Y-ACCDIST: An Automatic Accent Recognition System for Forensic Applications

Georgina Alice Brown

Master of Arts (by Research)

University of York

Language and Linguistic Science

September 2014

Abstract

This thesis introduces and explores the performance of the Y-ACCDIST system (the York ACCDIST automatic accent recognition system). Based on the ACCDIST metric (Huckvale, 2004), it is a newly developed accent recognition system intended for forensic applications. Accent has received a lot of research attention within speech technology as it is often to blame for automatic speech recognition errors. A lot of research has therefore targeted automatic accent recognition while taking the automatic speech recognition application into account. Little has been done, however, to research automatically recognising speakers' accents for forensic purposes. Such a task might involve identifying speaker properties (e.g. geographical origin) if no suspect is in the frame for making an incriminating telephone call.

The Y-ACCDIST system is applied to the forensic context in two main ways. Firstly, it is applied to geographically-proximate accents, where a predicted increase in similarity between varieties exists. This accent recognition task is therefore expected to be more difficult than tasks in previous studies. Secondly, the model is adapted in such a way which makes it possible to process spontaneous speech, instead of just highly comparable speech content (i.e. read prompts). The present thesis shows accent recognition results, distinguishing between four geographically-proximate accents, of up to 90%. Accent recognition results of spontaneous speech are lower (up to 59.3%). However, light is shed on clear research directions aiming to improve this result.

Contents

List of Figures

List of Tables

Acknowledgements

Firstly, I would like to thank my supervisor, Dom Watt, for giving me the guidance and freedom needed to complete this thesis. His encouragement along with valuable discussion have maintained the project's momentum and I look forward to continuing what has been started.

I would like to thank Paul Foulkes and Peter French for their useful advice and suggestions, particularly in the earlier stages, which have helped to shape the project.

I would also like to thank Tamar Keren-Portnoy, Marilyn Vihman and Eytan Zweig who granted me the opportunity to work on their research projects while conducting my own. In doing this, they gave me enjoyable and valuable research experience while simultaneously helping to fund the project presented here. Without these opportunities, completing this thesis would have been very difficult indeed.

Declaration

This is to declare that all work presented in this thesis is my own and other sources have been acknowledged. The work in this thesis has not been submitted for any other degree other than Master of Arts (by research) at the University of York.

Introduction

The overarching goal of this thesis is to explore the performance of the York ACCDIST automatic accent recognition system (the Y-ACCDIST system) with a forensic application in mind. This section aims to offer a background of the motivations behind the system within the field of forensic speech science. It also outlines the research directions of this thesis and why they are important steps to take for the system's forensic application.

Background

Literature on linguistic variation is vast and yet accent still poses problems for automatic speech recognition applications. If we take two speakers of English, where one is a speaker of Southern British English and the other a speaker of Glasgow Standard English, it may be problematic if the two speakers were to produce the word *pot* to an automatic speech recognition system. It could be very easy for the Glasgow Standard English speaker to have been mistaken for saying *port* due to the Glasgow Standard English vowel system (Stuart-Smith, 1999: 206). This, of course, is just one example of the repercussions brought about by accent variation and countless instances exist. For automatic speech recognisers, accounting for the variation is a key consideration. There is a large body of research committed to developing more effective accent adaptation strategies in the hope of improving speech recognition performance. Accent clearly exists as a major obstacle for automatic speech recognition and it is blamed for the significant drops in speech recognition success rates. Teixeira *et al* (1997) show a detrimental difference of around 15% when speakers with non-native accents of English were tested on a speech recognition system which was trained on native British English varieties.

The forensic application, however, can take advantage of this variation in the population. The area of forensic speech science brings together a number of subdisciplines (such as phonetics, statistics and acoustics) to assist in legal casework. Most casework tasks fall under the heading of *speaker comparison* (once known as 'speaker identification') (Broeders, 2001). This involves taking two speech samples and arriving at the likelihood of the two samples belonging to the same speaker. A typical case would entail comparing a police interview recording with an incriminating telephone call for example. For some cases, by exploiting accent-specific features, forensic speech scientists can offer information regarding a speaker's geographical origin which may be of value in a legal setting. Identifying properties of unknown speakers in this way is referred to as *speaker profiling.* Within the nature of forensic science, the specific application of speaker profiling is undoubtedly varied, but to set the scene, it is a relatively common task in kidnapping cases (Foulkes and French, 2001: 330). In these kinds of cases the offender may make a ransom telephone call which could subsequently be used as evidence. Gathering as much information about the culprit as possible can assist investigative teams to home in on specified populations, which in turn can respond to a case's urgency. Another typical scenario may be the profiling of masked robbers where speech samples have been captured by CCTV technology (French and Harrison, 2006: 248). However, the list of possible scenarios is of course endless in line with the unpredictable subfield. The present thesis seeks to develop an automatic accent recognition system which could objectively provide accent information to assist on such cases.

More specific cases have been reported in the literature. The 'Yorkshire Ripper' case is famous both nationally and within forensic speech science. Throughout the late 1970s, a number of women were murdered in Yorkshire. In 1979, George Oldfield, leading the investigation to find the serial killer responsible, was sent a cassette tape with a recorded message from a man claiming to be the 'Yorkshire Ripper'. At this point, Stanley Ellis, an expert in dialects, was brought in to offer expertise on the speaker's geographical origin. From the recording, it is likely that listeners could comfortably guess that the speaker was from the North-East of England. As a specialist, Ellis narrowed down the speaker's accent to be spoken in the Southwick or Castletown areas of Sunderland (Ellis, 1994: 202). It later became apparent that the recording was in fact of a hoaxer. However, when the hoaxer was identified much later in 2006, it turned out that the recorded speaker was indeed from the Wearside area of Sunderland near Ellis' estimations.

As an example of accent information used to provide evidence to either eliminate or confirm a suspect's involvment in a crime, Ash (1988) reports on a case she worked on. For this case, the police had four recorded telephone calls of false claims of bombs and a fire at one particular company premises. The police had a suspect for the offence and Ash was asked to provide her expertise. With a strong focus on the vowel system, Ash was able to classify the suspect as a typical speaker of white Philadelphia vernacular accent, while commenting that his accent aligned with this variety 'in every detail' (Ash 1988: 27). By plotting formant values in a vowel space, Ash clearly shows that the offensive caller and the suspect had very different vowel systems, therefore recommending that they belong to different speech communities (Ash, 1988: 29-31). By confirming the speaker's accent and demonstrating its difference from the incriminating data, Ash's analysis provided compelling evidence in support of the defence.

The roles of the above individuals in their respective cases are instances of expert analysts in the area of speech science. The rise in technology has led to a rise in demand for forensic speech experts to assist in legal cases. Within the bounds of the broader domain of forensic science, Dror *et al* (2013) and Kassin *et al* (2013) have challenged the reliability of expert witnesses as human beings subject to 'psychological contamination' (Dror *et al*, 2013: 79; Kassin *et al*, 2013: 48). The researchers in these papers are concerned with what is termed *forensic confirmation bias* (Kassin *et al*, 2013: 45), which is determined by a combination of the individual's prior beliefs, motivations, exposure to irrelevant but potentially biassing case information, etc. Although these witnesses are called upon as experts, Dror *et al* and Kassin *et al* highlight that forensic examiners are human and the effects of bias can seep into their conclusions. Kassin *et al* (2013) offer some recommendations as to how biassing factors can be reduced. Suggestions include developing ‘blind testing’ protocol to avoid exposure to irrelevant information and also the use of technology (Kassin *et al*, 2013: 49). Rhodes (2014) brings the issue of forensic confirmation bias to the surface with direct reference to forensic speech science. Rhodes highlights the fact that forensic speech scientists utilise a perceptual mechanism for their analysis, as well as the issue that the data being analysed often holds excess contextual information which may contribute to the examiner’s beliefs (e.g. a police interview recording). Psychological contamination does not pose an issue for a technological tool. One of the main advantages of incorporating technology is its objectivity in deriving an outcome given data. However, Kassin *et al* (2013: 49) correctly point out that technology could have a negative biassing effect itself. Of course, technology is not resistant to error and outcomes from a computational tool (erroneous or not) could add to a forensic practitioner’s false prior beliefs during an analysis. This is an issue for consideration when forensic examiners and technology come together in casework analysis. Guidelines should be put in place before a combination of methods are put into practice. If developed and used carefully, technology could offer reliable and objective information.

The introduction of new technologies to the forensic domain should not of course be welcomed without caution. Eriksson and Lacerda (2007) bring some shocking material to the forefront. They raise concerns regarding both the performance of some automatic ‘lie detectors’ available on the market coupled with the readiness of various institutions to invest in these technologies without sufficient checks. Inner software mechanisms should be carefully analysed, particularly when potentially very serious consequences are tied to the application.

The research conducted here aims to assess a technological tool's potential for forensic applications. As well as uncovering the Y-ACCDIST system's capabilities, this thesis aims to discover its limitations.

Research Aims

Experiments and system developments in the present thesis differ from previous automatic accent recognition studies in two main ways. These are manifested in the two key research aims below:

- **Recognition of geographically-proximate accents**

  Previous studies developing automatic accent recognisers have used varieties dispersed across the breadth of Britain or indeed across the globe. Since it is unlikely that such recognition tasks would be of use in the forensic context, this thesis will address the recognition of accents distributed within a geographically-proximate area (namely the *Scottish-English Border*). By using a corpus comprising varieties which are geographically closer together, this would mimic the sort of task Stanley Ellis conducted in the case of the 'Yorkshire Ripper'. Although at the time Ellis' analysis did not necessarily aid the 'Yorkshire Ripper' investigation, it still demonstrates an instance of speaker profiling and a specific case where an accent recognition system might have been applied had the technology been available.

- **Accent Recognition using spontaneous (incomparable) data**

  Making further steps towards forensic applicability this thesis aims to investigate the Y-ACCDIST system's performance on spontaneous speech samples. Although a shared reading passage is mostly used, the system presented here has made adaptations to the model to be able to process spontaneous speech. In contrast, previous systems using a similar model to the Y-ACCDIST system have relied on highly comparable speech content (Hanani *et al*, 2011, 2013; Huckvale, 2004, 2007a). It is highly unlikely that speech samples brought to a forensic practitioner will be directly comparable in this way, so the following system makes context-independent segmental comparisons. Its performance on spontaneous speech (incomparable data) is presented.

It is of interest to improve the system's performance under the testing conditions brought about by the above research aims. Along with addressing these objectives, this thesis goes on to make further modifications to the system in attempt to improve on the results. In doing so, this thesis also addresses the need to discover the optimum phonemic units or combinations which yield the highest recognition rates. Past systems have either selected all segments containing vowels (excluding schwa) or taken a list of the most frequent segments. These approaches towards selection may have been the most appropriate for distinguishing between a large number of geographically disperse accents, but for the purposes of a more localised accent recognition task, a more considered or informed approach should be taken. Some speech segments are of value to accent classification while others may create 'noise' and distort results. The present thesis will address this issue and will discuss the possible routes to identify optimum system configurations.

Simultaneously, the results in the recognition tasks provide a basis for sociophonetic interpretation, implying categorical behaviour of the spoken varieties involved. This reveals an additional use of the Y-ACCDIST system: a tool for sociophonetic research.

Thesis Outline

To introduce the Y-ACCDIST system, chapter 1 focusses on accent recognition in a broader context by first looking at manual accent recognition and then moving on to past computational models. These computational models have largely been developed with the aim of supporting automatic speech recognition systems. Chapter 2 outlines the baseline architecture of the Y-ACCDIST system. Chapter 3 delivers results assessing the Y-ACCDIST system's recognition performance on geographically-proximate accents (addressing the first research aim). The first two sections of chapter 4 address the second research aim, applying the system to spontaneous speech data. Following sections test its performance under different configurations. The thesis then concludes with a discussion of the numerous avenues available to further develop the system.

Chapter 1

Accent Recognition

Accent recognition is a task which is more widespread than its role within the forensic domain. As discussed in the *Introduction*, accent is a major cause for concern and performance degradation in speech recognition technologies. Many accent recognition methods and approaches in the past therefore have automatic speech recognisers in mind and researchers are looking to link accent recognition models to their speech recognisers.

Elements of accent recognition also occur in sociophonetic studies where specific features of interest are highlighted as typical of a particular speech community. From a manual point of view, approaches adopted by sociophoneticians will be considered in this chapter. The issue of lay persons' perceptions of speakers' accents is also of interest to forensic speech science. This is with regards to the ability of a witness to give an accurate accent label to what they may have heard in relation to a crime.

This chapter looks at both manual performance and automatic models which have been applied to accent recognition for purposes which may or may not go beyond the forensic application. We will finally arrive at the ACCDIST (Accent Characterization by Comparison of Distances in the Intersegment Similarity Table) metric (Huckvale, 2004, 2007a), which has been employed for the system developed and presented in this thesis.

1.1 Manual Accent Recognition

Quantity of literature surrounding forensic speaker profiling in relation to literature on speaker comparison seems to be reflective of the proportion of casework each task type presents. Speaker comparison is by far the more common task in forensic speech science. Köster *et al* (2012: 56), however, report on a speaker profiling experiment. They find a 15.7% error rate for a group of 15 phonetically trained individuals who came to accent identity conclusions on 20 German speech samples. They were classifying samples into one of 14 accent classes. All of the analysts were native speakers of German. This result can be compared with some past results involving lay listeners discussed below. In general, it appears that phonetically trained individuals perform at a higher level than lay listeners.

Without a forensic focus, Vieru *et al* (2011) look into human accent identification. They assess French speakers' native language classification of speech samples. They had French speech samples of native speakers of American English, English English, German, Italian, Portuguese, Spanish and French. Overall, the native French listeners were correctly classifying the read speech samples 60% of the time (Vieru *et al*, 2011: 296).

Accent recognition results of native language varieties in Hanani *et al*

(2013) tell a very similar story. They use the *Accents of the British Isles* (ABI) corpus (D'Arcy *et al*, 2004) which contains speech samples of 14 varieties taken from locations across Britain (a more detailed description is given in section 1.2.4). By testing 24 lay listeners on the ABI corpus, they record a human accent recognition accuracy of 58.24% (Hanani *et al*, 2013: 70) in a forced choice experiment between 14 British accents. They also make a distinction within this result by comparing human accent recognition performance of 'familiar' and 'unfamiliar' accents. These categories were determined by marking accents belonging to regions in which they had lived as 'familiar' accents. Other accents were marked as 'unfamiliar'. By treating these familiarity categories separately, human listeners correctly identified 76.2% familiar accents, while only achieving 51.7% when asked to classify unfamiliar accents. These human accent performance experiments were conducted on the ABI corpus which Hanani *et al* (2011, 2013) also used to train and test their automatic accent recognition systems. Two of these systems employ the same model implemented in the Y-ACCDIST system presented here: the ACCDIST metric (Huckvale, 2004, 2007a). This allows for a direct comparison to be made with the automatic models which are discussed in further detail in the sections below.

First though, we move away from lay listener performance and towards traditional linguistic analysis of speech. Some discussion on the use of formants as analytical tools is given.

1.1.1 Formants

Ellis (1994) gives a detailed perceptual account of the phonetic features in the 'Yorkshire Ripper' hoax tape recording. A fine-grained phonetic analysis like this by an expert can pick out distinguishing accent features. For instance, the vowel in *strike* surfaced as a distinctive indicator. The vowel was realised more as a PRICE vowel, rather than a FACE vowel which would be typically expected in some parts of North East England (Ellis, 1994: 200). Such details can therefore contribute to specific accent conclusions. It is more common though to rely on formant frequency values taken from a spectrogram. Formants reveal the points at which there is increased energy within the vocal tract. Formants displayed in the spectrogram can therefore capture the shape of the vocal tract offering a way to measure and record vowel realisations. Most attention is paid to the first and second formants (F1 and F2). These frequency values produce a picture of vowel quality since F1 largely represents vowel height and F2 largely represents vowel backness (Ladefoged, 2003: 131).

