition task, and it is to do with the type of question we are trying to answer. *Speaker identification* is a closed-set task, in which we decide which speaker out of a reference set is most likely to be the speaker of our unknown sample. *Speaker verification* is the task of determining how likely it is that two samples were produced by the same speaker, relative to how likely they were not. Equal Error Rate (EER) is standardly presented to evaluate the performance of a system, and this will be given in these experiments as well. Descriptions of these measures typically used in speaker recognition research were given in Chapter 7 (Section 7.3).

To summarise, for each set of experiments, we will present system performance through the following measures:

1. % correct for a *speaker identification* result.

2. EER along with DET curves for a *speaker verification* result.

8.4.4 Results and Analysis

This subsection first presents the performance results of a Y-ACCDIST-SVM speaker recognition system with just the good-quality recordings (sampling rate 44.1kHz) at different durations to train and test the system (on 1-minute, 2-minute and 5-minute samples). These results are then compared with those generated using the same data, but after they have been artificially degraded (using the degraded data from Chapter 3, which were downsampled to 8kHz, bandpass-filtered to 250-3500Hz and a-law compression was applied). The performances of all of these duration and quality combinations will first be presented through recognition rates (in terms of a sort of closed-set *speaker identification* task), which is more similar to the accent recognition tasks that have been presented previously in this thesis. Performance will also be expressed via DET curves and EERs.

Table 8.1: Results from speaker identification experiments on good-quality data.

Sample Duration	**Recognition Rate (% Correct)**
1 min	22.1%
2 mins	41.2%
5 mins	70.6%

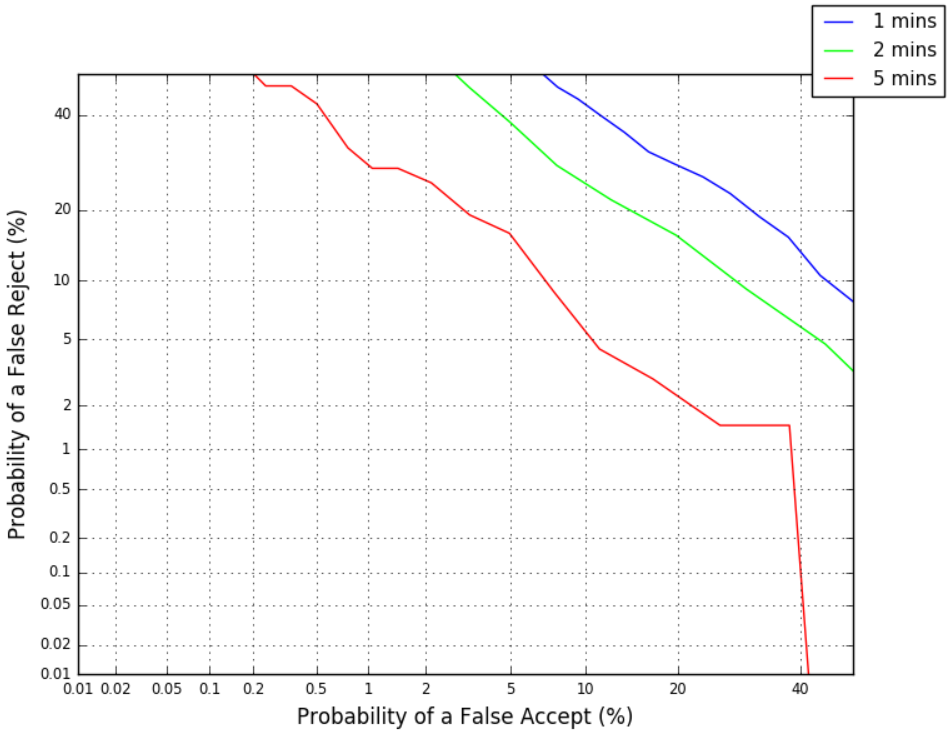

Figure 8.1: DET curves to compare the Y-ACCDIST-SVM system when it has been trained and tested on different durations of speech sample.

Table 8.2: Equal Error Rates (EERs) generated when training and testing the Y-ACCDIST system on different durations of speech sample.

Duration of Sample	**EER**
1 min	24.51%
2 mins	16.86%
5 mins	7.49%

These evaluation measures show an improvement in performance as we increase the sample duration. The discriminatory strength of the Y-ACCDIST speaker models is expected to increase and stabilise as we increase the amount of speech we use to build them. The abrupt drop-off we see in the 5-minute duration curve is down to the reduced amount of data we have for testing this duration.

Degraded Data

Below are the experimental outputs that result from having re-run these same experiments on the same recordings after artificially degrading them (this artificial degradation was conducted for the experiments in Chapter 3 of this thesis). For easier comparison, the DET curves for the experiments run on good-quality recordings (already given above) have been reproduced in the table of speaker identification recognition rates and DET plot below, along with the DET curves generated from the experiments run on the artificially degraded data.

Table 8.3: Results from a speaker identification experiments on good-quality data and degraded data.

Sample Duration	**Recognition Rate (% Correct)**	
	Good-quality data	**Degraded data**
1 min	22.1%	11.5%
2 mins	41.2%	18.8%
5 mins	70.6%	39.7%

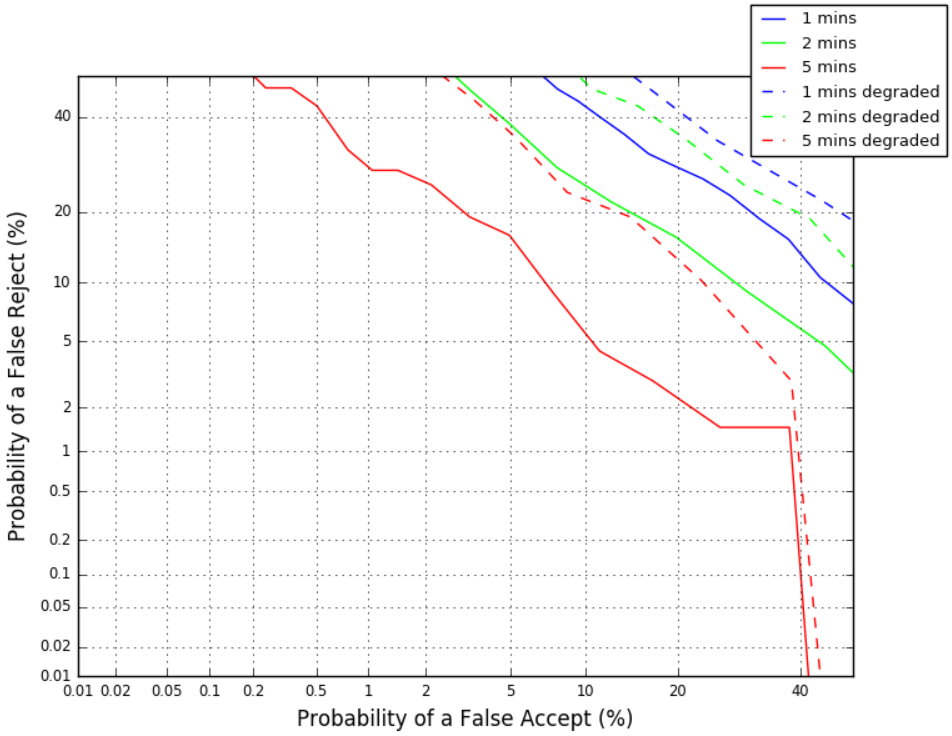

Figure 8.2: DET curves to compare the Y-ACCDIST-SVM system when it has been trained and tested on different durations and qualities of speech sample.

Table 8.4: Equal Error Rates (EERs) generated when training and testing the Y-ACCDIST system on different durations of speech sample after artificial degradation.

Duration of Sample	**EER**
1 min	30.12%
2 mins	27.18%
5 mins	15.14%

Overall, a degradation in the data has a negative effect on speaker recognition performance (as we would expect), and the hierarchy of performance among the different durational conditions remains.

8.5 Discussion

The results above show that Y-ACCDIST models have some speaker discriminatory power, although they have certainly not been found to be particularly powerful characterisations for speaker comparison purposes.

We should also keep in mind the difference in speech sample durations from those that other speaker recognition researchers tend to use, and compare them to the relatively long samples that have been used in the experiments in this chapter. A lot of speaker recognition research involves training and testing on 30-second speech samples or 10-second speech samples. The test samples that were provided for the NIST SRE 2012 campaign ranged between 20 seconds and 160 seconds. Samples of these shorter lengths could not be modelled well using Y-ACCDIST, as suggested by the results above, since it requires a number of tokens of the same phoneme for Y-ACCDIST to be able to generate a sufficient representation of that phoneme. This of course significantly limits the number of applications, or specific cases, that a tool like this could be used for.

It is interesting to observe that as an accent recognition system, Y-ACCDIST-SVM performs well in comparison to other types of accent recognition system, whereas in a speaker recognition task, Y-ACCDIST-SVM performs poorly. These results confirm that we cannot necessarily assume that a method that has been shown to be good for discriminating fairly similar accents is necessarily good for discriminating individual speakers (although, we should keep

in mind the small dataset that has been used in this chapter). Hughes and Foulkes (2016) presented findings on the same corpus used in this chapter, the Northern Englishes corpus, where they looked at the capacity of different formant frequencies (F1, F2 and F3) to distinguish between individual speakers, and also their capacity to distinguish between accent groups of speakers. Their results showed that features that seem to have the most speaker discriminatory power did not necessarily have the most power in discriminating between accents. F3 was shown by Hughes and Foulkes to be the best performing feature in discriminating speakers, but it was the worst feature for distinguishing between accents. These findings seem to be mirrored in this thesis. While we seem to have a good method for distinguishing accents (and that is what we see from the Y-ACCDIST-based systems), this method does not yield good results in distinguishing between speakers.

Although it has been an interesting question to consider, we have established that a Y-ACCDIST-based system does not perform well on speaker recognition tasks, and so we can conclude that this is not a research direction that is worth pursuing.

8.6 Summary

This chapter has presented initial results that use a Y-ACCDIST-based approach to speaker recognition. While a Y-ACCDIST-SVM speaker recognition system shows some capacity to discriminate speakers (i.e. it performs above chance level), it does not seem to produce results that are close to those produced by state-of-the-art i-vector speaker recognition systems, and when taking into account the lengths of speech samples that are required to model a speaker in Y-ACCDIST, it seems that the number of applications that we

could use it for might be limited.

However, we still do not know how a Y-ACCDIST-SVM speaker recognition system compares with an i-vector system in terms of the speakers they incorrectly classify. It would be of interest to see whether the two approaches are complementary in anyway, or whether there is overlap in terms of the specific speakers that cause system errors.

CHAPTER 9

Overall Discussion and Conclusions

What distinguishes this work from other automatic accent recognition studies is the repeated reference to forensic applications. This thesis has continually considered what the experimental results presented in this thesis might mean for forensically-realistic tasks. At this point, it is therefore important to offer an overall evaluation of how suitable the Y-ACCDIST system is for forensic tasks. We should also suggest steps that could be taken to further establish the Y-ACCDIST-SVM's suitability for forensic applications. To do this, the current chapter is divided into two main parts. The first part will bring together the key findings from the thesis, and summarise what they mean for accent recognition technology in the context of forensic applications. As is the nature of research, there are obviously countless research projects that could build on the work presented herein. The second part of this chapter will therefore present some of these possible research avenues.

9.1 Overview of Key Findings

This section takes findings and implications from across the chapters of this thesis to review both the Y-ACCDIST-SVM system and automatic accent recognition more broadly. These will of course be evaluated in the context of forensic applications. The key points for discussion have been broken down and organised under a number of headings.

Smaller and more similar datasets

One of the key observations in this work is how the main system in focus, the Y-ACCDIST-SVM system, has transferred between datasets with different properties. In doing this, we have been able to establish which kinds of task Y-ACCDIST seems to perform well on, and which kinds of task it does not. Unlike other applications, forensic ones particularly require methodologies that are robust to changes in the data under scrutiny. Some applications, more commercial ones for example, might be satisfied by only a fixed set of training data (i.e. the same set of accents), but technology developed for forensic applications should ideally be trained, and should function well, on numerous and diverse datasets. Chapter 2 demonstrated Y-ACCDIST-SVM's performance relative to other types of automatic accent recognition system. The size of the dataset used for testing was likely to play a large role in the low performance of the text-independent systems, as well as the nature of the accents themselves that were being used. Chapter 2 indicated that Y-ACCDIST-SVM shows promise for smaller-data tasks, as well as tasks where the degree of similarity between the accents is expected to be higher. These two traits in a system are favourable for forensic applications, as larger 'relevant' datasets

are unlikely to be available (ones of the size required to sufficiently train the kinds of text-independent systems that were tested in Chapter 2). Also, to be able to distinguish between accents that present more of a challenge is of greater value to forensic applications. If the task is easy, then it is perhaps pointless to invest in technology for that purpose.

There might be some interest in accommodating low-resource accent recognition tasks in future research plans. Particularly for forensic applications, in which it is not necessarily obvious or predictable which accent varieties will be relevant to future cases, we might want to consider research on more generalised 'low-resource' approaches. The idea of 'low-resource' tasks has been considered for automatic speech recognition (e.g. Thomas, Seltzer, Church and Hermansky, 2013), in cases in which we might be trying to develop speech recognition systems for languages for which we simply do not have the training resources to develop systems that use data-hungry algorithms. To a certain degree, and in a similar way, low-resource methods have been explored in relation to language recognition. Van Leeuwen and Brümmer (2008) focus on this in relation to the NIST 2005 Language Recognition Evaluation data, where they focus on the performance of different systems trained on only one hour of language data.

If we are considering developing accent recognition technology for forensic purposes, then this low-resource research theme should be continued in order to provide methodologies that are as versatile as possible.

Text dependency

One of the main features that distinguishes Y-ACCDIST from other ACCDIST-based systems (Huckvale 2004, 2007; Hanani, Russell and Carey, 2013) is that Y-ACCDIST has been developed to process content-mismatched speech. In

other words, the spoken content of the training speech samples does not need to match that of the test speech samples. We therefore needed to redefine text-dependent systems and divide them into two groups: those which require the spoken content of recordings of test speakers to match that of the training speakers, and those that do not. Although the latter category still requires some manual preprocessing to provide a transcription for an analysis take place, it is still much more suitable for forensic applications than the former category. We have seen throughout this thesis that an ACCDIST-based system that has been adapted to process content-mismatched speech can work to some level across different databases.

Regarding the transcription, in many of the experiments in this thesis we have been making a key assumption that, in the text-dependent condition, the transcription will always be correct. In terms of the practical implementation of a technology like this, we might be tempted to assume that only a lower level of transcription (orthographic transcription) is required, and so we might not necessarily need specially trained individuals to provide those transcriptions. Within forensic speech science, there is a call to avoid underestimating how difficult transcription can be. Fraser (2015) shows us the extent to which errors can occur in transcriptions of covert recordings, which are of poor quality (as is typical of forensic casework). Police officers might sometimes provide transcriptions of recordings for the court, and Fraser (2017) describes the quality of these transcriptions as "amazingly low". We should therefore aim to use transcriptions provided by trained specialists, and the risks of using low-quality transcriptions might be a factor we ought to consider in relation to the Y-ACCDIST system. Although it will be more costly only to use trained transcribers for the recordings that are fed into the system, it could prove valuable.

Chapters 6 and 7 showed that the Y-ACCDIST system can work to some extent with inaccurate transcriptions, because the input transcriptions were generated using an automatic speech recognition system. However, we do not know whether, or by how much, automatically generated transcriptions impact on accent recognition performance. Also, do these inaccuracies necessarily have the same impact across different sets of accents? If we were to estimate the transcriptions for a more similar set of accents (like the AISEB corpus in Chapter 2), would we see a large drop in performance because we lose some of the key accent-distinguishing information in the speech recognition errors? This thesis has largely only considered whether a transcription is present or not, regardless of its quality. There are further factors we should take into account with respect to the quality of these transcriptions.

Recording quality

It was expected that the Y-ACCDIST-SVM system would be able to cope with a degradation in recording quality. This is because of the intra-speaker segmental calculations that are computed to model the speaker. In contradiction to this hypothesis, Chapter 3 revealed a substantial reduction in performance under the degraded condition. It seems that the quality of the data has a greater impact on performance than was initially expected. A further look into how Y-ACCDIST processes degraded data showed the increase in similarity between all speaker accent models, regardless of the speakers' accent. The degradation of the data means that fewer differences are expressed through the Y-ACCDIST methodology. As a consequence, we achieve a lower performance.

Chapter 3 provided an opportunity to see what difference data quality can make to accent recognition performance, because of the controlled experiments conducted by degrading the same data to make a direct comparison. However,

there were problems with these experiments because the degraded recordings were only approximations of telephony and were not genuine telephone recordings. The NIST SRE dataset consisted of genuine telephone recordings for accent recognition experiments to take place, demonstrating that the Y-ACCDIST-SVM system can work in some cases on telephonic data.

Segmental Factors

Working from different directions, Chapters 4 and 5 observed the contribution that individual phonemes made to the successful classification of speech samples. Chapter 4 incorporated feature selection to reduce the Y-ACCDIST models to contain the most valuable features (the phoneme-pair distances) and use only those for classification. We saw feature selection improve accent recognition rates when we were using the AISEB corpus to train and test the Y-ACCDIST system, but did not see such an effect when we used the Northern Englishes corpus. This was partly put down to the fact that we might expect to find fewer features that help to distinguish between the AISEB varieties because these are thought to be more similar accents. Feature selection is expected to improve performance in the case of more similar varieties as there are more 'noisy' features to remove.

In running these feature selection experiments, we were also able to identify which specific features were estimated to be most valuable to a given accent recognition task. Through this, we could gather an idea of which phonemes are particularly distinctive among a set of accents. By showing us which phonemes are helpful to the system, this could also be offering sociophonetic information about our dataset of accents.

Phoneme frequency was shown to have some effect on the estimated value of a phoneme to an accent recognition task. The higher the frequency, the

more stable the features are expected to be in the models, and so if a phoneme is a good distinguisher for a set of accents, then it is more likely to contribute to a successful classification if it is also a frequent segment.