Returning to the case described by Ash (1988), already discussed in the *Introduction*, she plots vowels in a 2-Dimensional vowel space (where F1 and F2 occupy the x and y axes). Plotting vowels in this way provided a visually accessible account of how different the vowel realisations were between the suspect's speech sample and the incriminating telephone calls. We can clearly observe the distance between the same vowel in the different speech samples. This formant-based approach resonates throughout forensic speech reports.

Another advantage behind formants is that they are primarily used throughout sociophonetic literature. If variety-specific literature is relevant to a specific case, the consistent discussion of formants could enable cohesive comparisons.

Although they are useful vowel descriptors, formants come with disadvantages requiring a lot of time, and therefore expense, from the phonetician. The demand for population data in forensic casework means that more formants than those in the centrally involved recordings are often needed. This, then, may lead to extensive formant value extraction. There are automatic formant value estimators in existence, but these do not necessarily offer the reliability required in forensic casework. Harrison (2004) compares the formant estimations for different software packages showing some great discrepancies between formant values. Such variation led Harrison to conclude that forensic phoneticians should be wary of the software they use and the settings in place due to inconsistencies. This calls for either the development of much more reliable automatic formant estimators or an alternative method of capturing vowel quality which is much more reliable. The system presented in this study opts for the latter and uses an alternative to formants: Mel Frequency Cepstral Coefficients (MFCCs) (explained below in section 1.2.1). Interestingly, previous studies (namely Huckvale, 2004, 2007a) have trialled the performance of formant values in a similar accent recognition model to the one employed here. This enables a comparison between these two methods of phonetic representation. These past results are discussed in further detail in the sections below.

1.2 Automatic Accent Recognition

In the past, automatic accent recognition systems have been developed for different languages. For example, Biadsy *et al* (2010) applies combinations and adaptations of the techniques described below to distinguish between four Arabic dialects. Similarly, in Koller *et al*, (2010), recognition systems were tested on different varieties of Portuguese. A large proportion of these systems involve decisions from the developer to dictate distinct groupings of accents. This assumes clear-cut divisions between them. Zheng *et al* (2005) acknowledge the reality of different degrees of accent and explore an 'accentedness' detection model with a view to adapting a speech recognition system based on the outcome.

Previous automatic accent recognition systems have ranged in the different algorithms used to maximise success rate. We can categorise approaches into two types: *text-independent* and *text-dependent* systems. Text-independent systems do not require transcriptions of the speech sample to conduct the analysis, whereas text-dependent ones do. In the context of developing automatic speech recognition systems, where an accent recogniser is run before the speech recogniser in the hope of improving speech recognition performance, text-independent methods are crucial. This is because a transcription is effectively the objective of these systems. Text-dependent accent recognition systems therefore have no place in general automatic speech recognition applications. DeMarco and Cox (2013) take text-independency a step further and specifically target accent classification methods with no phonetic labelling at all. In training their system, just one accent label is assigned to entire stretches of speech samples, motviated by the practicalities involved in acquiring phonetic labels for the majority of applications.

More recently, Najafian *et al* (2014a, 2014b) have been exploring the bridge between the results generated from an accent recogniser which then go on to make adaptations to an automatic speech recognition system. Here they are concerned with accent recognition models' overall impact on error rate of an automatic speech recognition system. Najafian *et al* (2014b) report a reduction of 44% in automatic speech recognition error rate when accent-specific systems are used. Najafian and Russell (2014) also highlight results regarding the quantity of speech data required for this adaptation. They compare the performance of speech recognition systems when they have been adapted by the results of accent recognition systems and individual speaker adaptation models. They find that accent-based model adaptation outperforms unsupervised speaker adaptation even when the speaker adaptation model uses five times the amount of adaptation data. This demonstrates how effective accent information can be in the broader area of automatic speech recognition.

Obviously, to be able to be practically applied to automatic speech recognition in as many applications as possible, these accent recognition models need to be text-independent. Conversely, in forensic speech science a text-dependent accent recogniser could have an informative role to play, particularly if it is a text-dependent system which achieves higher accent recognition rates.

Before this chapter embarks on the details of past systems, below is a description of Mel-Frequency Cepstral Coefficients (MFCCs) regularly used across speech technology and the feature vector used here. Following on are examples of past text-independent and text-dependent automatic accent recognition approaches. The chapter ends with details on the ACCDIST metric, a text-dependent method.

1.2.1 Mel Frequency Cepstral Coefficients (MFCCs)

As already mentioned, MFCCs are extensively used across speech technology. They can represent the speech signal ready for further processing for various applications. As we will also see, they can exist as an alternative to formant values for sociophonetic studies. This section goes into the details of how they are extracted.

The first stage of MFCC extraction is preemphasis. This reduces the effects of 'spectral tilt' by boosting the energy at higher frequencies which would otherwise be very low. MFCCs are taken at overlapping intervals throughout a speech signal. Short overlapping frames of usually 25ms of the speech signal are taken to generate a single MFCC vector. These frames are captured by a *Hamming* window, rather than a rectangular window. A Hamming window reduces the abrupt boundaries of each frame by peatering off gradually. Abrupt boundaries would ensue more signal discontinuities at the edges of the frame which would present problems for a Fourier analysis, the next step of feature extraction. To gather the spectral information of the speech signal, we apply a Discrete Fourier Transform (DFT) to each windowed selection. From the DFT of each window, we can extract the magnitude of each frequency component. Information at every frequency band in the signal, however, offers much more spectral detail than required

(Holmes and Holmes: 160). For each frame, therefore, a mel-spaced filterbank is applied. This is based on the mel scale which which is an approximation of the human auditory system; reflecting the fact that humans perceive more spectral information at lower frequencies. Consequently, in the case of mel-spaced filterbanks, more filters exist at the lower frequencies of the spectrum to extract more information lower down the spectrum. It is at these lower frequencies where, as humans, we can differentiate between speech sounds. The cepstrum is then derived by a Discrete Cosine Transform (DCT). The cepstrum separates signal properties determined by the source and those determined by the shape of the filter. Since it is the changes in the speaker's filter that lead to changes in the phones, values largely representative of only the filter contribute to the composition of the feature vector. For this reason it is common for the first 12 cepstral values to be incorporated into the MFCC vector. Values beyond the first 12 may contain useless information for speech recognition purposes (i.e, information about the speech source), producing noise in any forthcoming results.

It is possible to add dynamic information to the MFCCs. It may be useful to monitor 'change' between neighbouring frames of a speech signal since telling phonetic cues may be manifested in these changes. Diphthongization, for example may be represented more effectively if dynamic factors were integrated. Delta coefficients log the difference between one mel-frequency cepstral coefficient to its corresponding coefficient extracted from neighbouring frames. A delta coefficient can therefore be generated for each of the cepstral coefficients. Taking this further, double delta coefficents can be added which log change in a similar way between corresponding delta coefficients of neighbouring MFCC vectors.

1.2.2 Text-independent Accent Recognition

This section outlines some commonly used models applied to the accent recognition task which do not require a transcription of the speech sample as input (text-independent). Such methods have the potential of being applied to automatic speech recognition systems.

Gaussian Mixture Models

Gaussian Mixture Models (GMMs) are a widely used acoustic modelling technique. Within the area of speech science, GMMs have been used for Language Identification (e.g. Torres-Carrasquillo *et al* (2002), where they conclude that the performance of a GMM-based system is comparable with phonotactic approaches), speaker recognition (e.g. Reynolds *et al* (2000) as just one example of a huge pool of speaker recognition systems), and speaker sex classification (e.g. Zeng *et al*, 2006). The possibilities continue given the appropriate data.

Using extracted feature vectors from training data, GMMs can formulate a picture of what could be typically expected of a category of data. GMMs statistically model the parameters of the training data using a weighted mixture of multivariate Gaussian distributions. In the case of speech, where often MFCCs are used for modelling, corresponding coefficients are used to compute multivariate Gaussian distributions of the data, defined by a vector of means and a covariance matrix. When given unknown data observations, probabilities can be computed from the models to discover the most likely identity or class that unknown data belong to. GMMs can take on a wide range of configuration combinations and are often fused with a number of techniques to produce numerous variant systems. Hanani *et al* (2013) trial a range of variant GMM-based systems, including fusing the outputs of four different GMM variant systems to generate overall recognition decisions. This acoustic fused system generated the highest recognition rate out of the variant GMM systems with a recognition accuracy of 77.32%. This was conducted on the 14-way recognition task using fourteen British accents.

Phone Recognition followed by Language Modelling (PRLM)

Phone Recognition followed by Language Modelling (PRLM) is a method adopted from Language Identification (LID). As its name suggests, speech data is first passed through a phone recogniser and then the hypothesised sequence of phones is analysed to identify distinguishing phonotactic cues. Zissman (1996) employs a PRLM model to conduct an LID task, distinguishing between ten different languages. The model relies on the use of *n-grams*. These are sequential units consisting of n phones. In the training data of multiple languages, it is expected that the distributions of n-grams differ according to the particular language. These differences in distributions provide the basis for language recognition. Phone sequences of an unknown speech sample are hypothesised by the phone recogniser and the distribution of n-grams is computed. Likelihoods based on these n-gram distributions are calculated to arrive at a classification decision. In comparing different LID systems, Zissman finds that in a task involving ten languages, his variant PRLM system outperfoms a GMM system with an error rate of 21%, compared to the GMM error rate of 47% (Zissman, 1996: 39).

In the case of accent recognition, we can imagine accent properties such as rhoticity being detected in this way. For instance, the word *park* would be pronounced by a rhotic speaker of English with an /r/ segment included - /pɑrk/. A non-rhotic speaker, on the other hand, would pronounce it without /r/ - /pɑk/. These differences in sequential patternings among pronunciation systems should be reflected through the distributions of n-grams. An example of one of these PRLM accent recognition systems was developed in Lincoln *et al* (1998). In Lincoln *et al*'s system, distributions of diphones (sequential pairs of phones) throughout entire pronunciation dictionaries for different accents are logged. These act as reference databases of diphone distributions for unknown speech samples to be compared against. These comparisons lead to probabilities regarding class membership. Lincoln *et al* report results and link performance in relation to the number or spoken sentences per test speaker. In a task which recognises individuals as either a speaker of American English or British English, they report that no speakers are misclassified or unclassified when using at least 4 sentences from the test speakers. We return to the issue of data quantity in section 4.3.

1.2.3 Text-dependent Accent Recognition

This section describes past examples of text-dependent systems, where the transcription of test speech samples is known.

In contrast to the above unsupervised methods, Woehrling *et al* (2009) determine which linguistic features are accounted for in the classification of different French accents. They make a number of measurements of features which are known to be identifying factors between the varieties within their corpora. They take average vowel formant values for F1 and F2, voicing rate of consonants and segmental duration information to name a few. The automatic part comes in the form of the classification mechanism. Using this information they train two types of classifier: decision trees (using the Classification and Regression Tree (CART) algorithm) and Support Vector Machines (SVMs). SVMs have been employed in the Y-ACCDIST system and are explained later in section 2.1.2. Woehrling *et al* found that the SVM classifier outperformed the CART model with a highest classification rate of 85% distinguishing between five accents spoken in parts spanning over a wide geographical range: Northern France, Southern France, Alsace, Belgium and Switzerland. As well as being text-dependent, these classifiers have been modelled using vectors consisting only of information which the authors have specifically selected as being likely candidates for differentiation. Along with being laborious, it is possible that an automatic system would be more effective if it were able to identify the most telling factors which distinguish one accent from another. From a practicality point of view, this certainly would be preferable.

Moving towards phonotactic methods, Chen *et al* (2014) look at comparing segmental sequences of given speech samples. Their model could be viewed as a text-dependent alternative to the PRLM models introduced in the previous section. The segmental sequences were given by a reference pronunciation dictionary. Specifically in relation to this pronunciation dictionary, they look for insertions, deletions and substitutions of phones to formulate sets of accent-specific phonotactic rules. These sequences are discovered through modelling on adapted Hidden Markov Models (HMMs) - sequential statistical models. Their best variant system, when conducting a two-way recognition task on African American Vernacular English and non-African American Vernacular English, was performing with an Equal Error Rate of 9.97%. It is speculated here that for the accent recognition tasks conducted in this thesis, these sorts of sequential methods would not prove useful. Because geographically-proximate accents (which are hypothesised to be more similar than accents in past studies) are being tested, the degree of similarity is expected to be too great for these sequential methods to overcome. It is predicted that the distributions in sequential units would be too similar. Instead, we need to focus on differences in the realisations of individual phones. However, this specific hypothesis requires experimental support.

We can observe structural ideas of the ACCDIST metric (the model at the centre of the Y-ACCDIST system) in Barry *et al* (1989). Here, they consider inter-speaker variation and how it could be addressed. They form three-element vectors representing formant values of F1, F2 and F3. We can make a broad comparison of these with MFCC vectors when later considering the ACCDIST metric (section 1.2.4). It was these three-element formant vectors which represented stressed vowels of words in deliberately constructed sentences, encouraging the expression of accent differences. Using these vectors, Barry *et al* (1989) created an 'acoustic space' for each speaker while reducing the effects of 'long-term articulatory idiosyncrasies' (Barry *et al*, 1989: 356) (i.e, vocal tract length determined by biological characteristics such as gender). This was achieved by calculating the Euclidean distance between these formant-value vectors between pairs of vowel tokens. Despite differing vocal tract sizes - which therefore produce ranging formant values for the same phonetic segments for different speakers - calculating intra-speaker Euclidean distance in this way disregards these absolute spectral values and produces the relative distances between the different segments an individual utters. It is a way of normalising speakers' pronunciation systems. It is assumed that these relative distances are much more similar among same-accented individuals, consequently harbouring relatively similar 'acoustic spaces'. These acoustic spaces can therefore be compared among a population to speculate which individuals fall into the same accent class. Barry *et al* incorporated a basic scoring system, where Euclidean distances were used to judge the comparative degree of similarity between sets of vowels. A tally-like system was effectively put in place to log the number of characteristic vowel similarity hierarchies for each of the accents involved. ACCDIST-based studies have made progress on this kind of classification mechanism as will be revealed in this thesis.

The methodology adopted in Barry *et al* (1989) achieved promising results. The system processed the speech of 58 speakers, correctly classifying 43 into one of four accent categories (North American English, Scottish English, Northern English and Southern English). The methodology and results from Barry *et al* (1989) has provoked further interest and research producing and making use of accent spaces.

1.2.4 The ACCDIST Metric

The ACCDIST metric (Accent Characterization by Comparison of Distances in the Inter-segment Similarity Table) (Huckvale, 2004, 2007a) delivers a strategy for accent recognition which incorporates the relationships between individual speech segments. It combines techniques from speech technology with distance measures to arrive at a label for a speech sample attached to an originally unknown identity. ACCDIST provides a normalisation framework for direct comparison of individuals' pronunciation systems, disregarding additional complicating factors associated with varying voice quality, such as gender or age. By expelling such factors, ACCDIST becomes an appealing method for its purpose. The ACCDIST metric has been adopted by other researchers. Hanani *et al* (2011, 2013) trial a range of accent recognition approaches on the ABI corpus, reporting that an ACCCIST-based model yields the highest accent recognition rate (95.18%).

Past Studies Using ACCDIST

Most studies involving ACCDIST have used the same speech corpus: Accents of the British Isles (ABI) corpus (D'Arcy *et al*, 2004). The difference in accent corpora is a key focus of the present thesis, which is outlined in the first research objective in the *Introduction*. The current study looks at more geographically-proximate accents which are thought to harbour fewer distinctive accent features. The ABI corpus, on the other hand, contains varieties which are spoken in locations with greater distances in between. A description of the ABI corpus is given below to give context to past AC-

CDIST studies and for comparison with the experiments presented in this thesis.