Chapter 5 observed the effect of individual segments and phoneme frequency from a different angle, in the context of the test data rather than the training data. It follows that the unknown speech sample is more likely to be correctly classified if it contains larger numbers of certain phoneme segments. We are therefore relying on the natural phoneme distributions of the language to provide enough indicative accent information for a Y-ACCDIST analysis. These experiments showed that the particular segmental composition of a test sample does have a significant effect on its likelihood of being successfully classified. This chapter did not go as far as discovering the criteria a speech sample needs to meet for it to be reliably analysed, but the results suggest that there is a need to conduct further research to uncover such criteria.

Y-ACCDIST Speaker Recognition

Despite Y-ACCDIST's relative success on accent recognition tasks, Chapter 8 revealed a rather poor speaker recognition performance obtained by Y-ACCDIST speaker models, in comparison to other types of automatic speaker recognition system. However, the types of system that do tend to perform well in automatic speaker recognition (e.g. GMM-UBM and i-vector systems) do not perform well in automatic accent recognition, as shown in Chapter 2. This is an interesting contrast to draw, and demonstrates how the different types of model capture different types of variation in a complementary way.

9.2 Further Directions for Research

This section offers just six of the many possible topics for further research.

Accent Disguise

One interesting research avenue, particularly in the context of forensic applications, is whether and how a system like Y-ACCDIST could process speech suspected of being disguised. There have been forensic cases where an individual has committed fraud and imitating another voice has aided the deception. Foulkes, French and Wilson (2018) discuss a specific case, the case of *Lord Buckingham*. In this instance, a North American individual falsely took on the identity of a British Earl, and so an imitated British accent was adopted to reinforce this identity. As an indication of volume, Künzel, Gonzalez-Rodriguez and Ortega-Garcia (2004) say that approximately 15% of cases submitted to the German Federal Police Speaker Recognition Department are suspected to involve some form of voice disguise. *Voice disguise*, of course, may not necessarily entail the specific form of *accent disguise*, but this may occasionally feature in forensic casework. Voice disguise can involve altering the voice in different ways (i.e. voice quality factors, pitch, etc). However, because Y-ACCDIST aims to exclude these sorts of voice quality properties, we can only discuss its application in relation to just one form of voice disguise: *accent disguise*.

Singh, Gencaga and Raj (2016) carried out a study that analysed the disguised speech production of a professional mimic. The imitations the mimic was asked to do were based on the voices of a number of public figures and so both voice quality and accent characteristics were involved in the disguise.

They took formant measurements to determine the nature of the variation of phonemes between the different imitations. They discovered three (/dʒ/, /tʃ/ and /j/) that did not vary across all the different imitations, while all the vowel phonemes varied greatly. The phonemes that do not seem to vary could be useful in cases of suspected voice disguise. However, further research on a larger number of professional mimics would be required to confirm this.

We could extend a study on voice disguise using Y-ACCDIST. Its segmental approach could analyse the effects of disguise, specifically in the accent domain. Also, it could be valuable to conduct an analysis which is based on MFCC features, rather than formants. Automatic formant extraction is not known to be wholly reliable (Duckworth, McDougall, de Jong and Shockey, 2011; Harrison, 2013). We may achieve a clearer picture of the production effects of accent disguise using a higher-resolution feature: MFCCs.

Jenkins (2016) conducted a small-scale study that compared Y-ACCDIST's capabilities of assessing accent-disguised speech against the capabilities of human listeners. She used as stimuli recordings of speakers imitating Scottish English, as well as genuine Scottish English speech. She discovered that Y-ACCDIST and humans perform in different (and possibly complementary) ways, and found that Y-ACCDIST made fewer false positive judgments than the human listeners. As Jenkins acknowledges, however, a larger amount of data would be required to confirm this comparison. Also, a larger number of speakers would be required to determine a reliable recognition rate that would indicate how well Y-ACCDIST could classify disguised and genuine accented speech recordings. By running this small-scale study, Jenkins (2016) has sparked curiosity around Y-ACCDIST's capabilities in the context of accent disguise with these initial findings.

Y-ACCDIST as a tool for sociophonetic research

Forensic speech science often turns to sociophonetic research literature to assist in casework. If a case involves speech from a particular variety of a language, it is important to know the features of the variety. However, of course it is more than possible to come across varieties which have not been researched, or varieties which have not been researched for a number of years. Accent varieties undergo change, and for sociolinguists to keep on top of these changes is a challenging prospect. In Brown (2014), the effect of a system trained on older accent varieties, when faced with newer accent varieties was shown. Under this kind of mismatched condition, we achieved a recognition rate of 65% on a four-way accent classification task. When the data are not mismatched in this way, however, the recognition rate is substantially higher at 83.3% correct. These results indicated the impact on accent recognition performance if the training data are not updated. These results also demonstrate the need to regularly update our knowledge of accent varieties, and so alternative and more efficient analytical methods can help us to achieve this.

Of course, more traditional sociophonetic techniques have value. Common techniques include measuring formants or Voice Onset Time (VOT), and these can lead to findings about linguistic varieties. The same techniques can be transferred to forensic casework and so, in many ways, the two subdisciplines go hand in hand. These analytical methods are relatively time-consuming and, consequently, more costly. Developing new and more efficient techniques could potentially benefit sociophonetic research, as well as forensic casework. Not only could introducing new techniques increase the efficiency of research, but it could also bring the researcher complementary methodologies, presenting different findings about the varieties that we would not otherwise find using the more traditional techniques alone. The intention here is to propose that we

look to combine different methods, rather than replace some with others. This was demonstrated in Vieru, de Mareüil and Adda-Decker (2011), for example, who combined acoustic measurements (e.g. vowel formant measurements and segmental duration) with data mining techniques to determine the most effective of these measurements in distinguishing between speakers of non-native French accents.

Nerbonne and Kleiwig (2007) express the importance of sociophonetic approaches which move away from a selective strategy of simply choosing a few linguistic variables that are expected to exhibit variation. They talk about 'dialectometry' and promote the advantages of more objectively incorporating a collection of features into our studies. They discuss and compare a number of similarity measures that can be applied to dialect variation studies. In particular, Nerbonne and Kleiwig apply a variant of Levenshtein distance to dialect data, which involves tracking the cost of each of the segmental differences (i.e. deletions, insertions and substitutions) between phonemic transcriptions of the the same word in two different dialects. The outcome is an overall score for the pair that reflects how similar or different they are. By approaching dialect difference in this kind of way allows us to incorporate a collection of linguistic variables, rather than a select few like we see across a range of sociophonetic studies. There are of course problems with this kind of approach, however, in that insertions, deletions and substitutions work at the phonemic sequential level of analysis. This might work when we make comparisons across some varieties, but a lot of accent research would benefit from looking more closely at the phonetic realisational differences which exist between accents. A Y-ACCDIST-based approach to sociophonetics could provide an analysis using a collection of linguistic variables, as well as measuring similarity through the different phonetic realisations between different accents. This approach might

be more suitable for analysing some corpora, whereas a Levenshtein-based approach might be more suitable for others, depending on the type of variation we wish to observe for the particular study. As mentioned earlier in this thesis, Y-ACCDIST has been developed so to allow us to process content-mismatched speech data. Levenshtein distance requires one-to-one sequences of words, so these tools are suitable for different types of corpora, as well as, probably, different types of variation.

Brown and Wormald (2017) demonstrate how Y-ACCDIST could be beneficial to sociophonetic research by producing outputs from the Y-ACCDIST system when trained on a corpus of Panjabi-English (Wormald, 2016). Because a detailed phonetic analysis of the corpus had already been conducted, we were able to see how Y-ACCDIST's outputs corroborated results gathered using more traditional sociophonetic research methods. The varieties in this corpus that were shown to be more similar to one another through a conventional sociophonetic analysis were also shown to be more similar to one another by the Y-ACCDIST outputs. However, Y-ACCDIST also highlighted some interesting features that were not identified by the phonetic analysis. For example, a feature ranking from running a feature selection task on the varieties in the corpus indicated that /ə/ could be of particular interest in the case of these accents. These sorts of outputs could uncover features that might otherwise go overlooked by more traditional phonetic methodologies.

Comparison with human accent recognition performance

One clear direction to take would be to compare the recognition performance of automatic accent recognition systems with that of human listeners. These kinds of comparisons have been made in past studies. For example, Hanani, Russel and Carey (2013) asked lay listeners to classify speech samples into

one of the 14 accent classes represented in the Accents of the British Isles (ABI) corpus (D'Arcy, Russell, Browning and Tomlinson, 2004). They could then compare human performance with the performance of automatic systems. They reported that human listeners obtained a lower performance score than nearly all of their automatic accent recognition systems, with 58.24% accuracy.

While it is interesting to observe the performance of lay listeners on a task like this, of interest to this research is how forensic speech practitioners perform in an accent recognition task, compared with how automatic systems perform. Perhaps the Y-ACCDIST system does not correctly classify the same speech samples as an expert. It would be of value to investigate whether there is any complementarity between expert analysts and automatic accent recognition systems.