Accents of the British Isles - ABI - Corpus

The ABI corpus comprises 14 different accents from across the breadth of the British Isles (D'Arcy *et al*, 2004: 116):

- Standard Southern English
- Midlands (Birmingham)
- Wales (Denbeigh)
- Scottish Highlands (Elgin)
- Republic of Ireland (Dublin)
- East Yorkshire (Hull)
- Lancashire (Burnley)
- Ulster (Belfast)
- North East England (Newcastle)
- Scotland (Glasgow)
- Inner London
- North West England (Liverpool)
- East Anglia (Lowestoft)

- West Country (Truro)

In each of the above locations, 10 males and 10 females were recorded. Each informant was required to have been born in the location and have lived there for the entirety of their lives. Each informant read a number of short prompts and longer pieces designed to elicit accent-specific features.

Huckvale's (2004, 2007a) ACCDIST metric could be viewed as a development of Barry *et al*'s three-formant vector approach, one key development being the use of MFCC vectors (described above in section 1.2.1). Using recordings of 274 speakers from the ABI corpus uttering 20 short sentences, Huckvale (2004, 2007a) takes the vowels from each speaker's sample and halves each vowel by time. The median MFCC vector from each half is taken. The resultant vectors from each half are concatenated to represent the vowel as a whole. Splitting and representing the vowel in this way makes sense to an accent recognition task. Taking measurements from two different temporal points of a vowel can capture dynamic movement within it which may prove significant when distinguishing between different varieties (i.e. capturing the difference between diphthongs and monophthongs). Once each vowel has been represented in this way, a distance table (an ACCDIST matrix) is formed by calculating the Euclidean distance between all possible vowel pairs. These intra-speaker vowel distances produce a normalised representation of the speaker's accent. Whether the speaker is male or female, for example, the distance between the TRAP and BATH vowels should be similar to others within the same accent class. Unknown speakers can therefore be classified on this basis. By employing a measure of similarity an unknown speaker's matrix can be compared against a reference set of speakers with known accent labels. Huckvale (2007a) uses correlation distance for this purpose. Chapter 2 gives a more detailed description of how an ACCDIST-based system works, obviously with reference to the specific mechanisms of the Y-ACCDIST system.

Like Barry *et al* (1989), Huckvale (2004) reports results where formant values have been used, incorporating them into intersegmental distance tables for comparison. He contrasts the use of these against the use of MFCCs for the same purpose and in the same model. The median first four formant frequencies were taken for each half of each vowel phone. By training and testing on the ABI corpus, a best score of 72.6% was generated. This greatly contrasts with the results gathered by using MFCCs as the representative vector for each phone, where a best score of 87.2% is reported (Huckvale, 2004: 31). One methodological difference which is likely to have a negative impact on the formant frequency vector method is formant value estimation. It is widely agreed that formant trackers on various spectrographic software are not completely reliable. This issue has already been touched upon in section 1.1.1.

Ferragne and Pellegrino (2007, 2010) also compare strategies for accent classification using the ABI corpus. Using the ACCDIST metric, they categorise speakers using the /hVd/ wordlist from ABI: a set of comparable and realistic word contexts for most of the vowel phonemes in English. This specific wordlist provides a common consonantal enviroment in which each of the traditional vowel phonemes can be expressed. These are frequently used to express the differences in accents. The vowel phonemes are famously outlined in Wells' lexical sets (Wells, 1982). Ferragne and Pellegrino (2007, 2010) demonstrate the ACCDIST metric's potential in sociophonetic research. They contrast its performance and practical advantages with that of more traditional formant analyses. It is shown that as well as being able to automatically extract MFCC vectors more reliably than formants, this system can display expected vowel space relationships between accents in line with sociolinguistic literature - such relationships are also uncovered in chapter 3 of this thesis. The individual vowel representations from each of the /hVd/ contexts are placed into an ACCDIST matrix for each speaker, and the matrices are compared using Pearson product-moment correlation. They generated a similar recognition rate to Huckvale (2004) of 89.7%. Ferragne and Pellegrino (2007: 251) suggest that by generating very similar recognition scores, it is possible that the ABI corpus has been exhausted for this particular accent recognition method.

Hanani *et al* (2011, 2013) compare the performance of a number of accent recognition systems. Included are two ACCDIST-based systems, GMM systems and human accent recognition performance. It was found that both the ACCDIST-based systems outperformed other system designs with recognition rates of 95.18% and 93.17%, while human performance was recorded at 58.24%. We saw above in section 1.2.2 that the highest performing GMM-based system achieved 77.32%. The ACCDIST-based systems were developed using triphones as the segments acting as pinpoints for distance calculations. This differs from Huckvale's (2004, 2007a) systems where he treated each vowel in the speech samples as unique vowels if they appeared in unique word contexts. Triphones, on the other hand, are sequential units of three phones which phonotactically exist within a language. They are often used in automatic speech recognition systems. (The difference in segmental units is returned to and tested later in section 4.1 of this thesis.) The resultant ACCDIST matrices were of a substantial size (105 x 105). They used the 105 most frequent triphones in the speech dataset, generating the mean MFCC vector across multiple tokens of the same triphone.

Hanani *et al's* highest performing accent recognition system was the ACCDIST-SVM system (ACCDIST coupled with a Support Vector Machine classifier). Since it was found to be a powerful system, the Y-ACCDIST system is based on this model combination. The specifics of the Y-ACCDIST system are outlined in chapter 2.

Chapter 2

The Y-ACCDIST Accent Recognition System

The Y-ACCDIST system's development can be explained in two phases: *accent modelling* and *accent classification*. The accent modelling phase describes how an individual's speech sample is taken and represented by a table (matrix) of distinctive values. Most importantly, how this table represents a speaker's accent is explained. These matrices are generated for all speaker samples involved for both the known reference speakers and the unknown speakers which are later classified. The classification phase describes the process of taking these representative speaker matrices and how they can be fed into a *Support Vector Machine*, ready for an unknown speaker to be compared against the reference population and to be assigned an accent label. Information about some previous classification methods is also given.

2.1 System Development

The baseline Y-ACCDIST system in this study was developed as follows. The numbered steps below are accompanied by figure 2.1 for the explanation.

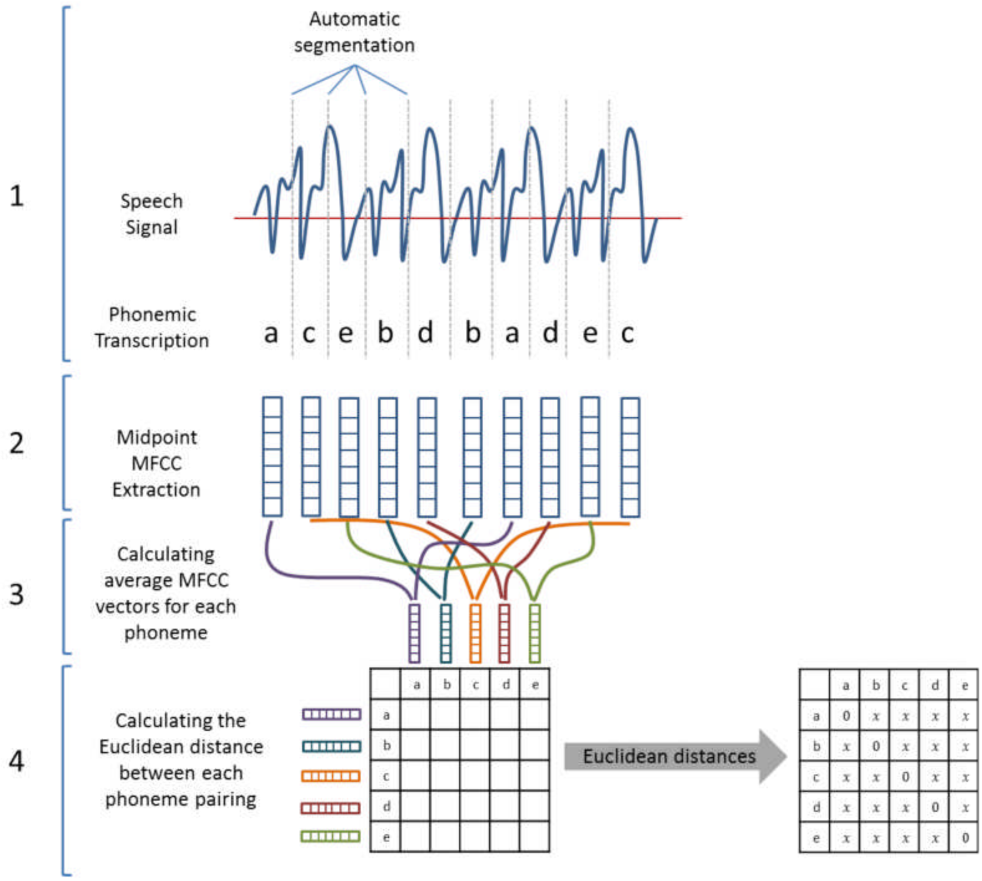
2.1.1 Phase 1: Accent Modelling

1. A forced aligner was built (using the Hidden Markov Model Toolkit (HTK) version 3.4 (Young *et al*, 2009)) to automatically segment each speaker's transcribed speech sample.

2. 12-element midpoint MFCCs are extracted for each phone segment.

3. For each vowel phoneme type, the corresponding phone MFCCs are pooled together to produce an average vector. This results in a mean vector to represent each vowel phoneme.

4. Each segment in the vowel inventory forms the template of a matrix (as shown in step 4 of figure 2.1 below). The Euclidean distance is then calculated between the averaged vectors for each combination of vowel pairs. To standardise the Euclidean distances, z-scores are calculated to generate relative distances for the specific speaker.



	a	b	c	d	e
a	0	x	x	x	x
b	x	0	x	x	x
c	x	x	0	x	x
d	x	x	x	0	x
e	x	x	x	x	0

Figure 2.1: The production of an ACCDIST matrix

Distance values can represent the degree of similarity between two vectors, showing how ‘close’ they are to each other. This collection of values represents an individual’s vowel pronunciation system. This can be illustrated by taking the vowels in FOOT and STRUT, as a classic example. Having heard a speaker of Northern English English utter these items as well as a Southern English English speaker, a listener is likely to agree that these two vowels sound more similar for the Nothern speaker than they do for the Southern speaker. In theory, distance values should reflect these differences. For the Northern speaker, the distance value between these two vowel phonemes is expected to be smaller than it would be for a Southern speaker. By extending this theory across a vowel inventory in the matrix shown above, the similarity relationships between speech segments is quantified for individual speakers, ready for comparison.

The above sequence is applied to speech samples from a number of speakers until we have a number of matrices to represent a variety of accents. This first phase has described how individuals' accents are represented. The next phase describes how these models are implemented in classification.

2.1.2 Phase 2: Accent Classification

Past ACCDIST-based systems have incorporated different classification mechanisms. Initially in Huckvale (2004, 2007a), correlation distance was used. Another correlation measure (Pearson product-moment correlation) was used by Ferragne and Pellegrino (2007, 2010). Similar results were achieved among these systems. Using a correlation measure expresses degree of similarity between an unknown speaker's matrix and a reference matrix. A high correlation between two matrices would therefore suggest that they belong to the same accent class. Hanani *et al* (2013) look at two alternative classifiers to follow ACCDIST modelling: correlation distance and Support Vector Machines (SVMs). They found SVMs to outperform the correlation method, where the SVM system achieved a recognition rate of 95.18% and the correlation system achieved 93.17%. Similar findings have been found during the Y-ACCDIST system's development. An earlier version of the Y-ACCDIST system was presented in Brown and Watt (2014) and Brown and Wormald (2014). The only difference between the earlier system and the system presented here is the classification mechanism. The earlier system implemented Pearson product-moment correlation (as per Ferragne and Pellegrino (2007, 2010)). Using the earlier version of the Y-ACCDIST system, Brown and Watt (2014), reported a recognition rate of 79.2% on the same four-way accent recognition task conducted later in this thesis. However, when a SVM replaced correlation, it was found that it served as a stronger classifier as results in chapter 3 clearly show.

Support Vector Machines

The Support Vector Machine is a widely used machine learning algorithm across numerous disciplines. It bases its classification conclusions on the plotting of classes of training data in high-dimensional space followed by the plotting of unknown data. Classification is governed by where unknown samples fall in relation to decision boundaries determined by the training data. The SVM used in this study to classify speakers into accent categories works as follows:

1. Each speaker's speech sample in the training database is converted into an ACCDIST matrix in the way shown in phase 1 (section 2.1.1).

2. For each of the different elements within a single ACCDIST matrix, a high-dimensional space is constructed for an SVM where the number of dimensions corresponds with the number of different matrix elements.

3. One SVM is constructed for each of the four classes involved in this

recognition task and a 'one-against-the-rest' structure is employed. A decision boundary is formulated in the form of a 'hyperplane'. Since an SVM is an optimisation algorithm, the *optimal hyperplane* is calculated. Obviously, in such classification problems, there are numerous possible hyperplanes. The optimal hyperplane is the one which sits with the largest possible margin between the two classes of data. Using a simplified 2-Dimensional space, the diagram below illustrates just one SVM modelling one accent category:

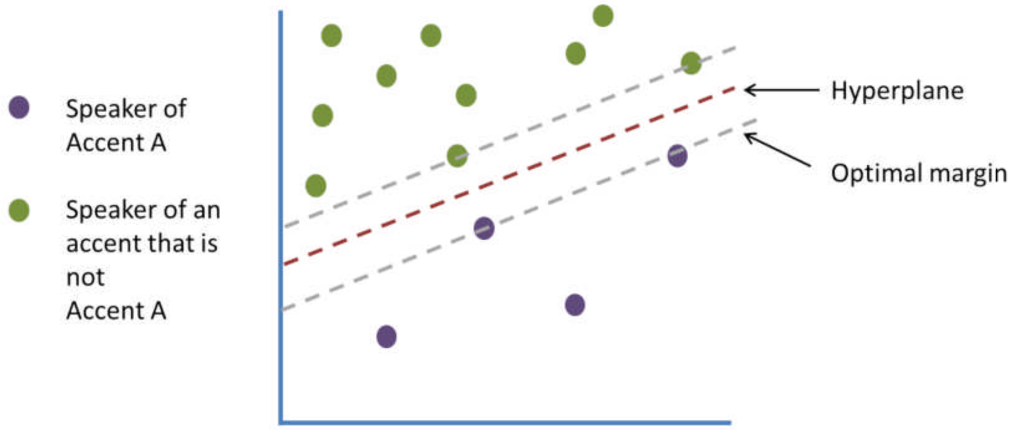


Figure 2.2: Simplified illustration of a Support Vector Machine

Figure 2.2 above shows a support vector machine where training speakers of one accent class (Accent A) are modelled against all other speakers of the remaining accent classes in the reference set which take on a collective category (hence 'one-against-the-rest').

4. An SVM is constructed for each accent where each takes on the role of Accent A in the figure above. Each accent is also integrated into the

collective ‘the-rest’ category for each of the other three models.

5. To classify an unknown speaker, the speech sample is processed to form a representative ACCDIST matrix. It can therefore be compared and classified within each of the four SVMs by plotting the speaker’s matrix in these high-dimensional spaces. Classification is derived from the model in which the unknown speaker achieves the clearest margin in, indicating accent class membership.

Figure 2.3 below illustrates an overview of the whole process explained by the two phases above.

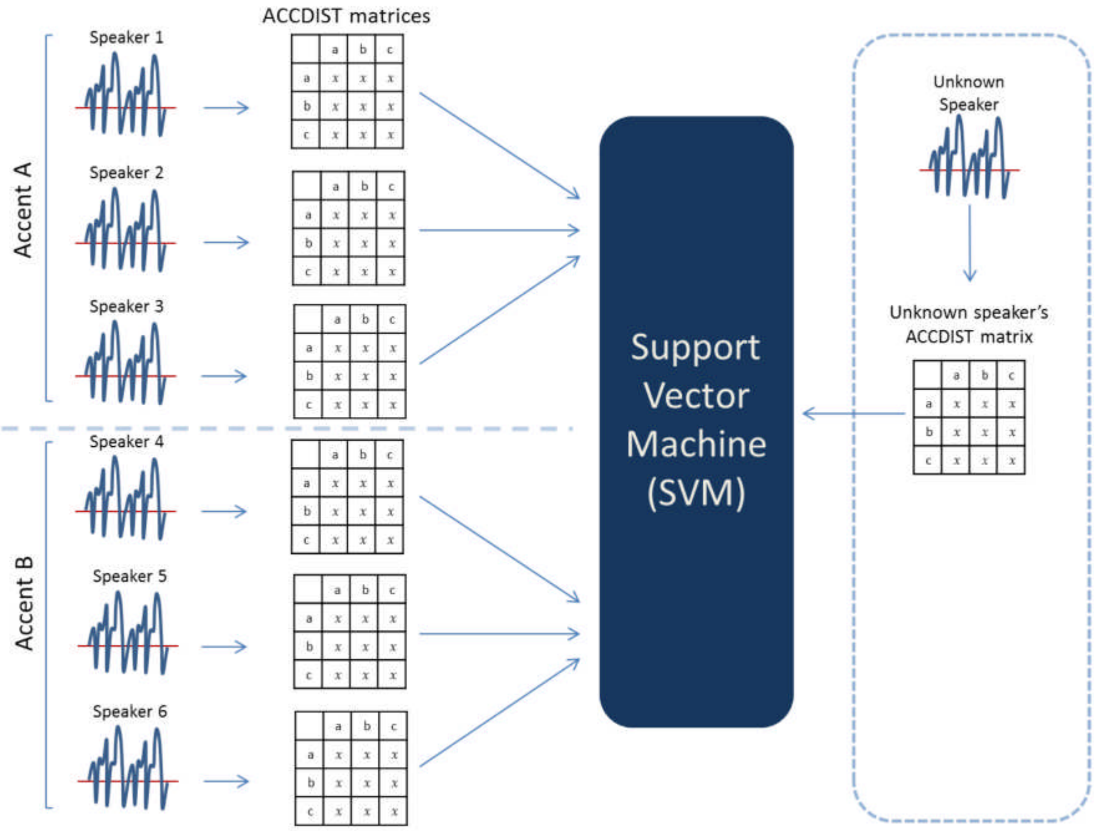


Figure 2.3: Overview of the Y-ACCDIST accent recognition process

2.2 Testing Protocol

Outlined below are the experimental details which recur throughout the experiments in this thesis. Of course, the system is tested under different configurations. The corresponding methodological manipulations required to test the system are described in their resepective sections.

2.2.1 Data Processing

All experiments within this thesis have been conducted using data from the AISEB (Accent and Identity on the Scottish/English Border) corpus (Watt *et al*, 2014). A more detailed account of the AISEB corpus is given below in section 3.1. For the purposes of describing the testing process, here is a short summary of the data used throughout.

Most experiments are based on recordings of 120 informants reading a passage, *Fern's Star Turn*. These 120 speakers can be divided into four accent groups based on their geographical origin: Berwick-upon-Tweed, Carlisle, Eyemouth and Gretna. Within each of these geographical groups, two further subcategories of 15 informants aligning with speaker age exist: older and younger speakers. The younger category contains speakers ranging from ages 14 to 27. The older category ranges from 54 to 93. Recordings took place in each of the four locations, in varying environments (often in individuals' homes). Although the recordings are of a high quality (with a sampling rate of 44.1 kHz), there is some variability between speakers' recordings.

Each speaker's reading passage recording was divided into indvidual sound files for each sentence (although some sentences were judged as being significantly longer than others, so were cut into two clauses, assisting the forthcoming forced alignment process). Some sound files were rejected from the study due to one of a number of reasons: the presence of multiple reading disfluencies (e.g, false starts, filled pauses, etc.), background noise or disruption, coughing, laughing or another emotion affecting the reading quality. If one of these factors occurred throughout the recording or on multiple occasions, the informant was discarded for this particular study. This is largely due to the effects impacting on the forced alignment process. The AISEB corpus is large enough for discarding speakers in this way as it holds a total of 160 speakers. Consequently, for this study, a total of 120 speakers could be processed while keeping each speaker origin category and speaker age subcategory equal across the board. In the worst cases of the remaining 120 speakers, where only specific sentences were discarded, all informants still maintain at least 35 of the total 39 partitions of the reading passage. Although a selective mindset was at play here, some sentences containing such errors are included in the present study's dataset. If only one disfluency or cough, for example, occurred within a sound file, then it has retained a place in the dataset. The reading passage provided approximately 3-4 minutes of speech per speaker. This is a quantity which may be appropriate for comparison with forensic casework. In practice, speech samples of between 90 seconds and five minutes would a reasonable target length when conducting real-life analysis (Foulkes and French, 2012: 563).

The forced aligner used in this study was built using the Hidden Markov Model Toolkit (HTK) version 3.4 (Young *et al*, 2009). The HTK toolkit was also used to extract the midpoint MFCCs described in stage 2 of the *Accent Modelling* process (see section 2.1.1 above). These contained 12 cepstral coefficients, which were extracted every 10ms with a window length of 25ms.

2.2.2 Testing

For testing, the *leave-one-out* cross-validation approach was employed (as per Huckvale (2004, 2007a)) to maximise training data. In turn, each speaker in the corpus becomes the ‘test’/‘unknown’ speaker and is removed from the reference set. The test speaker is then treated in the way described and shown above to contribute to an overall result. This is repeated on rotation for each speaker within the corpus.

Throughout this thesis numerous modifications will be made to the overall process, be it changes to the data or to the system composition itself. The percentage of correct classifications will be logged as a measure of the Y-ACCDIST system’s performance under various settings.

Chapter 3

Automatic Recognition of Geographically-Proximate Accents

The previous studies involving the ACCDIST metric (discussed in chapter 1) all make use of the Accents of the British Isles (ABI) corpus (D'Arcy *et al*, 2004). As already stated, this corpus is composed of varieties taken from a wide dispersion of locations across Britain. Through cluster analyses, Huckvale (2007b) shows that accent varieties are more distinct when they are spoken at greater geographical distances apart. There appears to be a relationship between relative physical distance and degree of accent similarity among the ABI varieties. This chapter focusses on the ACCDIST metric's sensitivity in this respect by testing the Y-ACCDIST system's performance on a corpus of varieties spoken within shorter geographical distances. This falls in line with tasks which may be of relevance to the forensic field (as we saw in the ‘Yorkshire Ripper’ case). The current study challenges the system by observing its performance on more similar varieties. The AISEB corpus (introduced fully in section 3.1) has been chosen to serve this purpose.

As well as exploring ACCDIST’s capabilities on a more localised scale, this study also explores its potential in speaker age estimation tasks. This aims to explore the model’s sensitivity further. Investigating the potential for speakers to be classified by age using an ACCDIST model may also prove advantageous for forensic applications. Accents change through time and so it is reasonable to expect accent differences between generations.

From an acoustic perspective, Schötz and Müller (2007) assessed a range of features manifested in the speech signal as hypothesised indicators of age. Out of the selection, they identify speech rate and sound pressure level as the most accurate age estimators. The factors tested were largely associated with the physiological traits of aging. Another study was conducted by Metze *et al* (2007) where they compared different computational models to classify speakers according to age and gender group. They found that the system which employed a phoneme-based recogniser (i.e, taking phonemic information into account) combined with Hidden Markov Models (HMMs) outperformed other systems designed for the same purpose. The other systems were modelled on other feature types such as prosodic indicators (jitter and shimmer) and fundamental frequency. These systems did not take phonemic information into account and used a much more general modelling procedure using GMMs. Such findings suggest that more phonemically dependent approaches to age classification should be considered in system development. This study, combined with the outcomes of sociophonetic research, calls for the Y-ACCDIST system, a phonemic model, to be trialled in the context of age recognition. Results for this kind of accent recognition task, where age is incorporated into accent groupings, are presented in this chapter.

Metze *et al* also show, however, their phonemic-based model's drop in performance when the quantity of speech data is reduced. By contrast, the system which uses prosodic indicators to estimate speaker age and gender appeared to be relatively robust to a shorter duration of speech data. The topic of speech sample duration is obviously a crucial talking point when developing a compuational tool in the context of forensic speech science. This issue is revisited later in this thesis in section 4.3.

The current chapter presents the Y-ACCDIST system's classification performance on geographically-proximate location categories and further subcategories determined by age group. To do this, the sytem will be trained and tested on the AISEB corpus which is described in detail in the section below.

3.1 The Accent and Identity on the Scottish/English Border (AISEB) Corpus

Interest in the effects of a political border dividing two countries led to the *Accent and Identity on the Scottish/English Border* (AISEB) project (Watt *et al*, 2014). The AISEB corpus consists of speech production data, as well as perception data collected from informants living in one of four locations closely situated either side of the Scottish-English border: Berwick-upon-Tweed, Eyemouth, Carlisle and Gretna. Recorded readings and interviews were involved in data collection. The AISEB locations are 'paired' (Watt *et* *al*, 2014: 80) in terms of their geographical proximity and differing national affiliation. Berwick-upon-Tweed, an English town, is only approximately ten miles away from Eyemouth sitting on the Scottish side of the border. Likewise, Carlisle is approximately ten miles away from Gretna, but Carlisle is south of the border while Gretna is on the Scottish side. These pairings can be seen more clearly in figure 3.1 below:

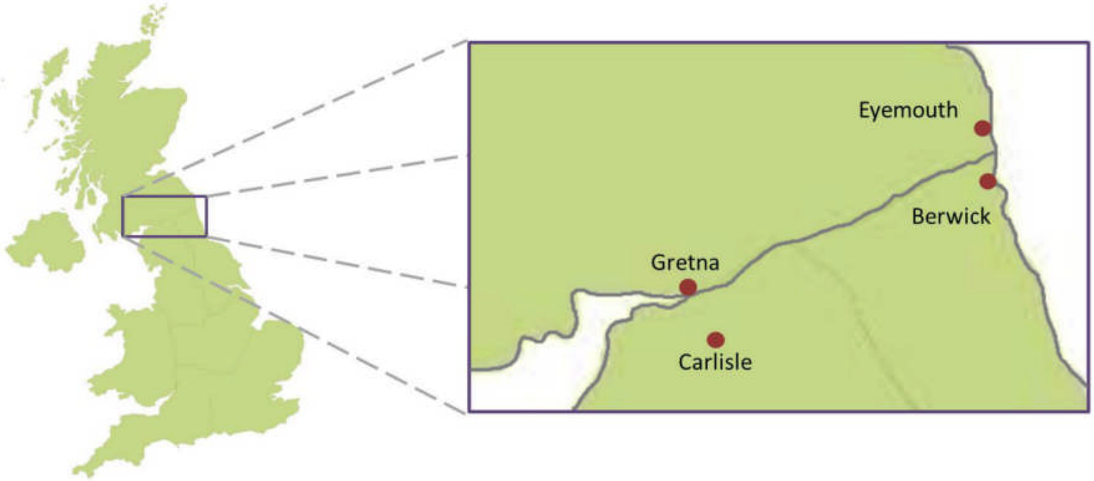


Figure 3.1: The four locations of the AISEB project

It is largely the reading passage which is used throughout this thesis. The reading passage *Fern's Star Turn* is a story specifically written for the AISEB project. It contains all of the words in Wells' lexical sets (Wells, 1982), eliciting accent differences through vowel realisations.

Some sociolinguistic analysis has already been conducted on the corpus. As a focussed study, Watt *et al* (2014) look at rhoticity and the /r/ variants used in the speech among the four locations. Rhoticity is a key distinguisher between the broad generalised varieties spoken in Scotland and England, with Scottish varieties associated with the presence of /r/ in a syllable's coda (Hughes *et al*, 2005: 45). With such a typical feature separating the two countries' spoken varieties, it is of great interest to explore its place in speech communities lying so close to the national border. Despite the very short distance between the Scottish towns and their English counterparts, Watt *et al* found that rhoticity was almost completely absent in the English speech communities, whereas it is a retained feature in the Scottish communities. Speakers of Eyemouth, in particular, show a much higher proportion of rhotic forms (roughly 80% rhotic), while Gretna, the other Scottish town in the corpus, collectively produces a lower proportion of just under 50% (Watt *et al*, 2014: 88). Findings such as this suggest a sort of continuum with Gretna sitting midway between Eyemouth and the English towns regarding rhoticity. Given the towns' respective histories, we can expect realisational continuums to be in place and that is in fact what is shown in results below. As suggested by Watt *et al*'s rhoticity results, it can be hypothesised that Gretna can fall as a 'hybrid' between the Eyemouth accent and the English varieties because of its relative youth as a place. Gretna formed in the First World War due to housing demand from new employees coming to the area to work in the world's largest munitions factory (Watt *et al*, 2014: 82). This involved people coming together from a range of areas surrounding Gretna and so linguistic contact would have taken its toll. It may well be due to Gretna's history and relatively young age that its spoken variety behaves as a hybrid in relation to other spoken varieties within the corpus. Such relationships between the spoken varieties in this corpus are reflected in the automatic accent recognition results shown in this chapter.

3.2 Methodology

The first classification task aimed to recognise speakers by their geographical origin. This was achieved by following the testing protocol outlined in 2.2.2. The AISEB corpus however, as explained above, offers an additional classification dimension as two different age categories exist among speakers. Following recognition tasks therefore consider age groups. System perfomance has been assessed by training and classifying data by geographical origin from just one age category at a time, as well as adding age category as an additional group variable to create eight speaker groups.

3.3 Results

Presented here is the classification performance of the system under different arrangements and groupings of the data. The results produced by these different data configurations are shown in the table below.

Classification Task	**% Correct**
4-way classification according to geographical origin, all speakers, N=120	86.7
4-way classification according to geographical origin, younger speakers, N=60	83.3
4-way classification according to geographical origin, older speakers, N=60	83.3
8-way classifcation according to geographical origin and age category, N=120	69.2

Table 3.1: Recognition of accents under different data groupings

The first recognition task tested all speech samples, collapsing age groups, classifying speakers according to location alone. As a four-way recognition task, a chance-level recognition rate of 25% would be expected, but here a result of 86.7% exceeds chance expectations. Although the recognition rates are not as high as previous studies which employ ACCDIST-based models, the difference in geographical proximity of these spoken varieties is expected to be to blamed for this reduction. Originally, it was predicted that by restricting the reference system's age categories (i.e, narrowing the variational focus by conducting recognition tasks within the respective age group) would ouput higher recognition rates. The results in table, however, show reasonably consistent recognition rates, irrespective of separating age categories or combining them. We can observe a slightly higher recognition rate in the first recognition task, but this could be down to twice the number of speakers used overall. This leads to a more substantial quantity of training data. All three sets of recognition tasks here demonstrate rates of well above chance level.