Cross-linguistic research into automatic accent recognition

Throughout this thesis, we have mentioned that automatic accent recognition performance is likely to depend on the nature of the set of accents we are trying to distinguish between and the degree of similarity (or number of commonalities) that exists between them. We have been looking at how one accent recognition system transfers between different accent datasets. Another consideration might be the specific language itself. It could be that some languages have fewer of the kinds of differences that lead to good differentiation using Y-ACCDIST models, and more of other kinds that would not be captured so well.

The experiments in this thesis have been on various accents of English. Depending on the sets of English accents we are focussing on, the phoneme inventories have consisted of around 42 phonemes. Some languages have smaller phoneme inventories, and others have larger ones. Because the Y-ACCDIST

system relies so heavily on the language's phoneme inventory to provide the foundations for the accent models, it would be interesting to explore whether performance deteriorates when we only have access to a small phoneme inventory. A study of accent recognition across a number of languages would be an interesting task to carry out.

Incorporating prosodic features

We know that prosody can help to discriminate accents (Grabe, 2004), a feature the Y-ACCDIST system is not designed to account for. Vieru, de Mareüil and Adda-Decker (2011) acknowledge prosody in their analysis of foreign-accented French speech. They measured prosodic cues for their accented speech data, as well as segmental cues (vowel formants, segmental duration, etc.). Using data mining techniques, they were able to produce a ranking of which cues are most valuable to distinguishing between the six accents of French in their database. In this ranking task, they found that prosodic cues are of less value to the task than the segmental cues. Of course, it is quite possible that for the task of classifying these six specific non-native accents of French, prosody does not play a large role. Prosody might play a greater role in distinguishing between another set of accents. Alternatively, the measures that Vieru *et al* used to characterise prosody in the speech samples might not have been the most effective.

Biadsy and Hirschberg (2009) incorporated prosodic features into their automatic accent recognition system, which was developed to classify Arabic speech samples into dialect groups. They showed an improvement in performance when prosodic information was added to the process. Although prosodic information is not necessarily expected to help in all accent classification tasks, it would be of interest to integrate and compare methods of

quantifying prosodic patterns and combine them with the Y-ACCDIST-based methods shown in this thesis.

Looking beyond the phoneme

We could suggest that the segmental generalisations that the Y-ACCDIST system makes throughout this thesis are too broad. By assigning a phoneme class to each speech sound, we are possibly losing some resolution in the variation that the data offer. If we were to use syllable-based units, we could uncover more specific segments which carry distinctive value in a text-dependent accent classification or speaker recognition task. Some phonemes are systematically realised in a particular way depending on its position within a syllable. For example, /l/ in many varieties of English is produced as a 'light'-/l/ when in a syllable onset (at the beginning of a syllable) and as a 'dark'-/l/ in a syllable coda (at the end of a syllable) (Sproat and Fujimura, 1993).

At a more subtle level, Greenberg (1999) demonstrates how consonants can be produced differently depending on whether they are situated within a syllable onset or coda. He demonstrates how, particularly in conversational speech, sounds in a syllable coda are much more likely to be deleted than the sounds in an onset. He describes the syllable onset as the "survivor" component of the syllable (p. 163), whereas the coda is described as being "disposable" in comparison (p. 166). In the context of the Y-ACCDIST system then, it could be of interest to form our phoneme representations using only sounds in the onset, because by including sounds in the coda, we could be modelling sounds based on their deleted forms in spontaneous speech. Deletions could subsequently create noise within the Y-ACCDIST matrices. Alternatively, it could be valuable to split consonant phoneme classes into two: one for those phoneme tokens that are found in a syllable onset, and those that are found

within a syllable coda. Naturally, this would come at the expense of the system becoming more text-dependent by requiring much more specific contexts. Another consideration related to this is the issue of syllable stress. Sounds are produced differently according to whether they fall within a stressed syllable or not. This could also impact on the phoneme representations that form the Y-ACCDIST matrices. Perhaps conditioning the phoneme classes to only include segments within stressed syllables, and exclude those in unstressed syllables, could have an effect on accent recognition rates.

9.3 Conclusion

The key strengths of the Y-ACCDIST-SVM system are its relative sensitivity to fairly similar accents and the ability to work with smaller datasets, compared with other types of system. These traits are expected to be favourable to a system intended for forensic applications, because conducting accent recognition on an easy task is unlikely to be of interest to forensic analysts, and also a methodology that can be transferred between accent recognition tasks is more likely to be useful to a domain that operates on a case-by-case basis. This is because different resources are available for different cases.

However, we still do not see an accent recognition performance that compares with the performance of automatic speaker recognition systems. When we applied the conclusion frameworks that are thought to be a more appropriate way of presenting the outcome of an analysis (likelihood ratios), we achieved 19% EER using the Y-ACCDIST system. Although this is comparable with the results of other automatic accent recognition studies, it is not comparable with the performance of different types of recognition system (namely speaker recognition systems). In the *Introduction* of this thesis, the

issues concerned with using technology within forensic science and the criminal justice system were emphasised. Implementing a new methodology into this sensitive context is not a straightforward process. We have discussed the problems that an automatic speaker recognition analysis encountered in one particular case, Slade & Ors v Regina [2015], in the *Introduction*. While this research has contributed to the area by establishing some of the strengths and weaknesses of one particular accent recognition system, we certainly cannot conclude that it is ready to use for forensic casework. The findings in this thesis are simply a step forward towards developing new methodologies for the purpose of forensic speaker profiling. It seems that Y-ACCDIST would not be appropriate to use for all accent recognition tasks. One promising line of inquiry could be to advance research into identifying the kinds of cases and speech samples that are suitable for a particular methodology, rather than developing one single methodology that works for all data types. Perhaps a "one-size-fits-all" approach is simply unachievable in this context.

Even if an automatic accent recognition system cannot be used in all cases, we cannot ignore the fact that it presents qualities that are desirable in a forensic analytical methodology. The most viable means to take advantage of the objective, testable and data-driven properties of an automatic system is to use it in conjunction with a trained human analyst, but there are plenty of questions over how this kind of collaboration would be put into practice. Of all the possible future research directions, considering the issues and sensitivities surrounding forensic applications, how we can combine human expertise with the strengths of an automatic system is one that could greatly benefit the field.

Legal Cases

Daubert v Merrell Dow Pharmaceuticals [1993] 509 US 579

Slade & Ors v Regina [2015] EWCA Crim 71

Bibliography

Alam, F. & Stuart-Smith, J. (2011), Identity and ethnicity in /t/ in Glasgow-Pakistani high school girls, *in* 'Proceedings of the 17th International Congress of the Phonetic Sciences', Hong Kong, pp. 216–219.

Alexander, A., Botti, F., Dessimoz, D. & Drygajlo, A. (2004), 'The effect of mismatched recording conditions on human and automatic speaker recognition in forensic applciations', *Forensic Science International* **146**, S95–S99.

Alexander, A., Forth, O., Atreya, A. & Kelly, F. (2016), VOCALISE: A forensic automatic speaker recognition system supporting spectral, phonetic, and user-provided features, *in* 'Proceedings of Odyssey: the Speaker and Language Recognition workshop', Bilbao, Spain. Paper number 88.

Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G. & Vinyals, O. (2012), 'Speaker diarization: A review of recent research', *IEEE Transactions on Acoustics, Speech and Language Processing* **20**, 356–370.

Asher, J. & García, R. (1969), 'The optimal age to learn a foreign language', *The Modern Language Journal* **53**, 334–341.

Atkinson, N. (2015), Variable factors affecting voice identification in forensic contexts, PhD thesis, University of York, UK.

Bahari, M. H., Saeidi, R., Hamme, H. V. & Leeuwen, D. V. (2013), Accent recognition using i-vector, gaussian mean supervector and gaussian posterior probability supervector for spontaneous telephone speech, *in* 'Proceedings of the International Conference on Acoustics, Speech and Signal Processing', Vancouver, Canada, pp. 7344–7348.

Baranowski, M. & Turton, D. (2016), Manchester English, *in* R. Hickey, ed., 'Researching Northern English', John Benjamins, pp. 293–316.

Beck, J. & Schaeffler, F. (2015), Voice quality variation in Scottish adolescents: gender versus geography, *in* 'Proceedings of the 18th International Congress of Phonetic Sciences', Glasgow, UK. Paper number 737.

Behravan, H., Hautamäki, V. & Kinnunen, T. (2013), Foreign accent detection from spoken Finnish using i-vectors, *in* 'Proceedings of Interspeech', Lyon, France, pp. 79–82.

Behravan, H., Hautamäki, V. & Kinnunen, T. (2015), 'Factors affecting i-vector based foreign accent recognition: A case study in spoken Finnish', *Speech Communication* **66**, 118–129.

Bellman, R. (1957), *Dynamic programming*, Princeton University Press, Princeton, US.

Biadsy, F. & Hirschberg, J. (2009), Using prosody and phonotactics in Arabic dialect identifcation, *in* 'Proceedings of Interspeech', Brighton, UK, pp. 208–211.

Biadsy, F., Soltau, H., Mangu, L., Navratil, J. & Hirschberg, J. (2010), Discriminative phonotactics for dialect recognition using context-dependent

phone classifiers, *in* 'Proceedings of Odyssey: The Speaker and Language Recognition Workshop', Brno, Czech Republic, pp. 263–270.

Bilmes, J. (1988), 'A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian Mixture and Hidden Markov Models', *Technical Report from the International Computer Science Institute* . URL: http://melodi.ee.washington.edu/people/bilmes/mypapers/em.pdf [Accessed: 15/05/2017].

Boakye, K., Trueba-Hornero, B., Vinyals, O. & Friedland, G. (2008), Overlapped speech detection for improved speaker diarization in multiparty meetings, *in* 'Proceedings of the International Conference on Acoustics, Speech and Signal Processing', Las Vegas, US, pp. 4353–4356.

Bocklet, T. & Shriberg, E. (2009), Speaker recognition using syllable-based constraints for cepstral frame selection, *in* 'Proceedings of the International Conference on Acoustics, Speech and Signal Processing', Taiwan, pp. 4525–4528.

Boersma, P. & Weenink, D. (2017), 'Praat: doing phonetics by computer'. www.praat.org. [Accessed: 3/01/2017].

Bolón-Canedo, V., Sánchez-Maroño, N. & Alonso-Betanzos, A. (2015), 'Recent advances and emerging challenges of feature selection in the context of big data', *Knowledge-Based Systems* **86**, 33–45.

Botti, F., Alexander, A. & Drygajlo, A. (2004), 'On compensation of mismatched recording conditions in the Bayesian approach for forensic automatic speaker recognition', *Forensic Science International* **146**, S101–S106.

Broeders, A. (2001), Forensic speech and audio analysis forensic linguistics, *in* 'Proceedings of the 13th INTERPOL Forensic Science Symposium', Lyon, France, pp. 54–82.

Brown, G. & Wormald, J. (2017), 'Automatic sociophonetics: Exploring corpora using a forensic accent recognition system', *Journal of the Acoustical Society of America* **142**, 422–433.

Brümmer, N. (2007), 'FoCal Multi-Class: Toolkit for evaluation, fusion and calibration of multi-class recognition scores - tutorial and user manual'. Available at: https://sites.google.com/site/nikobrummer/focalmulticlass [Accessed: 17/04/2017].

Brümmer, N., Burget, L., Černocký, J., Glembek, O., František Grézl, Karafiát, M., van Leeuwen, D., Matějka, P., Schwarz, P. & Strasheim, A. (2007), 'Fusion of heterogenous speaker recognition systems in the STBU submission for the NIST Speaker Recognition Evaluation 2006', *IEEE Transactions on Audio, Speech and Language Processing* **15**, 2072–2084.

Brümmer, N. & du Preez, J. (2007), 'Application-independent evaluation of speaker detection', *Computer Speech and Language* **20**, 230–275.

Brümmer, N. & van Leeuwen, D. (2006), On calibration of language recognition scores, *in* 'Proceedings of Odyssey: The Speaker and Language Recognition Workshop', San Juan, Puerto Rico.

Burkhardt, F., Schüller, B., Weiss, B. & Weninger, F. (2011), Would you buy a car from me? - on the likability of telephone voices, *in* 'Proceedings of Interspeech', Florence, Italy, pp. 1557–1560.

Byrne, C. & Foulkes, P. (2004), 'The 'mobile phone' effect on vowel formants', *The International Journal of Speech, Language and the Law* **11**, 83–102.

Cambier-Langeveld, T. (2010), 'The role of linguists and native speakers in language analysis for the determination of speaker origin', *The International Journal of Speech, Language and the Law* **17**, 67–93.

Campbell-Kibler, K. (2010), 'Sociolinguistics and perception', *Language and Linguistics Compass* **4**, 377–389.

Canavan, A. & Zipperle, G. (1996), 'CallFriend corpus'. http://yki-korpusjyu.fi. Accessed: 07/05/2015.

Chen, T., Huang, C., Chang, E. & Wang, J. (2001), Automatic accent identification using Gaussian Mixture Models, *in* 'Proceedings of IEEE workshop on Automatic Speech Recognition and Understanding', Italy.

Cheshire, J., Kerswill, P., Fox, S. & Torgerson, E. (2011), 'Contact, the feature pool and the speech community: The emergence of Multicultural London English', *Journal of Sociolinguistics* **15**, 151–196.

Clopper, C. & Pisoni, D. (2004), 'Some acoustic cues for the perceptual categorization of American English regional dialects', *Journal of Phonetics* **32**, 111–140.

Clopper, C. & Smiljanic, R. (2011), 'Effects of gender and regional dialect on prosodic patterns in American English', *Journal of Phonetics* **39**, 237–245.

D'Arcy, S., Russell, M., Browning, S. & Tomlinson, M. (2004), The Accents of the British Isles (ABI), corpus, *in* 'Proceedings of Modélisations pour l'Identification des Langues', Paris, France, pp. 115–119.

Davis, S. & Mermelstein, P. (1980), 'Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences', *IEEE Transactions on Acoustics, Speech and Signal Processing* **28**, 357–366.

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

Dehak, N., Torres-Carrasquillo, P., Reynolds, D. & Dehak, R. (2011), Language recognition via i-vectors and dimensionality reduction, *in* 'Proceedings of Interspeech', Florence, Italy, pp. 857–860.

DeMarco, A. & Cox, S. (2012), Iterative classification of regional British accents in i-vector space, *in* 'Proceedings of Machine Learning in Speech and Language Processing', Portland, Oregon, USA, pp. 1–4.

Docherty, G. & Foulkes, P. (1999), Derby and Newcastle: instrumental phonetics and variationist studies, *in* P. Foulkes & G. Docherty, eds, 'Urban Voices: Accent Studies in the British Isles', Routledge, London, pp. 47–71.

Doddington, G., Liggett, W., Martin, A., Przybocki, M. & Reynolds, D. (1998), SHEEP, GOATS, LAMBS and WOLVES: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation, *in* 'Proceedings of the 5th International Conference of Spoken Language Processing', Sydney, Australia, pp. 1351–1354.

Dror, I. (2015), Cognitive neuroscience in forensic science: Understanding and utilising the human element, *in* 'Paper presented at The Paradigm Shift for UK Forensic Science meeting', The Royal Society, London, UK.

Dror, I., Charlton, D. & Péron, A. (2006), 'Contextual information renders experts vulnerable to making erroneous identifications', *Forensic Science International* **156**, 74–78.

Dror, I. & Hampikian, G. (2011), 'Subjectivity and bias in forensic DNA mixture interpretation', *Science and Justice* **51**, 204–208.

Drummond, R. (2012), 'Aspects of identity in a second language: ING variation in the speech of Polish migrants living in Manchester, UK', *Language Variation and Change* **24**, 107–133.

Drygajlo, A. (2007), 'Forensic automatic speaker recognition', *IEEE Signal Processing Magazine* pp. 132–135.

Drygajlo, A., Jessen, M., Gfroerer, S., Wagner, I., Vermeulen, J. & Niemi, T. (2016), 'Methodological guidelines for best practice in forensic semiautomatic and automatic speaker recognition'. Report published by the European Network of Forensic Science Institutes.

Duckworth, M., McDougall, K., de Jong, G. & Shockey, L. (2011), 'Improving the consistency of formant measurement', *International Journal of Speech, Language and the Law* **18**, 35–51.

Ellis, S. (1994), 'The Yorkshire Ripper Enquiry: Part 1', *Forensic Linguistics* **1**, 197–206.

Fecher, N. (2014), Effects of forensically-relevant facial concealment on acoustic and perceptual properties of consonants, PhD thesis, University of York, UK.

Ferragne, E. & Pellegrino, F. (2007), Automatic dialect identification: A study of British English, *in* C. Müller, ed., 'Speaker Classification', Vol. 2

of *Lecture Notes in Computer Science*, Springer-Verlag, Berlin Heidelberg, pp. 243–257.

Ferragne, E. & Pellegrino, F. (2010), 'Vowel systems and accent similarity in the British Isles: Exploiting multidimensional acoustic distances in phonetics', *Journal of Phonetics* **38**, 526–539.

Flege, J. E., Yeni-Komshian, G. & Liu, S. (1999), 'Age constraints on second-language acquisition', *Journal of Memory and Language* **41**, 78–104.

Foulkes, P. & French, P. (2012), Forensic speaker comparison: A linguistic-acoustic perspective, *in* P. Tiersma & L. Solan, eds, 'The Oxford Handbook of Language and Law', Oxford Handbooks in Linguistics, Oxford University Press, Oxford, pp. 557–572.

Foulkes, P., French, P. & Wilson, K. (2018), LADO as forensic speaker profiling, *in* P. Patrick, M. Schmid & K. Zwaan, eds, 'Language Analaysis for the Determination of Origin', Springer.

Foulkes, P., Scobbie, J. & Watt, D. (2010), Sociophonetics, *in* W. Hardcastle, J. Laver & F. Gibbon, eds, 'The Handbook of Phonetic Sciences', 2nd edn, Blackwell Handbooks in Linguistics, Wiley-Blackwell, Oxford, pp. 703–754.