Confusion matrices of each of the recognition tasks are found in the tables below. They offer some insight into the sociolinguistics of these accents and these are discussed.

loc.	**Ber**	**Car**	**Eye**	**Gre**
Ber	**24**	1	1	4
Car	1	**28**	0	1
Eye	1	0	**29**	0
Gre	1	4	2	**23**

Table 3.2: Confusion matrix of four-way recognition task between locations

loc.	**Ber**	**Car**	**Eye**	**Gre**
Ber	**9**	2	1	3
Car	0	**14**	0	1
Eye	0	0	**15**	0
Gre	2	0	1	**12**

Table 3.3: Confusion matrix of younger speakers' four-way recognition task between locations

loc.	**Ber**	**Car**	**Eye**	**Gre**
Ber	**13**	0	1	1
Car	0	**14**	0	1
Eye	2	0	**12**	1
Gre	1	3	0	**11**

Table 3.4: Confusion matrix of older speakers' four-way recognition task between locations

loc.	**BY**	**BO**	**CY**	**CO**	**EY**	**EO**	**GY**	**GO**
BY	**4**	4	2	0	2	0	3	0
BO	5	**8**	0	0	0	1	0	1
CY	1	0	**11**	1	0	0	2	0
CO	0	0	1	**12**	0	0	0	2
EY	1	0	0	0	**14**	0	0	0
EO	0	2	0	0	0	**12**	0	1
GY	1	0	1	0	1	0	**11**	1
GO	0	0	1	0	1	0	2	**11**

Table 3.5: Eight-way recognition task between combined location and age categories - BY=Younger Berwick speakers, BO=Older Berwick speakers, CY=Younger Carlisle speakers, CO=Older Carlisle speakers, EY=Younger Eyemouth speakers, EO=Older Eyemouth speakers, GY=Younger Gretna speakers, GO=Older Gretna speakers

3.4 Discussion

The results presented here fall in line with sociophonetic expectations. Generally speaking, the confusion matrices (in particular, see table 3.2) show a greater proportion of Eyemouth speakers being correctly classified. Speakers from Carlisle tend to follow Eyemouth in this respect. Berwick and Gretna speakers on the other hand are correctly classified on fewer occasions. This pattern is largely echoed throughout the confusion matrices (apart from table 3.4 showing results for only older speakers). This consistent patterning is put down to the histories of each location. As already stated, Gretna was established later than Carlisle and Eyemouth, and so it is expected that spoken varieties are less distinct in these locations. In the way that Gretna formed in the First World War, the influences of the different spoken varieties coming together seem to be having effects on the probability of a Gretna speaker being correctly classified. This aligns with what has already been suggested in section 3.1 from Watt *et al*'s (2014) findings: Gretna's 'hybrid' nature as an accent variety. Berwick has historically swapped between national identities a number of times which may have impacted on the number of confusions the system makes with these speakers.

With the relative geographical distance in mind, no confusions occur between Carlisle and Eyemouth speakers at all. Sociolinguistically, this is to be expected between the two more established locations which exist on either side of the border and at opposite ends of its length. As a result, we would expect fewer accent characteristics to overlap between these varieties which holds true in the above tables. A distinctive classification ranking between different accents was also found among the ABI varieties. DeMarco and Cox (2012: 3) find in their experiments that speakers from the Scottish Highlands hold the highest classification rate as a group with 95.24%, while speakers from Cornwall are collectively recognised 22.22% of the time. These substantial differences suggest that accent recognition is also influenced by the varieties themselves.

The recognition rate generated by the eight-way system, where age factors into the grouping as well, suggests that there are marked accent differences between age groups. We see a correct classification rate of 69.2%, which is well above the expected chance classification rate of 12.5% attached to eight-way recognition tasks. This result suggests a reasonably high degree of accent distinctiveness between age groups within a single location. The worst performing speaker group, in this respect, seems to be Berwick speakers, showing only four correct classifications of younger speakers and eight correctly classified older speakers. From the confusion matrix, shown as table 3.5, we can see a number of confusions occurring between age categories within speaker location. For example, five older Berwick speakers which were incorrectly classified were instead assigned the younger Berwick speaker accent category. This is not surprising as we would expect the most accent similarity to exist between age group varieties within the same location. Another interesting point regarding these age group confusions is that all incorrect classifications which occur between locations remain within the same age category. For example, taking the worst performing speaker group, younger Berwick speakers, most incorrect classifications are with older Berwick speakers. However, all other incorrect classifications of these speakers occur in only the younger categories of the other locations. This is a pattern which resonates throughout all speaker categories in the confusion matrix. The consistency of this patterning is suggestive of the Y-ACCDIST system's potential as a sociophonetic tool, as well as a forensic tool. The confusion matrices express patternings which largely fit with the varieties' sociolinguistic profiles and is therefore showing a novel and objective method for analysing accent corpora.

For forensic purposes, it may be of interest to simulate a situation where a test speaker is compared against an old corpus and the effects this may have. A shortage in population data may tempt the use of old corpora, or the closest that can be found, but here we find a large decrease in recognition rate. French and Harrison (2006: 248) rightly point out that the 'shelf life' of information on regional variation is very short since 'accents are in a constant state of flux'. To demonstrate, all 60 speakers from the older category were used to form the reference matrices of each of the four geographical locations. All 60 younger speakers were used as test speakers and recognition rate is based on whether they were correctly classified according to geographical location. This experimental configuration could be seen to mimic a scenario where a recently produced speech sample is compared against a database collected some time ago. The result is compared directly with that collected when all the younger speakers were tested only on reference matrices produced from only other younger speakers (the recognition task shown in table 3.3).

Age category of reference data	**% Correct - Younger speakers tested**
Younger speakers	83.3
Older speakers	65.0

Table 3.6: Comparison of results with mismatched age groups

These recognition results clearly echo the short 'shelf life' of informative accent material and the importance of collecting 'fresh' speech data for analysis. Even though the data still needs to be collected, one major advantage of using compuational tools is the speed at which they can process the data. Computational tools can help to keep up with the inevitable changes embraced by language.

Agglomerative Hierarchical Clustering

A cluster analysis of individual speakers' ACCDIST matrices supports the conclusions drawn from the confusion matrices above. In the development of the Y-ACCDIST system, it is effectively trained on accent groups chosen by its developer. In experiments throughout this thesis, reference ACCDIST matrices are grouped based on speaker properties (i.e, geographical origin and some experimentation with age group). A speaker's accent, however, can be influenced by a number of factors. Shown here is a hierarchical cluster analysis of every individual's ACCDIST matrix and how the speakers naturally group without the developer's decisions and assumptions dictating a specific direction. Huckvale (2007b) shows the 'complete' linkage method to be a satisfactory clustering method using ACCDIST and is subsequently the one employed here.

Agglomerative hierarchical clustering works in a ‘bottom-up’ fashion, meaning that starting at the lowest level of clusters (beginning with each of the individual speakers), it then clusters in stages based on highest degree of similarity between pairs. Clusters are formed and similarity measurement is continued between larger and larger clusters. It is the relative positioning of speakers along the x-axis which is meaningful here. By assumption, the further away two speakers are along the x-axis, the lower the correlation between the speakers’ ACCDIST matrices, indicating less similarity between the speakers. The dendrogram below displays the cluster analysis of all 120 speakers. True speaker properties regarding geographical origin and age are revealed along the x-axis.

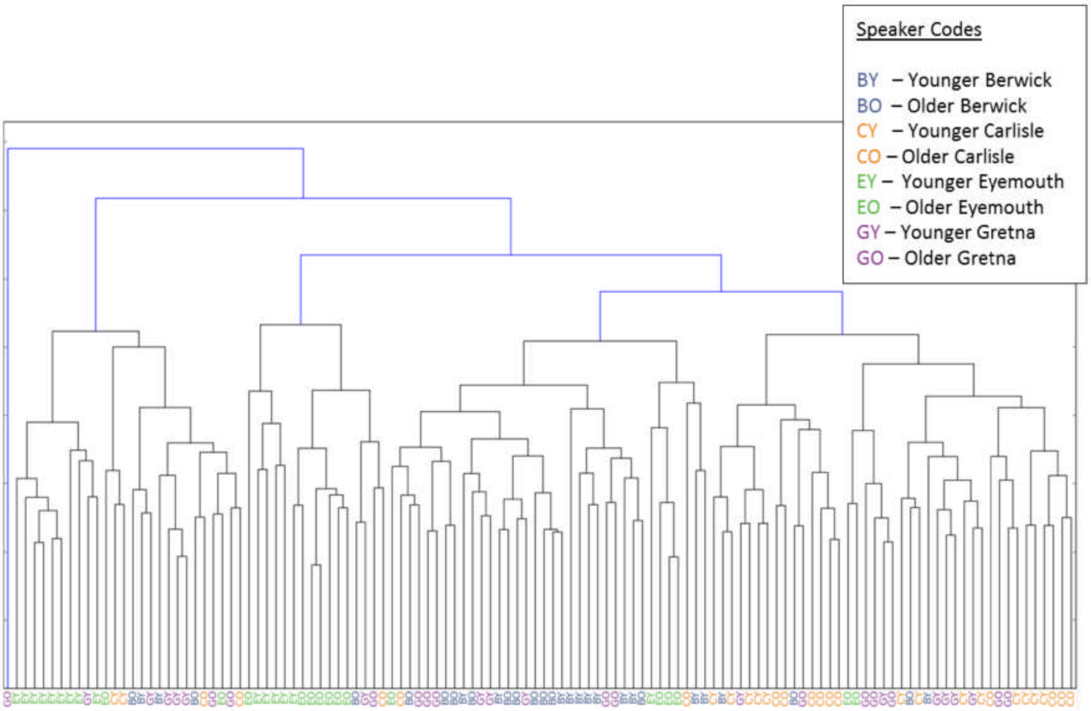


Figure 3.2: Agglomerative hierarchical cluster analysis of 120 AISEB speakers

Obviously at present, due to the number of speakers, the outcome is difficult to observe. This section takes a closer look using close-up windows. The cluster analysis in figure 3.2 above shows a similar overall picture to the one predicted by sociolinguistic speculation over these AISEB accent varieties. It reveals a similar story manifested in the confusion matrices already seen above. Earlier in this chapter we can recall an expected 'continuum' of these varieties, where it could be expected that Gretna speakers and Berwick speakers generally fall between those in Eyemouth and Carlisle. We also saw no confusions existing between Eyemouth and Carlisle, indicating that there is limited overlap between the spoken varieties. This possibly suggests that these sit on opposing ends of a continuum. The dendrogram here supports this expectation. Figure 3.3 shows four clusters, arranged into three (two smaller ones to the left are put into one division). These divisions of the clusters best highlight the significant arrangement and ordering of speaker identities along the x-axis. Figure 3.4 overleaf presents close-up windows of each of these divisions for a clearer viewing of the true speaker properties.

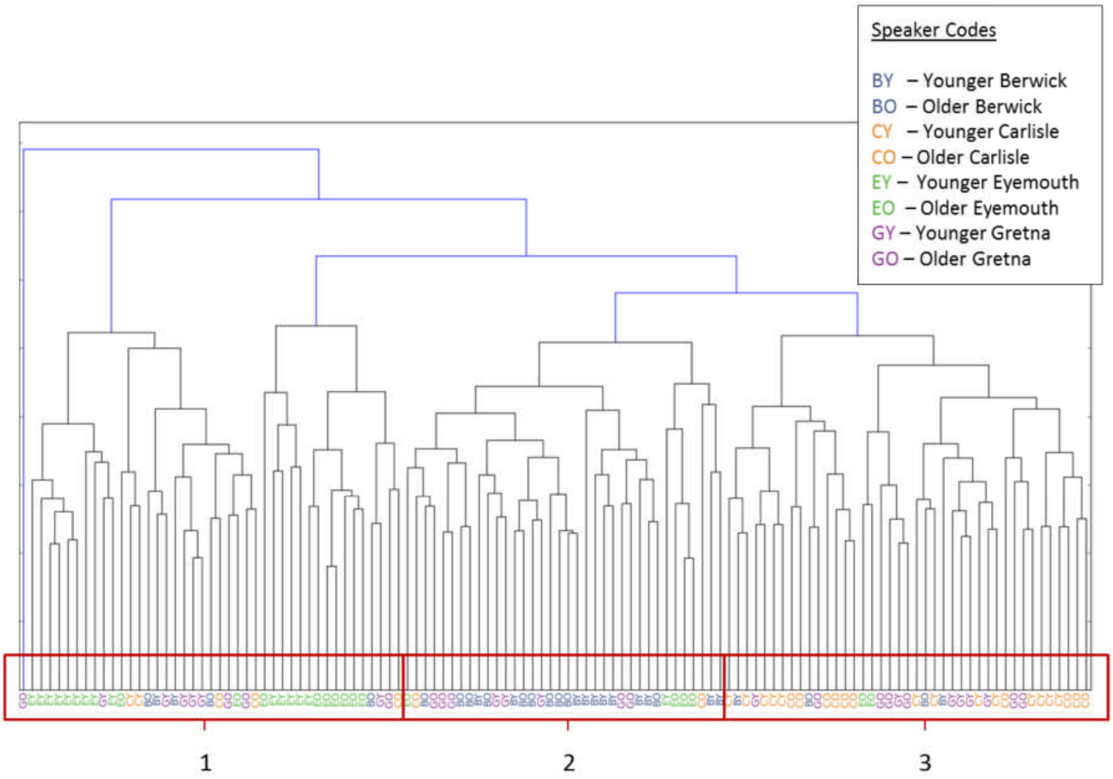


Figure 3.3: Agglomerative hierarchical cluster analysis of 120 AISEB speakers with observational divides

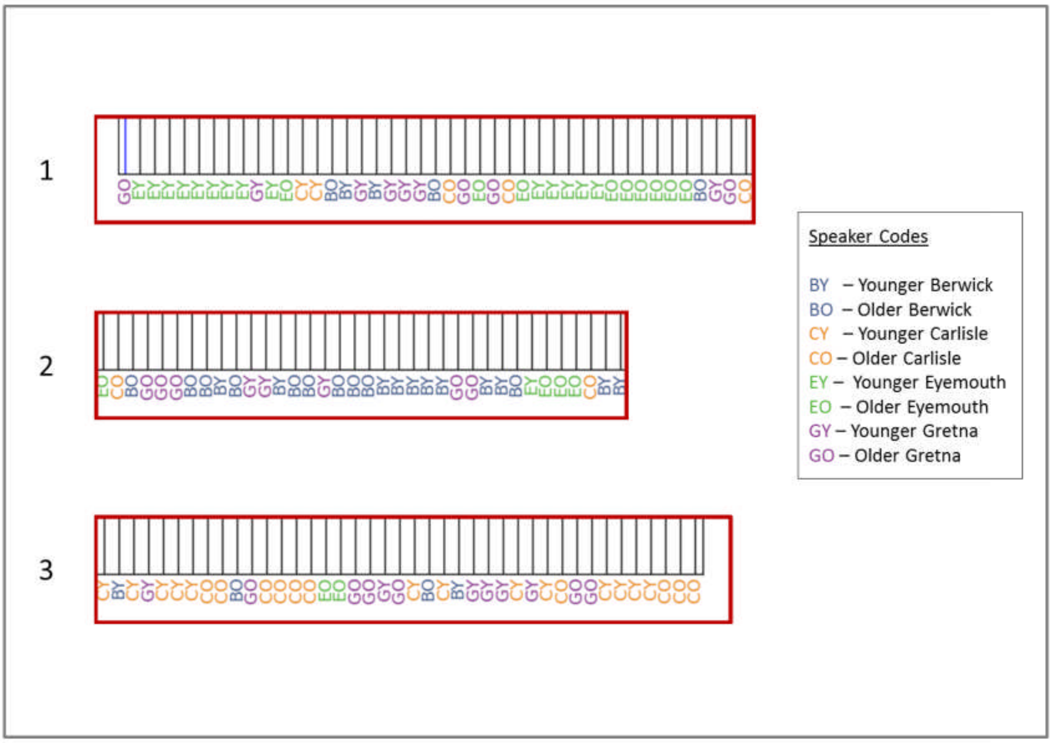


Figure 3.4: Close-up windows of the cluster analysis

Taking a closer look at these divisions in close-up windows, we can see the distinctive variety spoken in Eyemouth by the clear Eyemouth clusters towards the left of the dendrogram (shown more clearly in block 1 of figure 3.4). Likewise, block 3 of figure 3.4 shows that Carlisle speakers are largely found towards the right of the dendrogram, although there appears to be much more interference from other varieties. This could be seen as supportive of preliminary observations of the AISEB varieties in Llamas (2010). With respect to coda /r/ again, Llamas (2010: 234) speculates that 'as regards to the Scottish-English border, the divide is stronger and more stable on the east side both linguistically and in terms of social categorisations'. By scanning the x-axis of the cluster analysis above, we can get a feel that Eyemouth and Berwick appear to have a more stabilised and predictable posting. In a general grouping shown in figure 3.4(2), Berwick speakers appear to be straddling between the Eyemouth and Carlisle clusters, which is suggestive that a continuum of these varieties is indeed in place. In contrast, Gretna speakers are found spanning across the whole length of the x-axis without forming a distinctive cluster. Again, this can be put down to the location's later date of establishment.

Overall, this chapter has further investigated an ACCDIST-based model's sensitivity on varieties thought to be more similar than those used in past studies. This has opened up the possibility of using the Y-ACCDIST system in forensic applications where it is the aim of the practitioner to obtain speaker properties to do with place of origin, or even a general age group. This chapter has established that the Y-ACCDIST system can work to some extent on the geographically-proximate AISEB varieties. Chapter 4 builds on the baseline system presented so far to investigate whether adapting the system in different ways can improve recognition performance. First, however, chapter 4 looks at the possibility of using the Y-ACCDIST system on spontaneous speech data.

Chapter 4

System Modifications

The previous chapter has demonstrated the Y-ACCDIST system's baseline performance on the AISEB data. It showed results revolving around how the system performed on different groupings of speakers and how an accent corpus could be analysed using the system. While chapter 3 mainly dealt with changes in the sociolinguistic groupings of the data, this chapter will assess accent recognition with changes to the system and some of its limitations.

Firstly, this chapter will address one key feature of the Y-ACCDIST sytem which differs from previous ACCDIST-based systems: segmental context independency. This is explained further in section 4.1 below, but its appeal is down to the Y-ACCDIST system's potential applicability to incomparable speech content. The Y-ACCDIST system is therefore tested on spontaneous speech data in section 4.2 as this is of relevance to forensic casework. Together, sections 4.1 and 4.2 therefore address the second research objective given in the *Introduction*.

Also of relevance to forensic casework is the quantity of data required for optimum classification performance. This is addressed in 4.3. In the baseline Y-ACCDIST system, midpoint MFCC vectors are used to represent the segmental data. Different ways of representing the segments are trialled. A different feature vector type, Perceptual Linear Predictive coefficients (PLPs) are trialled in 4.4, while an attempt to capture the dynamic nature of vowels is made in section 4.5. This is done by using concatenated MFCC vectors taken from two separate temporal points within the phoneme segment. It becomes apparent that when dealing with a smaller number of varieties which are spoken within a more limited geographical area, some segments are more important than others when distinguishing between varieties. By focussing on /r/, section 4.6 shows how the addition of a single phoneme can have an effect on recognition results.

4.1 Segmental Context Dependency

The present thesis refers to *segmental context dependency* with respect to a speech segment's phonological environment. By returning to figure 2.1 in chapter 2, we can recall symbols *a*-*e* representing speech segments which form the ACCDIST matrix. In past systems, these speech segments are not necessarily individual phonemes. These speech segments in previous ACCDIST-based systems have been sequential units of phonemes, making these segments more context-dependent and less frequent in a speech sample. Further details and examples are given in section 4.1.1 below. One of the key features of the Y-ACCDIST system is its segmental context independency. ACCDIST as a model relies on the relationship between different segments within a single speaker's speech sample. To be able to then compare speakers to reference populations, or even other speakers, ACCDIST requires the same segments to be present in all speech samples involved. In the previous ACCDIST-based systems discussed in this paper (Huckvale, 2004, 2007a; Hanani *et al*, 2011, 2013), highly context-dependent speech segments are used which of course have repercussions on the sort of speech material required. When highly context-dependent segments are used in the model, identical read prompts or passages are needed. In other words, the more context-dependent the speech segments are, the more comparable the content of the speech samples need to be. When considering the forensic application, there is obviously an extremely limited number of contexts where comparisons between identical read passages can be made. We cannot assume full co-operation from individuals if we were to ask for a recorded reading. The advantage of using context-dependent segments is that they are expected to reduce the effects of coarticulation. It is well known that neighbouring speech segments affect the realisation of a phone. This section compares the performance of highly context-dependent word-context vowel segments (used in Huckvale (2004, 2007a)), context-dependent phone-vowel-phone triphone segments (used in Hanani *et al*, (2011, 2013)), and finally context-independent phoneme segments (used in the Y-ACCDIST system). Each of these types are illustrated and the performance of all three segment types are trialled on the reading passage recordings from speakers of the AISEB corpus.

4.1.1 Segmental Context Types

This section gives a demonstration of each segmental type. Taking the first clause, ‘Fern was a nurse from Harrogate’, from the reading passage, each segmental context type is illustrated below. Only vowels (excluding schwa) are taken into account in this section, in line with previous studies. These can of course alter depending on the phonemic transcriptions used. Section 4.6, however, shows the effects of including phones beyond just the vowel inventory.

Vowel Segment Type	**Specific segments found in *Fern was a nurse from Harrogate***	**No. unique segments**
Word-context vowels	vowel in *Fern* vowel in *was* vowel in *nurse* vowel in *from* vowel in first syllable of *Harrogate* vowel in third syllable of *Harrogate*	6
Phone-vowel-phone triphones	/ɜ/ in /fɜn/ /ɒ/ in /wɒz/ /ɜ/ in /nɜs/ /ɒ/ in /rɒm/ /a/ in /har/ /eɪ/ in /ɡeɪt/	6
Context-independent vowels	/ɜ/ in both *Fern* and *nurse* /ɒ/ in both *was* and *from* /a/ in *Harrogate* /eɪ/ in *Harrogate*	4

Table 4.1: Demonstration of different segmental types

During the ACCDIST process, when a unique segment is reencountered, within a speech sample the segments are simply clustered together to form an average MFCC representation of that segment. Euclidean distances are then calculated from these average representations to form an ACCDIST matrix. As we can see from the table above, this happens more regularly in the case of context-independent phonemes than for the word and triphone context segments. When only counting all word-context vowel segments (excluding schwa) the reading passage, *Fern's Star Turn*, contains a total of 342 unique segments. This can be compared with the 15 unique vowel segments within the context-independent phoneme inventory (this can change depending on the particular phoneset used, which is usually dependent on accent). In effect, these counts impact on the size of the ACCDIST matrices computed in the accent recognition process.

When considering the effects of coarticulation on the realisation of vowels, it could be expected that the more detailed word-context segments will perform very well, grasping the specific quality of a vowel. In the case of the ABI corpus, this is certainly shown to be true as Huckvale (2007a) presents an impressive recognition rate of 92.3%. However, in the case of more localised varieties, this section shows that this level of detail is unnecessary and in fact produces unwanted 'noise' in the model. Not only would context-independent phonemes be the preferred choice for comparing accents across incomparable spoken content, but it also appears to be the preferred choice when comparing accents across comparable data as well.

4.1.2 Methodology

To investigate the effects of segmental context-dependency, two further variant Y-ACCDIST systems were built to compare with the context-independent baseline system. For all three of the systems the recorded reading passage for each speaker was forced aligned and the midpoint MFCC vectors for every vowel token were extracted. Vowel segment contexts were then coded for the word-context and triphone systems. For each of the variant systems, ACCDIST matrices were generated for all 120 speakers. To mirror segmental quantities of previous works, only the first sixteen divides of the reading passage (which is less than half of the entire passage) were taken, which left approximately 150 unique word-context vowels for each speaker. This allows for comparison with Huckvale (2007a), who reported that approximately 140 word-context segments were incorporated into his system. Only approximate quantities are given because some portions of a speaker's sample may have been discarded as a result of background disruption or significant reading errors. These are factors which would have had a negative impact on the forced aligner. For comparison, each of the variant systems were trained and tested on the same first portion of the reading passage. Hanani *et al* (2013) used the 105 most frequent segments which is the same as what has been employed here for the triphone variant system. Recall that recognition rates of over 90% have been reported using these context-dependent segments on the ABI corpus. As mentioned above, the phoneme-based system only incorporates 15 different types. Each of the three variant systems conducted a four-way recognition task on all 120 speakers, following the testing protocol outlined in section 2.2.2, classifying them into AISEB's four geographical locations.

4.1.3 Results

The recognition rates of each variant system are shown below:

Segmental context-dependency variation	**% Correct**
Word-context vowels	74.2
Phone-vowel-phone triphones	75.0
Context-independent vowel phonemes (baseline)	76.7

Table 4.2: Y-ACCDIST's performance using different segmental types

A four-way classification task naturally carries a 25% chance recognition rate (i.e, if the model was not working the way we predicted, 25% is approximately the recognition rate we would expect if the system were assigning categories to speakers at random). Each of the three variant systems achieve recognition rates well above 25%. Marginal differences exist between the three variant systems. However, a hierarchy seems to be in place. The more context-dependent the segmental type is, the lower the recognition rate. This hierarchy is more evident when these experiments are conducted on an earlier version of the Y-ACCDIST system presented in Brown and Watt (2014) and Brown and Wormald (2014). The only difference between the earlier system and the present system is the difference in classification procedure. It appears that the current Y-ACCDIST system is less sensitive to such modifications than an earlier version. As described in section 2.1.2, the current Y-ACCDIST system uses Support Vector Machines to classify speakers, whereas the earlier system used Pearson product-moment correlation to compare and classify speakers. The earlier system processes and classifies the AISEB data in the same way and the pattern above appears to be more significant and worthy of note:

Segmental context-dependency variation	**% Correct**
Word-context vowels	50.0
Phone-vowel-phone triphones	56.7
Context-independent vowel phonemes (baseline)	60.0

Table 4.3: An earlier Y-ACCDIST system's performance using different segmental types

4.1.4 Discussion

The above results suggest that at the very least, more context-dependent segments do not entail better recognition results for this particular corpus. Coupled with results from an earlier system, there is even suggestion that the more context-dependent the segments are, the lower the recognition rate. This outcome is put down to a larger number of context-specific segmental units seemingly creating 'noise' within the model. This may well be a side effect of classifying geographically-proximate varieties. Hanani *et al* (2013: 67) identify the reduction in vowel context information, when comparing their triphone segment system to Huckvale's (2007a) word-context system as a likely disadvantage to performance. Interestingly, the results shown above do not align with findings made by Huckvale (2007a: 266-267), where he built a comparable system to ACCDIST-based systems. The key difference was that vectors of formant values replace MFCCs, and in the formant-based system he pools together averages across phoneme formant frequency values in a similar way to the pooling of MFCC vectors in the context-independent baseline Y-ACCDIST model. Huckvale reports that there was a reduction in best accent recognition rate in the averaged formant frequency system, dropping from 89.4% to 79.6%. In the accent recognition task here however, it is shown that this pooling of speech segments into phonemic categories is not necessarily a degrading factor. The recognition results displayed above may even suggest a spectrum of segmental context-dependency and its interaction with a similarity scale of the particular varieties involved.

Another likely contributing factor is the frequency of segmental units. As well as adding larger numbers of irrelevant rows and columns to ACCDIST matrices, the use of more context-dependent units reduces the number of segments which produces the average representative feature vector. This minimises room for error which may be required, especially when using automatic segmenters (which are not wholly reliable) as part of the overall system.

In sum, this section has revealed that the context-independent phoneme segments are the preferred choice when conducting recognition tasks among geographically-proximate accents.

4.2 Spontaneous Speech

The results from the above section show the recognition advantage of using context-independent phoneme segments. They appear to outperform more context-dependent segments in a geographically-proximate accent recognition task. Phoneme segments have a practical advantage as well which is that they are more likely to create segmental overlap between spontaneous speech samples (incomparable spoken content). This overlap is essential for classification. As the present research is conducted with the forensic application in mind, where the likelihood of receiving comparable speech samples is low, the context-independent phoneme-based system (Y-ACCDIST) has been tested on spontaneous speech.

As well as the practical motivations, it is also critical to test the system on spontaneous speech as more realistic data. Watt *et al*, (2014: 91-93) show a difference in their results from /r/ variant analysis of the AISEB varieties. They show that phonetic realisations do in fact alter between read speech and spontaneous. Particularly in the Gretna speakers, they point out the higher proportion of rhotic realisations in word list recordings compared with the same speakers' conversational recordings. Such differences may change recognition performance.

In terms of automatic systems, past ACCDIST-based systems have only been tested on comparable read speech. Hanani *et al* (2013) predict a rise in accent recognition rates when using spontaneous speech data as they suspect that paralinguistic information, like accent, would be more explicit. The results shown in this section do not support this prediction, but it is likely that quantity of spontaneous speech data plays a role in accent recognition performance. In Woehrling *et al* (2009), automatic accent classifiers were tested on French spontaneous speech data. However, in their study, while comparing with approximately 3 minutes of read speech per speaker, they used approximately 13 minutes of spontaneous speech. Using these unequal quantities brought about similar results between the two modes of speech.

In this section, approximately the same duration of speech data per speaker was used for both the spontaneous speech system and the reading passage system (approximately 3 minutes).

4.2.1 Methodology

For each speaker, the whole reading passage is processed to gain a recognition rate for comparable data (which is approximately 3-4 minutes long), and to gain a result for incomparable data, the first three minutes (approximately) of unoverlapped speech from the informant's interview session is taken. The interview sessions involved the researcher asking questions about the informants' local areas and their opinions regarding national identity and such topics. It was not uncommon for informants to go off topic during interview sessions. However, conversations remained fairly neutral regarding subject and temperament. Under both conditions, speech samples were automatically aligned. Each speaker for each experiment was taken from the reference population and treated as the unknown test speaker (again, following the testing protocol).

4.2.2 Results

The recognition rate generated from spontaneous speech data is directly compared with that generated by the reading passage data below:

Speech data type	**% Correct**
Reading Passage (comparable)	86.7
Spontaneous (incomparable)	52.5

Table 4.4: Comparison of recognition rate on comparable and incomparable data

4.2.3 Discussion

Of course, there is a large discrepancy between the system's performance on comparable and incomparable data, but the model appears to be working to some extent since the recognition rate still sits well above chance expectations. This task is likely to require larger quantities of speech data than reading passage data to achieve similar recognition rates. This is due to varying coarticulation effects and their impact on the averaging processes which feature in the Y-ACCDIST system. Recall from chapter 3 that average midpoint MFCCs are generated to represent each phoneme in the inventory. With more variation in the phoneme contexts in which phoneme tokens occur, more variation in phonetic realisations (and therefore resultant MFCC vectors) is expected from the broad ranging coarticulation effects. Such variation is obviously reduced under the shared reading passage condition as wide contextual variation is controlled between speakers. The sort of variation brought about by spontaneous speech, then, means that more tokens per phoneme are required to produce a 'stabilised' average representative midpoint MFCC vector. 'Unstable' average MFCCs obviously then have pejorative effects on following calculations, namely calculating the Euclidean distances to form a speaker's ACCDIST matrix. Even though the results in this section do not support Hanani *et al*'s (2013) predictions regarding improved recognition on conversational data, here it is suggested that a larger quantity of data may be required to make full use of the paralinguistic information expressed.