Foulkes, P. & Wilson, K. (2011), Language analysis for the determination of origin, *in* 'Proceedings of the 17th International Congress of Phonetic Sciences', Hong Kong, pp. 692–694.

Franco-Pedroso, J. & Gonzalez-Rodriguez, J. (2016), 'Linguistically-constrained formant-based i-vectors for automatic speaker recognition', *Speech Communication* **76**, 61 – 81.

Fraser, H. (2015), 'Transcription of indistinct forensic recordings: Problems and solutions from the perspective of phonetic science', *Language and Law/Linguagem E Direito* **1**, 5–21.

Fraser, H. (2017), Forensic transcription: Where to from here to create a better system for handling of indistinct covert recordings?, *in* 'Paper presented at the International Association for Forensic Phonetics and Acoustics conference', Split, Croatia.

French, P. & Harrison, P. (2007), 'Position statement concerning use of impressionistic likelihood terms in forensic speaker comparison cases', *The International Journal of Speech, Language and the Law* **14**, 137–144.

French, P., Harrison, P., Kirchhübel, C., Rhodes, R. & Wormald, J. (2017), From receipt of recordings to dispatch of report: Opening the blinds on lab practices, *in* 'Paper presented at the International Association for Forensic Phonetics and Acoustics conference', Split, Croatia.

French, P., Harrison, P. & Lewis, J. W. (2006), 'R v John Samuel Humble: The Yorkshire Ripper Hoaxer Trial', *The International Journal of Speech, Language and the Law* **13**, 255–273.

French, P., Nolan, F., Foulkes, P., Harrison, P. & McDougall, K. (2010), 'The UK position statement on forensic speaker comparison: a rejoinder to Rose and Morrsion', *The International Journal of Speech, Language and the Law* **17**, 143–152.

Furui, S. (1997), Recent advances in speaker recognition, *in* 'Proceedings of the First Conference on Audio- and Video-based Biometric Person Authentification', Crans-Montana, Switzerland, pp. 237–252.

Garcia-Romero, D. & Espy-Wilson, C. (2011), Analysis of i-vector length normalisation in speaker recognition systems, *in* 'Proceedings of Interspeech', Florence, Italy.

Gardner, R. C. (2007), 'Motivation and second language acquisition', *Porta Linguarum* **8**, 9–20.

Gauvain, J.-L. & Lee, C.-H. (1994), 'Maximum a-posteriori estimation for multivariate gaussian mixture observations of Markov chains', *IEEE Transactions on Speech and Audio Processing* **2**, 291–298.

Giles, H. & Billings, A. (2004), Assessing language attitudes: Speaker evaluation studies, *in* A. Davies & C. Elder, eds, 'The Handbook of Applied Lingusitics', Blackwell Handbooks in Linguistics, Blackwell, Oxford, UK, pp. 187–209.

Gold, E. & French, P. (2011), 'International practices in forensic speaker comparison', *The International Journal of Speech, Language and the Law* **18**, 293–307.

Goldman, J.-P. (2011), EasyAlign: an automatic phonetic alignment tool under Praat, *in* 'Proceedings of Interspeech', Florence, Italy, pp. 3233–3236.

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

Grabe, E. (2004), Intonational variation in urban dialects of English spoken in the British Isles, *in* P. Gilles & J. Peters, eds, 'Regional Variation in Intonation', Linguistische Arbeiten, Tuebingen, pp. 9–31.

Greenberg, C., Martin, A., Brandschain, L., Campbell, J., Cieri, C., Doddington, G. & Godfrey, J. (2010), Human assisted speaker recognition in NIST SRE10, *in* 'Proceedings of Odyssey: The Speaker and Language Recognition Workshop', Brno, Czech Republic, pp. 180–185.

Greenberg, S. (1999), 'Speaking in shorthand - a syllable-centric perspective for understanding pronunciation variation', *Speech Communication* **29**, 159–176.

Guyon, I. & Elisseeff, A. (2003), 'An introduction to variable and feature selection', *Journal of Machine Learning Research* **3**, 1157–1182.

Haddican, W., Foulkes, P., Hughes, V. & Richards, H. (2013), 'Interaction of social and linguistic constraints of two vowel changes in Northern England', *Language, Variation and Change* **25**, 371 – 403.

Hall-Lew, L., Friskney, R. & Scobbie, J. (in press), 'Accommodation or political identity: Scottish members of the UK parliament', *Language variation and change* .

Hanani, A., Russell, M. & Carey, M. (2011), Computer and human recognition of regional accents of British English, *in* 'Proceedings of Interspeech', Florence, Italy, pp. 27–31.

Hanani, A., Russell, M. & Carey, M. (2013), 'Human and computer recognition of regional accents and ethnic groups from British English speech', *Computer Speech and Language* **27**, 59–74.

Harrison, P. (2013), Making Accurate Formant Measurements: An Empirical Investigation of the Influence of the Measurement Tool, Analysis Settings

and Speaker on Formant Measurements, PhD thesis, University of York, UK.

Hasan, T., Saeidi, R., Hansen, J. H. L. & van Leeuwen, D. (2013), Duration mismatch compensation for i-vector based speaker recognition systems, *in* 'Proceedings of the International Conference on Acoustics, Speech and Signal Processing', Vancouver, Canada, pp. 7663–7667.

He, Q., Wornell, G. & Ma, W. (2016), A low-power text-dependent speaker verification system with narrow-band feature pre-selection and weighted dynamic time warping, *in* 'Proceedings of Odyssey: The Speaker and Language Recognition Workshop', Bilbao, Spain.

Helgason, P., Ringen, C. & Suomi, K. (2013), 'Swedish Quantity: Central Standard Swedish and Fenno-Swedish', *Journal of Phonetics* **41**, 534–545.

Herd, W., Jongman, A. & Sereno, J. (2010), 'An acoustic perceptual analysis of /t/ and /d/ flaps in American English', *Journal of Phonetics* **38**, 504–516.

Hermansky, H. (1990), 'Perceptual Linear Predictive (PLP) analysis of speech', *Journal of the Acoustic Society of America* **87**, 1738–1752.

Holmes, J. & Holmes, W. (2001), *Speech Synthesis and Recognition*, 2nd edn, Taylor and Francis, New York.

Howell, P. & Kadi-Hanifi, K. (1991), 'Comparison of prosodic properties between read and spontaneous speech material', *Speech Communication* **10**, 163–169.

Huckvale, M. (2004), ACCDIST: a metric for comparing speakers' accents, *in* 'Proceedings of the International Conference on Spoken Language Processing', Jeju, Korea, pp. 29–32.

Huckvale, M. (2007), ACCDIST: An accent similarity metric for accent recognition and diagnosis, *in* C. Müller, ed., 'Speaker Classification', Vol. 2 of *Lecture Notes in Computer Science*, Springer-Verlag, Berlin Heidelberg, pp. 258–274.

Hughes, A., Trudgill, P. & Watt, D. (2012), *English Accents and Dialects*, 5 edn, Hodder, London.

Hughes, V. (2014), The definition of the relevant population an the collection of data for likelihood ratio-based forensic voice comparison, PhD thesis, University of York, UK.

Hughes, V. (2017), 'Sample size and the multivariate kernel density likelihood ratio: how many speakers are enough?', *Speech Communication* **94**, 15–29.

Hughes, V. & Foulkes, P. (2015), 'The relevant population in forensic voice comparison: Effects of varying delimitations of social class and age', *Speech Communication* **66**, 218–230.

Hughes, V. & Foulkes, P. (2016), Speaker- and group-specific information in formant dynamics: a forensic perspective, *in* 'Paper presented at LabPhon 15 Satellite Workshop: Speech dynamics, social meaning and phonological categories', Ithaca, USA.

Hughes, V. & Wormald, J. (2017), Assessing typicality in forensic voice comparison, *in* 'Paper presented at the 2nd Innovative Methods in Sociophonetics Workshop', Edinburgh, UK.

Humphries, J. & Woodland, P. (1997), Using accent-specific pronunciation modelling for improved large vocabulary continuous speech recognition, *in* 'Proceedings of Eurospeech', Rhodes, Greece, pp. 2367–2370.

Jansen, S. (2013), "I don't sound like a Geordie!": Phonological and morphosyntactic aspects of Carlisle English, *in* N.-L. Johannesson, G. Melchers & B. Björkman, eds, 'Of butterflies and birds, of dialects and genres: Essays in honour of Philip Shaw', pp. 209–224.

Jenkins, M. (2016), Identifying an imitated accent: Humans vs. computers, MSc dissertation, University of York, UK.

Jobling, M. & Gill, P. (2004), 'Encoded evidnce: DNA in forensic analysis', *Nature Reviews Genetics* **5**, 739–752.

Johnson, J. & Newport, E. (1989), 'Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language', *Cognitive Pyschology* **21**, 60–99.

Junqua, J.-C., Fincke, S. & Field, K. (1999), The Lombard Effect: A reflex to better communicate with others in noise, *in* 'Proceedings of the International Conference on Acoustics, Speech and Signal Processing', pp. 2083–2086.

Jurafsky, D. & Martin, J. (2009), *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*, 2nd edn, Pearson Education International, New Jersey.