Also of course, like read prompts in the general landscape of speech corpora, the reading passage was specifically designed to elicit accent differences. It includes a distribution of sounds which is perhaps not typical of conversational phonemic content. Providing a variety of segments like this may also contribute to a higher performance achieved in the case of read speech.

4.3 Quantity of Speech Data

For the sake of clearer comparison with previous ACCDIST studies, the results showing the effects of segmental context dependency in table 4.2 have been generated using a smaller quantity of speech data than that used in chapter 3. In fact, less than half of the reading passage was used. We can compare the recognition rates between the two different quanitites of speech data below. These are the results generated when taking into account all the vowel segments (excluding schwa).

% Correct using the whole passage (3-4 mins)	**% Correct using approx. one third of the passage (approx. 1 min)**
86.7	76.7

Table 4.5: Recognition rates using two different quantities of speech data

This puts into question approximately how much data is required to achieve optimum accent recognition performance. Using a HMM-based accent recognition system to distinguish between non-native accents of English, Arslan and Hansen (1996: 363) show the effect of ‘utterance’ length on accent classification scores. ‘Utterance’ in Arslan and Hansen’s study refers to strings of randomly selected words from their database of recordings, rather than naturally produced word sequences typical in conversation. They show that by increasing the number of words per test sample, accent recognition plateaus at 93% for utterances beyond seven words long. It is also speculated that individual word length has an effect on accent recognition. They report that when the system is faced with longer words, such as *aluminium* and *communication*, generally speaking classifcation rate is higher. In section 4.6 of this thesis, however, it would be suggested that the segmental content itself has an impact on recognition rate (i.e, the presence or absence of particular phonemes can determine classification success rates). It can also be argued that it would also depend on the particular varieties being distinguished between.

As it stands, it seems that the current system performing on the AISEB varieties requires more data to perform at the level it does than Hanani *et al*'s (2011, 2013) system and Huckvale's (2004, 2007a) system. To generate Hanani *et al*'s (2011, 2013) highest recognition rate, 95.18%, on the 14-way recognition task, they used a shorter reading passage lasting between 30 and 45 seconds in duration. The difference in performance, while also considering the difference in reading passage lengths, is likely to be an effect of the different corpora and varieties in use. Since we expect that the AISEB varieties are more similar than the ABI varieties, we expect that discriminatory elements would crop up throughout a speech sample at less regular intervals. This leads to requiring a longer sample with more segments to offer enough discriminatory information.

4.3.1 Methodology

Using the reading passage speech sample, recognition tasks were conducted with varying proportions of it. Beginning with the first 10%, the Y-ACCDIST system was run to recognise each speaker's geographical location. Percentages were based on proportion of segmental tokens within the reading passage. In total, *Fern's Star Turn* contains 2074 phoneme tokens and these were the measure of speech data quantity. Cumulative increments of 10% were used to train and test the system.

4.3.2 Results

The results are represented in the graph below:

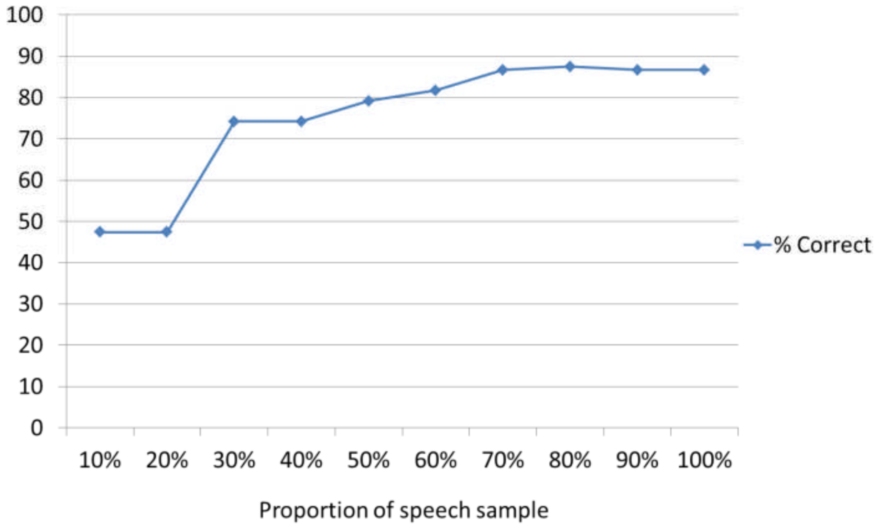


Figure 4.1: Accent recognition rates with varying quantities of speech sample

4.3.3 Discussion

The graph shows a general rise in success rate as the proportion of the reading passage used increases. Correct classifications start to plateau at around 70%, which equates to roughly requiring 1450 phoneme segments. However, it is proposed here that segmental quantity is not the only criterion for a speech sample and its probability of being correctly classified. It is suspected here that it is also dependent on which specific phones are present within the sample (and adding to the list of considerations, this will of course depend on the varieties involved). The issue of phonemic content is addressed in section 4.6.

Another area regarding data quantity which is of interest to forensic speech science is the number of speakers required in the reference dataset. This thesis has already shown the importance of collecting 'fresh' speech data to form our reference populations, but it would of course be useful to have a criterion stating how many speakers are required to form a stable and representative sample of a single variety. Previous results within this thesis may suggest that the answer to this would depend on the variety. We have already seen in chapter 3 that different varieties yield different likelihoods of correct speaker classification. More specifically, from the results in chapter 3, it is less likely that a speaker from Gretna will be correctly classified than a speaker from Eyemouth. It is likely that such a trend will be echoed in the number of speakers required to account for speaker variation within a variety. This topic requires further research.

4.4 Feature Vectors

It is obviously crucial to represent the speech signal as effectively as possible for an application. In order to interpret a speech signal, we can take spectral information in a number of forms. Each form is assumed to characterise the speech signal and deliver specific information about the shape of the vocal tract at a particular point in time. In phonetic analyses formant frequency values are traditionally taken. These can often feature in forensic casework (e.g, in Ash (1988)). However, here we will look at how automatically extracted feature vectors perform. Commonly used are Mel Frequency Cepstral Coefficients (MFCCs) particularly for automatic speech recognition tasks. These are the feature vectors implemented in the baseline Y-ACCDIST system and a description of their composition is given in section 1.2.1. Another prominent vector type for front-end signal processing are Perceptual Linear Predictive coefficients (PLPs) (Hermansky, 1990). These are thought to offer a more noise-robust alternative to MFCCs (Bach *et al*, 2011). However, often both varieties of vector are trialled in the emergence of new recognition systems.

PLPs are described below by highlighting the key features which distance them from MFCCs.

4.4.1 Perceptual Linear Predictive Coefficients

Perceptual Linear Predictive coefficients (PLPs) are another common feature vector used in speech technology. The general aim of PLPs is the same as MFCCs. Meyer and Kollmeier (2009), however, highlight the three key differences which separate the two types of feature vector. One key difference of PLPs is the adoption of the Bark scale, rather than the mel scale. Both are employed for their approximation of the human perceptual system, but it is thought that the Bark scale brings a finer representation to the process. Another difference is the nature of the pre-emphasis which takes place in the extraction of PLPs. Hermansky (1990) developed an adapted pre-emphasis technique called *equal loudness preemphasis* which is thought to offer a more representative treatment of the power spectrum of that of human hearing (Hönig *et al*, 2005: 2999). Finally, an additional key difference is the methods used for compression. The compression strategy in the formation of MFCCs is a logarithmic scale, whereas PLPs use 'Stephen's Law'. The combination of these differences in speech representation have often shown PLPs to outperform MFCCs when assessing the feature vectors in noisy recognition tasks (e.g, Bach *et al*, 2011). It is due to their reputation of outperforming MFCCs in noise that it seems worthwhile to trial PLPs here in this accent recognition model. Since the Y-ACCDIST system has been developed with the forensic application in mind, it is reasonable to enquire whether these more noise-robust features can function in this particular model. Obviously, background noise features heavily in forensic speech casework and therefore it makes sense to take the most noise-robust measures if possible.

4.4.2 Methodology

Using the same automatic alignments, two variant Y-ACCDIST systems were compared, differing only by feature vector. One system (the baseline system) took the midpoint MFCC vectors, while the other took midpoint PLPs to represent phoneme segments. Other aspects of the methodology followed the testing protocol. Both types of feature vector were extracted using the HTK toolkit.

4.4.3 Results

The PLP variant Y-ACCDIST system is compared with the results generated by the baseline Y-ACCDIST system under the different data configurations trialled in chapter 3 (table 3.1). The results are logged in the table below:

Recognition Task	**MFCCs (% Correct)**	**PLPs (% Correct)**
4-way classification according to geographical origin, all speakers, N=120	86.7	87.5
4-way classification according to geographical origin, younger speakers, N=60	83.3	75.0
4-way classification according to geographical origin, older speakers, N=60	83.3	83.3
8-way classification according to geographical origin and age category, N=120	69.2	73.3

Table 4.6: Comparison of recognition rates using MFCCs and PLPs

4.4.4 Discussion

When looking at the more general recognition rate (the first recognition task), involving all speakers and categorising them only by location, we can see a slight improvement in recognition rate, from 86.7% to 87.5%. However, when conducting the four-way recognition task within the specific age groups, we see either a reduced or sustained recognition rate. Of note though, is the promising rise from 69.2% to 73.3% in the eight-way recognition task. Even though there does not appear to be consistency among the above results, they do suggest that alternative feature vectors, such as PLPs, are worth testing and bringing into the analysis. Obviously, particularly for the forensic application, it is preferable to discover the optimum configurations at all stages of the process. Because of PLPs' reputation of being a more noise-robust feature vector, it would be a useful experiment to test the model's performance using PLPs and MFCCs on low-quality recordings. From the results above it appears that PLPs offer a marginal improvement for the broader accent recognition task.

4.5 Dynamic Phoneme Representation

Past systems have also differed in terms of where feature vectors are extracted from within the phone segment. Here, we have seen the midpoint used. Of course, in accent comparisons we might consider capturing the dynamic quality of the vowels as some varieties can be distinguished this way. For example, some areas of Yorkshire may typically include a monophthongal GOAT vowel, whereas plenty of other varieties around the country realise this vowel as a diphthong. For this reason, Hanani *et al* (2011, 2013) and Huckvale (2004, 2007a), when distinguishing between the 14 varieties taken from locations across Britain, take average MFCCs from each temporal half of the vowel. These two feature vectors are then concatenated to form one long feature vector to represent a segment. These dynamic vectors then go on to be averaged (if required) and the ACCDIST process continues.

The Y-ACCDIST system will be trialled using a similar dynamic approach. Here we will use what will be called 25-75 MFCC vectors, where each segment of interest will be halved and the midpoint of each half will be extracted and concatenated together to form a longer vector. Keeping the points at which MFCCs are extracted consistent aims to capture more refined differences in the dynamic quality of a vowel.

The two feature vector measurement types (midpoint MFCCs and 25-75 MFCCs) are more clearly presented in the diagram below:

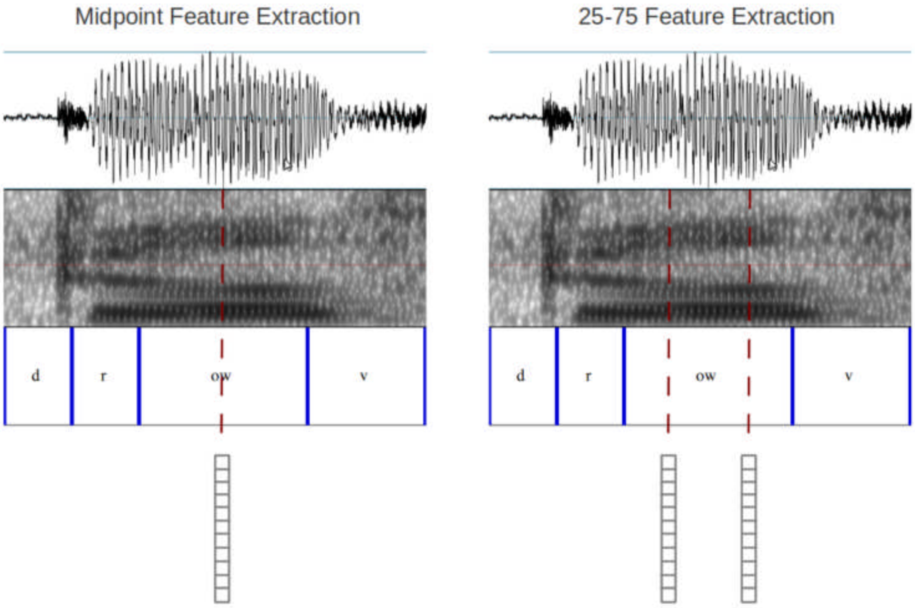


Figure 4.2: MFCC extraction points from the vowel in 'drove' for the midpoint and 25-75 variant systems

The use of these 'dynamic' effects is also witnessed in forensic speech science. McDougall (2004) makes use of formant dynamics in the context of individual speaker comparisons by focussing on Australian English speakers' /aɪ/ vowel. The findings indicate that more contour points throughout the vowel improves classification in the case of characterising speakers (McDougall, 2004: 117). It is also claimed that just taking the midpoints of the vowel is not sufficient for speaker characterisation as 'most speaker-pairs overlap considerably' (McDougall, 2004: 110). This baseline Y-ACCDIST development decision to use midpoint MFCCs was influenced by the earlier version of the system (where correlation was used to classify speakers rather than SVMs). When developing the system for previous studies (Brown and Watt, 2014; Brown and Wormald, 2014), taking 25-75 MFCCs either had no effect or slightly reduced success rates. Maintaining consistency with the effects of other system modifications, this was put down to creating more 'noise' in the model by introducing unnecessary information. This effect is confirmed by the current system in the results below.

4.5.1 Methodology

To compare a dynamic measurement Y-ACCDIST system with the baseline, for each phone token in the speech samples, an MFCC was extracted 25% of the way through the vowel and then 75% of the way through the vowel. These were concatenated to represent a single phone. These concatenated MFCCs went on to be grouped according to phoneme and then a mean concatenated MFCC vector was calculated to represent each phoneme within the speakers' phoneme inventories. The ACCDIST process continued as explained in chapter 2.

4.5.2 Results

A comparison of the two variant Y-ACCDIST systems is displayed below:

Baseline midpoint MFCC phoneme system - all vowels, N=120 (% Correct)	**25-75 concatenated MFCC phoneme system - all vowels, N=120 (% Correct)**
86.7	83.3

Table 4.7: Comparison of midpoint MFCCs and 25-75 MFCCs

4.5.3 Discussion

Despite success in previous systems, using midpoint MFCCs appears to bring about a higher recognition rate than taking MFCCs from two segmental halves. Above we see a drop in successful recognitions when concatenating 25-point and 75-point feature vectors. This is likely to be due to the high degree of similarity between the AISEB accents. This therefore means these varieties have more vowel realisations in common, leading to a greater degree of overlap between ACCDIST matrices. It is assumed that this leaves fewer matrix elements with discriminative power. By creating longer feature vectors to represent each phoneme, it is likely that this introduces 'noise' to the model by adding unnecessary information.

The suggestion of additional 'noise' also implies that removing individual phonemes unhelpful to recognition may also improve results. Equally, adding particularly useful phonemes to the process may prove beneficial. The issue of segmental selection is addressed in the next section.