Kahn, J., Audibert, N., Rossato, S. & Bonastre, J.-F. (2010), Intra-speaker variability effects on speaker verification performance, *in* 'Proceedings of Odyssey: The Speaker and Language Recognition Workshop', Brno, Czech Republic, pp. 109–116.

Kajarekar, S. S., Scheffer, N., Graciarena, M., Shriberg, E., Stolcke, A., Ferrer, L. & Bocklet, T. (2009), The SRI NIST 2008 speaker recognition evaluation

system, *in* 'Proceedings of the International Conference on Acoustics, Speech and Signal Processing', Taipei, Taiwan, pp. 4205–4208.

Kassin, S., Dror, I. & Kuckucka, J. (2013), 'The forensic confirmation bias: Problems, perspectives, and proposed solutions', *Journal of Applied Research in Memory and Cognition* **2**, 42–52.

Kelly, F. (2014), Automatic Recognition of Ageing Speakers, PhD thesis, Trinity College Dublin, Ireland.

Kenny, P., Boulianne, G., Ouellet, P. & Dumouchel, P. (2007), 'Speaker and session variability in gmm-based speaker verification', *IEEE Transactions in Audio, Speech and Language Processing* pp. 1–14.

Keogh, E. & Mueen, A. (2011), Curse of dimensionality, *in* C. Sammut & G. Webb, eds, 'Encyclopedia of Machine Learning', Springer, US, pp. 257–258.

Kinnunen, T. & Li, H. (2010), 'An overview of text-independent speaker recognition: from features to supervectors', *Speech Communication* **52**, 12–40.

Knowles, G. (1973), Scouse: the urban dialect of Liverpool, PhD thesis, University of Leeds, UK.

Kuhn, T. S. (1962), *The Structure of Scientific Revolutions*, University of Chicago Press, Chicago, IL.

Künzel, H., Gonzalez-Rogriguez, J. & Ortega-Garcia, J. (2004), Effect of voice disguise on the performance of a forensic automatic speaker recognition system, *in* 'Proceedings of Odyssey: The Speaker and Language Recognition Workshop', Toledo, Spain.

Labov, W. (1963), 'The social motivation of language change', *Word* **19**, 273–309.

Labov, W. (1966), *The Social Stratification of English in New York City*, Center for Applied Linguistics, Washington DC.

Larcher, A., Lee, K. A., Ma, B. & Li, H. (2014), 'Text-dependent speaker verification: Classifiers, databases and RSR2015', *Speech Communication* **60**, 56–77.

Laver, J. (1980), *The Phonetic Description of Voice Quality*, Cambridge University Press, Cambridge.

Lei, Y., Scheffer, N., Ferrer, L. & McLaren, M. (2014), A novel scheme for speaker recognition using a phonetically-aware deep neural network, *in* 'Proceedings of the IEEE International Conference for Acoustics, Speech and Signal Processing', Florence, Italy, pp. 1714–1718.

Lenneberg, E. (1967), *Biological Foundations of Language*, Wiley, New York.

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

Liu, G. & Hansen, J. (2011), A systematic strategy for robust automatic dialect identification, *in* 'Proceedings of the 19th European Signal Processing Conference', Barcelona, Spain, pp. 2138–2141.

Llamas, C. (2010), Convergence and divergence along a national border, *in* C. Llamas & D. Watt, eds, 'Language and Identities', Edinburgh University Press, Edinburgh, pp. 227–236.

Llamas, C., Watt, D. & Johnson, D. E. (2009), 'Linguistic accommodation and the salience of national identity markers in a border town', *Journal of Language and Social Psychology* **28**, 381–207.

Llamas, C., Watt, D. & MacFarlane, A. E. (2016), 'Estimating the relative sociolinguistic salience of segmental variables in a dialect boundary zone', *Frontiers in Psychology* **7**.

Mack, S. & Munson, B. (2012), 'The influence of /s/ quality of ratings of men's sexual orientation: Explicit and implicit measures of the 'gay lisp' stereotype', *Journal of Phonetics* **40**, 198–212.

Mandasari, M. I., McLaren, M. & van Leeuwen, D. (2012), The effect of noise on modern automatic speaker recognition systems, *in* 'Proceedings of the International Conference on Acoustics, Speech and Signal Processing', Kyoto, Japan, pp. 4249–4252.

Martin, A., Doddington, G., Kamm, T., Ordowski, M. & Przybocki, M. (1997), The DET curve in assessment of detection task performance, *in* 'Proceedings of the European Conference on Speech Communication and Technology', pp. 1895–1898.

McAllister, R., Flege, J. & Piske, T. (2002), 'The influence of L1 on the acquisition of Swedish quantity by native speakers of Spanish, English and Estonian', *Journal of Phonetics* **30**, 229–258.

McDougall, K. (2013), 'Assessing perceived voice similarity using Multidimensional Scaling for the construction of voice parades', *The International Journal of Speech, Language and the Law* **20**, 163–172.

McDougall, K. & Nolan, F. (2007), Discrimination of speakers using the formant dynamics of /u:/ in British English, *in* 'Proceedings of the 16th International Congress of Phonetic Sciences', Saarbrücken, Germany, pp. 1825–1828.

Mohammadi, G. & a Vinciarelli (2012), 'Automatic personality perception: prediction of trait attribution based on prosodic features', *IEEE Transactions on Affective Computing* **3**, 273–284.

Montgomery, M. (2001), British and Irish antecedents, *in* 'The Cambridge History of the English Language', Cambridge University Press, Cambridge, pp. 86–153.

Morrison, G. S. (2013), 'Tutorial on logistic-regression calibration and fusion: converting a score to a likelihood ratio', *Australian Journal of Forensic Sciences* **45**, 173–197.

Müller, F. & Mertins, A. (2011), Noise robust speaker-independent speech recognition with invariant-integration features using power-bias subtraction, *in* 'Proceedings of Interspeech', Florence, Italy, pp. 1677–1680.

Najafian, M., DeMarco, A., Cox, S. & Russell, M. (2014), Unsupervised model selection for recognition of regional accented speech, *in* 'Proceedings of Interspeech', Singapore, pp. 2967–2971.

Najafian, M., Safavi, S., Weber, P. & Russell, M. (2016), Identification of British English regional accent using fusion of i-vector and multi-accent phonotactic systems, *in* 'Proceedings of Odyssey: The Speaker and Language Recognition Workshop', Bilbao, Spain.

Nerbonne, J. (2009), 'Data-driven dialectology', *Language and Linguistic Compass* **3**, 175–198.

Nerbonne, J. & Kleiweg, P. (2007), 'Toward a dialectological yardstick', *Journal of Quantitative Linguistics* **14**, 148–166.

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

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

Nolan, F., McDougall, K., Jong, G. D. & Hudson, T. (2009), 'The DyViS Database: style-controlled recordings of 100 homogenous speakers for forensic phonetic research', *The International Journal of Speech, Language and the Law* **16**, 31–57.

Peppé, S., Maxim, J. & Wells, B. (2000), 'Prosodic variation in Southern British English', *Language and Speech* **43**, 309–334.

Piske, T., MacKay, I. & Flege, J. (2001), 'Factors affecting degree of foreign accent in an L2: a review', *Journal of Phonetics* **29**, 191–215.

Poblete, V., Espic, F., King, S., Stern, R., Huenupán, F., Fredes, J. & Yoma, N. B. (2015), 'A perceptually-motivated low-complexity instantaneous linear channel normalization technique applied to speaker verification', *Computer Speech and Language* **31**, 1–27.

Pohjalainen, J., Räsänen, O. & Kadioglu, S. (2015), 'Feature selection methods and their combination in high-dimensional classification of speaker likability, intelligibility and personality traits', *Speech Communication* **29**, 145–172.

Przybocki, M. & Martin, A. (2004), NIST Speaker Recognition Evaluation chronicles, *in* 'Proceedings of Odyssey: The speaker and language recognition workshop', Toledo, Spain.

Rajan, P., Kinnunen, T. & Hautamäki, V. (2013), Effect of multicondition training on i-vector PLDA configurations for speaker recognition, *in* 'Proceedings of Interspeech', Lyon, France, pp. 3694–3697.

Ramos-Castro, D., Gonzalez-Rodriguez, J. & Ortega-Garcia, J. (2006), Likelihood ratio calibration in a transparent and testable forensic speaker recognition framework, *in* 'Proceedings of Odyssey: The Speaker and Language Recognition Workshop', San Juan, Puerto Rico.

Reynolds, D. & Rose, R. (1995), 'Robust text-independent speaker identification using gaussian mixture speaker models', *IEEE Transactions of Speech and Audio Processing* **3**, 72–83.

Rhodes, R. (2012), Assessing the strength of non-contemporaneous forensic speech evidence, PhD thesis, University of York, UK.

Rhodes, R. (2016), Cognitive bias in forensic speech science: a survey on risks and proposed safeguards, *in* 'Paper presented at the International Association for Forensic Phonetics and Acoustics conference', York, UK.

Richardson, F. & Campbell, W. (2008), Language recognition with discriminative keyword selection, *in* 'Proceedings of the IEEE International Conference for Acoustics, Speech and Signal Processing', pp. 4145–4148.

Rong, J., Li, G. & Chen, Y.-P. P. (2009), 'Information processing and management', *Information Processing and Management* **45**, 315–328.

Rose, P. (2002), *Forensic Speaker Identification*, Forensic Science Series, Taylor Francis, London.

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

Rubin, J. (1975), 'What the "Good Language Learner" can teach us', *TESOL Quarterly* **9**, 41–51.

Sadjadi, S. O., Slaney, M. & Heck, L. (2013), 'MSR identity toolbox v1.0: A MATLAB toolbox for speaker-recognition research', *Speech and Language Processing Technical Committee Newsletter, IEEE* .

Saeys, Y., Abeel, T. & de Peer, Y. V. (2008), Robust feature selection using ensemble feature selection techniques, *in* 'European Conference on Machine Learning and Knowledge Discovery in Databases, Part II', Berlin, Germany, pp. 313–325.

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

Sankoff, G. & Blondeau, H. (2007), 'Language change across the lifespan: /r/ in Montreal French', *Language* **83**, 560–588.

Scheffer, N., Ferrer, L., Graciarena, M., Kajarekar, S., Shriberg, E. & Stolcke, A. (2011), The SRI NIST 2010 Speaker Recognition Evaluation system, *in* 'Proceedings of the International Conference on Acoustics, Speech and Signal Processing', Prague, Czech Republic.

Schilling, N. & Marsters, A. (2015), 'Unmasking Identity: Speaker profiling for forensic linguistic purposes', *Annual Review of Applied Linguistics* **35**, 195–214.

Schüller, B., Rigoll, G. & Lang, M. (2004), Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine - belief network architecture, *in* 'Proceedings of the International Conference on Acoustics, Speech and Signal Processing', Montreal, Canada, pp. 577–580.

Shriberg, E. (2007), Higher-level features in speaker recognition, *in* C. Muller, ed., 'Speaker Classification', Vol. 1 of *Lecture Notes in Computer Science*, Springer-Verlag, Berlin Heidelberg, pp. 241–259.

Singh, R., Gencaga, D. & Raj, B. (2016), Formant manipulations in voice disguise by mimicry, *in* 'Proceedings of the 4th International Workshop on Biometrics and Forensics (IWBF)', Limassol, Cyprus.

Sproat, R. & Fujimura, O. (1993), 'Allophonic variation in english /l/ and its implications for phonetic implementation', *Journal of Phonetics* **21**, 291–311.

Stuart-Smith, J. (1999), Glasgow: accent and voice quality, *in* P. Foulkes & G. Docherty, eds, 'Urban Voices: Accent Studies in the British Isles', Routledge, London, pp. 203–222.

Suh, J.-W. & Hansen, J. H. L. (2012), 'Acoustic hole filling for sparse enrollment data using a cohort universal corpus for speaker recognition', *Journal of the Acoustical Society of America* **131**, 1515–1528.

Thomas, S., Seltzer, M., Church, K. & Hermansky, H. (2013), Deep Neural Network features and semi-supervised training for low resource speech recognition, *in* 'Proceedings of the International Conference on Acoustics, Speech and Signal Processing', Vancouver, Canada, pp. 6704–6708.

Toor, A. (2017), 'Germany to use voice analysis software to help determine where refugees come from'. Accessed 28/06/2017.
URL: *https://www.theverge.com/2017/3/17/14956532/germany-refugee-voice-analysis-dialect-speech-software*

Torres-Carrasquillo, P., Singer, E., Kohler, M., Greene, R., Reynolds, D. & Jr, J. D. (2002), Approaches to language identification using Gaussian Mixture Models and Shifted Delta Cepstral features, *in* 'Proceedings of the International Conference on Spoken Language Processing', pp. 89–92.

Tranter, S. & Reynolds, D. (2006), 'An overview of automatic speaker diarization systems', *IEEE Transactions on Audio, Speech and Language Processings* **14**, 1557–1565.

Tully, G. (2016), Forensic Science Regulator Codes of Practice and Conduct: for forensic science providers and practitioners in the criminal justice system, Technical report, The UK Government. https://www.gov.uk/government/publications/forensic-science-providers-codes-of-practice-and-conduct-2016.

Tully, G. (2017), Forensic Science Regulator Annual Report, Technical report, The UK Government. https://www.gov.uk/government/publications/forensic-science-regulator-annual-report-2016.

van der Molen, L., van Rossum, M., Jacobi, I., van Son, R., Smeele, L., Rasch, C. & Hilgers, F. (2012), 'Pre- and posttreatment voice and speech outcomes in patients with advanced head and neck cancer treated with chemoradiotherapy: Expert listeners' and patients' perception', *Journal of Voice* **26**, 25–33.

van Leeuwen, D. & Brümmer, N. (2007), An introduction to application-independent evaluation of speaker recognition systems, *in* C. Müller, ed., 'Speaker Classification', Vol. 2 of *Lecture Notes in Computer Science/Artificial Intelligence*, Springer, Heidelberg, New York, Berlin.

van Leeuwen, D. & Brümmer, N. (2008), Building language detectors using small amounts of training data, *in* 'Proceedings of Odyssey: The Speaker and Language Recognition Workshop', Stellenbosch, South Africa.

Vapnik, V. (1998), *Statistical Learning Theory*, Wiley, New York.

Vergyri, D., Lamel, L. & Gauvain, J.-L. (2010), Automatic speech recognition of multiple accented English data, *in* 'Proceedings of Interspeech', Makuhari, Chiba, Japan, pp. 1652–1655.

Vieru, B., de Mareüil, P. B. & Adda-Decker, M. (2011), 'Characterisation and identification of non-native French accents', *Speech Communication* **53**, 292–310.

Vogt, T. & André, E. (2005), Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition, *in* 'Proceedings of Multimedia and Expo', Amsterdam, The Netherlands, pp. 474–477.

Watt, D. (2000), 'Phonetic parallels between the close-mid vowels of Tyneside

English: Are they internally or externally motivated', *Language Variation and Change* **12**, 69 – 101.

Watt, D. (2002), ''I don't speak with a Geordie accent, I speak, like, the Northern accent': Contact-induced levelling in the Tyneside vowel system', *Journal of Sociolinguistics* **6**, 44–63.

Watt, D. (2010), The identification of the individual through speech, *in* C. Llamas & D. Watt, eds, 'Language and Identities', Edinburgh University Press, Edinburgh, pp. 76–85.

Watt, D. & Allen, W. (2003), 'Tyneside English', *Illustrations of the IPA: Journal of the International Phonetic Association* **33**, 267–271.

Watt, D., Llamas, C., French, P., Braun, A. & Robertson, D. (2016), Forensic aspects of spectral and durational variability in English schwa at the individual, community and regional levels, *in* 'Paper presented at the International Association for Forensic Phonetics and Acoustics conference', York, UK.

Watt, D., Llamas, C. & Johnson, D. E. (2014), Sociolinguistic variation on the Scottish-English Border, *in* R. Lawson, ed., 'Sociolinguistics in Scotland', Palgrave Macmillan, London, pp. 79–102.

Watt, D. & Milroy, L. (1999), Patterns of variation and change in three newcastle vowels: is this dialect levelling?, *in* P. Foulkes & G. Docherty, eds, 'Urban Voices: Accent Studies in the British Isles', Routledge, London.

Wells, J. (1982), *Accents of English 2*, Cambridge University Press, Cambridge.

Wood, S., Hughes, V. & Foulkes, P. (2014), Filled pauses as variables in speaker comparison: dynamic formant analysis and duration measurements improve

performance for 'um', *in* 'Paper presented at the International Association for Forensic Phonetics and Acoustics conference', Zurich, Switzerland.

Wormald, J. (2016), Regional Variation in Panjabi-English, PhD thesis, University of York, UK.

Wu, T., Duchateau, J., Martens, J.-P. & Compernolle, D. V. (2010), 'Feature subset selection for improved native accent identification', *Speech Communication* **2**, 83–98.

Xue, S. & Hao, G. (2003), 'Changes in the human vocal tract due to aging and the acoustic correlates of speech production: A pilot study', *Journal of Speech, Language and Hearing Research* **46**, 689–701.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V. & Woodland, P. (2009), *The HTK Book for HTK Version 3.4*, Cambridge University Engineering Department, Cambridge.

Zapf, P. & Dror, I. (2017), 'Understanding and mitigating bias in forensic evaluation: Lessons from forensic science', *International Journal of Forensic Mental Health* pp. 1–12.

Zheng, Y., Sproat, R., Gu, L., Shafran, I., Zhou, H., Su, Y., Jurafsky, D., Starr, R. & Yoon, S.-Y. (2005), Accent detection and speech recognition for Shanghai-accented Mandarin, *in* 'Proceedings of Interspeech', Lisbon, Portugal, pp. 217–220.

Zissman, M. (1996), 'Comparison of four approaches to automatic language identification of telephone speech', *IEEE Transactions on Speech and Audio Processing* **4**, 31–44.

Zue, V. & Laferriere, M. (1979), 'Acoustic study of medial /t,d/ in American English', *Journal of the Acoustical Society of America* **66**, 1039–1050.