4.6 Segmental Selection

Pronunciation lexicons play a vital role across many speech technology applications. These are dictionaries which hold phonemic transcriptions of all the words systems may face. This is of course crucial information to speech recognition and synthesis systems. The lexicon also has a huge role to play in the Y-ACCDIST system. The Y-ACCDIST system heavily relies on the categorisation of individual phone tokens into phonemic types. This categorisation and classification is required for the formation of an ACCDIST matrix (i.e, the phonemes present in the pronunciation lexicon form the rows and columns of the ACCDIST matrices). As we know, different accents may have different phonemic inventories. To therefore maximise system performance, it is important to identify the best pronunciation lexicon for the varieties involved. Interestingly, the results generated in this thesis so far have been backed by the VoxForge lexicon[1], which is a pronunciation dictionary of General American pronunciations. Performance of a British English pronunciation dictionary (the British English Example Pronunciation (BEEP) dictionary[2]) has also been tested. For the AISEB varieties concerned, the VoxForge lexicon performs slightly better for this particular accent recognition task (as the table below shows).

General American phonemic transcriptions (VoxForge dictionary), N=120	**British English phonemic transcriptions (BEEP dictionary), N=120**
86.7	85.8

Table 4.8: Comparison of pronunciation dictionaries designed for two different accents of English: American English and British English

The results above do not necessarily show a great difference. Nonetheless, the VoxForge lexicon was selected due to its slight lead in recognition results. Also, unlike the BEEP dictionary, the VoxForge lexicon (providing General American phonemic transcriptions) obviously incorporates rhoticity into its transcriptions. This is a feature of interest when it comes to distinguishing between English and Scottish varieties like we are here.

As well as the phonemic transcriptions themsevles, Huckvale (2007a) acknowledges that the removal or addition of individual speech segments can alter the system's performance. For the current application, it is imperative that the optimum combination of segments is included in the classification task. Equally though, including some segments may produce 'noise' in the model. Brown and Wormald (2014) investigated the value of selecting the most telling speech units for an ACCDIST-based system to perform on a specific dataset. Here, the earlier version of the Y-ACCDIST system was trained and tested on a corpus of Panjabi-English speakers in Bradford and Leicester. It was found that to distinguish between two age groups (older and younger) within second generation Panjabi-English speakers in each city, it was absolutely crucial to only specify phonemic segments which were of value. As just one example, to distinguish between older and younger Panjabi-English speakers in Bradford, a two-way classification task, incorporating the entire vowel inventory only correctly recognised 5 out of 20 speakers. However, by using a sociolinguistically considered and restricted set of vowel phonemes (FLEECE, GOOSE, GOAT and BATH) 14 speakers out of 20 were correctly classified. These decisions were based on careful formant analyses by the second author. Even though such analyses are laborious, the findings showed how formant values within accent corpora can reflect the performance of the Y-ACCDIST system, based on which vowels are included.

In Brown and Watt (2014), again an earlier version of the Y-ACCDIST system, the baseline system simply included all vowel phonemes (-schwa) and the approximants /r/, /w/ and /l/. The decision for this phonemic combination was based on a comparison of results during development. A classification rate of 72.5% was achieved on the four-way recognition task of the AISEB varieties when just the vowels were incorporated, and 79.2% with the addition of the approximants. It is of course possible to trial any number of combinations and observe results. This section, however, focusses on /r/ alone as a carefully selected phoneme based on the particular varieties involved (namely Scottish and English accents). Additionally, this thesis has already identified past sociolinguistic literature which has pointed out rhoticity as a particular feature of interest among the AISEB varieties (Watt *et al*, 2014; Llamas, 2010). Below looks at the incorporation of /r/ into the current Y-ACCDIST system.

4.6.1 Rhoticity

Recent sociophonetic research has been focussing on the gradience of postvocalic /r/ realisation among speakers. Lawson *et al* (2014) address this gradience as a development of more traditional rhoticity studies, such as Labov's famous New York study (Labov, 1966). Labov's findings came about as a result of documenting the presence or absence of /r/ and any realisations which were questionable were discarded. However, Lawson *et al* (2014) show that the rhotic 'strength' has a story to tell among Scottish varieties, with particular reference to social class (i.e, Lawson *et al* report derhoticisation in working-class speakers from Glasgow). We therefore need to be working with a continuum of these realisations. Watt *et al*, (2014) indirectly comment on the the potential limitations of a manual IPA-based transcription methodology when classifying /r/ variants. They say, in relation to their study's methodology, that a 'classification scheme at this level of detail is unwieldy, so infrequent variants were pooled into broader variant categories' (Watt *et al*, 2014: 86). The ACCDIST metric, in its continuous nature, can model the realisation spectrum of a phoneme without having to categorise realisations or deal with borderline cases. Equally, if clear-cut categorical distinctions are in place between varieties, the ACCDIST model can cater for these too, without requiring knowledge of fine-grained differences beforehand. Overall, the way in which ACCDIST works allows it to make more precise distinctions than more categorical analyses.

Huckvale's (2004, 2007a) and Hanani *et al*'s (2011, 2013) ACCDIST systems, repeatedly referred to throughout this thesis, only take vowel segments into account. In the case of Huckvale's word-context variables, all vowel MFCCs are incorporated (excluding schwa). Hanani *et al*'s system only draws feature vectors from the phone-vowel-phone triphones, where again, only vowels are brought into processing. More localised accent recognition tasks, like the ones conducted here, can afford to specify valuable segments beyond the vowel inventory. Speakers of the AISEB varieties, for example, would be expected to group into their respective geographical classes on more occasions with the introduction of /r/. Rhoticity can play a key role in distinguishing between varieties and in the case of AISEB, could be expected to resolve the varieties either side of the Scottish-English border.

4.6.2 Results

The table below shows how the addition of the /r/ phoneme helps to distinguish between the four AISEB varieties. When adding /r/ to the process, firstly /r/ was included as a categorical phoneme which pools together both onset and coda /r/. /r/ was then split into the two categories of onset /r/ and coda /r/ to be treated as separate segmental categories in ACCDIST calculations. These additions and recategorisations are observed below:

Phoneme Configuration	**% Correct**
All vowels (- schwa)	86.7
All vowels (-schwa) + /r/	89.2
All vowels (-schwa) + onset /r/	86.7
All vowels (-schwa) + coda /r/	85.8
All vowels (-schwa) + onset /r/ + coda /r/	90.0

Table 4.9: The effects of the /r/ phoneme on recognition rates, distinguishing between speakers of four different locations

4.6.3 Discussion

Given Watt *et al*'s (2014) findings regarding rhoticity among the AISEB varieties (discussed in section 3.1), the results above are perhaps not surprising. They confirm the suspected distinguishing effects of the /r/ phoneme's presence in the ACCDIST model. As a general distinguishing feature separating English and Scottish accents, we would expect the results including coda /r/ to offer an improvement. Here we actually see a small drop in recognition rate when coda /r/ is added alone. However, it seems as if the general addition of /r/, be it separating /r/ into two syllable context categories or not, brings an overall improvement to recognition rates. Sociolinguistically, this may be suggesting that there is an overall systematic patterning of /r/ realisation among the AISEB varieties.

Since a marked improvement can be witnessed with the addition of /r/, an extra experiment was run to observe whether it can improve the spontaneous speech condition:

Phoneme Configuration	**% Correct**
All vowels (- schwa)	52.5
All vowels (- schwa) + /r/	59.3

Table 4.10: The effect of the /r/ phoneme on recognition rates of spontaneous speech

The table above confirms that /r/ also has a positive effect on the recognition of spontaneous speech. It is therefore expected that further work on segmental selection can push the Y-ACCDIST system towards forensic use.

Obviously, /r/ was selected based on justifications which can be gathered from sociolinguistic literature. Rhoticity is thought to be a feature of Scottish English, but not usually of English English. However, in the context of forensic speech science, it would be an ideal feature of the Y-ACCDIST system to be applied to varieties which are not necessarily in sociolinguistic literature and little is known about them. From this point of view, it would be preferable if the system could identify which specific phoneme segments to include and which to filter out. A future direction for the Y-ACCDIST system is to look towards *dimensionality reduction* techniques. Essentially, dimensionality reduction is a group of techniques in machine learning which aim to identify the most distinguishing features of training data, so these can remain in the process while unhelpful features to the task can be dismissed - therefore reducing ‘noise’. These can include methods familiar to statistics like principle component analysis or linear discriminant analysis. However, it is important to discover the most effective one for the specific process. Specifically in terms of the Y-ACCDIST system, it would be a great system feature to be able to initially include all phonemes within the inventory (both vowels and consonants), and for the system to identify the most telling segments for the specific recognition task involving a specific set of accent varieties. This area of investigation requires further work, but would be extremely useful to the forensic application.

Chapter 5

Discussion and Further Directions

The previous chapters have assessed the Y-ACCDIST system's performance under a number of settings and data configurations. However, to fully explore the Y-ACCDIST system's potential, there is clearly scope for further research. This section brings some of these potential research directions to the surface.

The bulk of the forensic practitioner's workload is the speaker comparison task, where two different recordings are compared to assess the likelihood that it is the same speaker in each one. The present thesis has demonstrated the potential of an ACCDIST-based model for the rarer task of speaker profiling, which aims to gather information regarding a speaker's identity. A possible next step for the Y-ACCDIST system is to explore and test its potential in speaker comparison tasks. Chapter 3, while highlighting the Y-ACCDIST system's capabilities with geographically-proximate accents, simultaneously highlighted sociolinguistic trends expected of the AISEB data. The sensitivity unveiled here may be showing promise for the Y-ACCDIST system's use in speaker comparison tasks. Throughout the forensic speech science literature, we can see individual segmental variables being identified as valuable speaker discriminators. For example, Rose *et al* (2006) show the use of diphthongs to distinguish between individual Australian English speakers. Especially as the speaker comparison task is much more common, it is of interest to trial such as system on these sorts of tasks.

To put the Y-ACCDIST system into practice, the conclusions it produces would need to conform to widespread forensic evidence formats. A position statement (French and Harrison, 2007), signed by most forensic speech scientists practising within the UK, was produced to address concerns arising from the presentation of expert conclusions. Particularly in the statement's *Foreword*, there is a preference towards conclusions expressing 'comparison' outcomes, rather than 'identification' outcomes. It is for the court or jury to decide anything beyond the 'comparative' evidence an expert puts forward. In light of this, the system should be modified in the way it operates its conclusions. The results within this thesis are based on the system conducting an 'identification' task (i.e, assigning speech samples to a discrete category). Further developments are therefore required to reinterpret the classification process upon which the identifications are based to produce valid comparison conclusions. These developments are likely to call on the increasingly used likelihood-ratio framework. The likelihood-ratio framework, having been used for DNA analysis since the 1990s (Morrison, 2012: 3), is a broadly accepted form of evidential presentation across numerous forensic subdisciplines. As a very general description, likelihood-ratios form a numerical probability that a feature of evidence would be found given its typicality within a population.

With the implementation of likelihood-ratios comes the need for population data to refer to, and with the unpredictable nature of forensic speech casework, inevitably comes the need for nonexistent reference databases. One upshot of using computational models, like the one here, is that new databases can be processed relatively quickly on mass. However, the issue of collecting these databases still remains. The problems of using old corpora were reinforced in section 3.4 where only younger speakers were treated as unknown speakers and the older speakers formed the reference population. This data setup aimed to simulate a situation where recently recorded speech samples were compared against a corpus collected in the past. As predicted, recognition rates suffered under this condition.

An ACCDIST modelling technique can also enable us to analyse the extent of variation involved and where it occurs. As seen above, we can conduct a cluster analysis on an entire population and identify varieties which are less likely to be correctly classified. Here, the variety spoken in Gretna shows comparatively little categorical association with regards to geographical origin. Information like this should be accounted for when deriving conclusions.

An obvious direction to move in would be towards foreign accent recognition. To be able to identify a speaker's first language from their production of a second language could also prove useful. Jessen (2007: 187) comments on the increased proportion of forensic casework involving non-native speech. Foreign accent has also been researched with the motivation to improve speech recognition systems. Many studies have largely worked with a two-way classification: native and non-native. Given the data, this may be extended to a more specific classification task to identify the native language of a speaker. As well as the general task of speaker profiling, this may be of particular use to a more specific cause: Language Analysis for the Determination of Origin (LADO). This is a very specifc application relating to asylum seeker cases. When asylum seekers apply for refugee status, government officials need to assess whether the applicant is genuinely from the place claimed in the application. Information regarding the applicant's native language could obviously be a good indicator (or perhaps eliminator) in some cases. Testing the Y-ACCDIST system with the LADO application worked into the experimental design could therefore be worthwhile. Foulkes and Wilson (2011) conducted an experiment on groups of listeners determined by different categories of expertise. Listeners were asked to assess whether speech samples played to them were Ghanaian English speakers or not. In doing this task, they found that LADO professionals in fact performed at approximately chance level, while native speakers performed with the highest success rate. It would be of interest to add an extra automatic 'listener' to such an experiment to assess its performance in comparison. Providing an objective method in the form of a computational tool could support the overall reliability of such assessments.

Another pressing condition for the Y-ACCDIST system to be tested on is with degraded signal. This thesis has tested the system on good-quality recordings, testing the model's performance under a range of different settings concerned with data groupings and system configurations. However, for the purposes of forensic speech analysis, where recordings may have been made under adverse recording conditions, it is imperative that the system is tested on realistic degraded data typical of casework. Also of interest, with regards to degraded recordings, is the ability for the system to perform when faced with data of varying qualities: mismatched recordings. It is preferable to train a system and then use the system on recordings of similar qualities to maximise performance. Obviously, however, the forensic application is unlikely to allow for these ideal conditions. It is therefore important to discover the Y-ACCDIST system's limitations when it comes to degraded speech samples.

Conclusion

Although the primary motivation for the present research was to develop a system for use in forensic casework, the potential purposes revealed above have been two-fold. The second use for the Y-ACCDIST system could be for sociophonetic analysis. A computational tool like the Y-ACCDIST system can allow for a broad and fast analysis of accent corpora, rather than undertaking traditional labour-intensive tasks on smaller pools of data.

Both of the key objectives originally outlined in this thesis have been addressed. The Y-ACCDIST system was initially tested on geographically-proximate accents, a capability thought to be of interest to the forensic speech application. A promising result of 86.7% was generated for the general four-way accent recognition task between speakers from four locations. However, with modifications to do with incorporating the /r/ phoneme, a highest recognition rate of 90.0% was reached. This result is very promising considering the relative physical distances between the speaker locations involved. This finding has also sparked a great interest in identifying the most fruitful combination of vowels.

The second objective of investigating the Y-ACCDIST system's performance on spontaneous (incomparable) speech data has also taken place in the course of this thesis. However, with a highest score of 59.3% on the four-way recognition task, we can conclude that further developments need to be made. However, the fact that this recognition rate rests well above the recognition rate we would expect by chance (25%), we can see that the model works to some extent on incomparable data. With further research and development, it is hoped that advances can be made on 59.3%. As a number of the results above show, it may be a case of filtering out irrelevant segmental components of the speech signal. It is often too much information which is to blame for a reduction in recognition rates.

This thesis has presented an automatic accent recognition system with the potential of assisting with forensic casework. However, a number of forensically relevant directions for further research have been identified. Given the weight of some of the consequences involved in forensic speech casework, it is important that a thorough investigation of these directions takes place.

Notes

[1] http://www.repository.voxforge1.org/downloads/SpeechCorpus/Trunk/Lexicon/ accessed on 14/12/2012

[2] http://svr-www.eng.cam.ac.uk/comp.speech/Section1/Lexical/beep.html accessed on 30/05/2013

Bibliography

Arslan, L. & Hansen, J. (1996), 'Language accent classification in American English', *Speech Communication* **18**, 353–367.

Ash, S. (1988), Speaker identification in sociolinguistics and criminal law, *in* 'Proceedings of the 16th Annual Conference of New Ways of Analyzing Variation', University of Texas, Austin, pp. 25–33.

Bach, J. H., Anemuller, J. & Kollmeier, B. (2011), 'Robust speech detection in real acoustic backgrounds with perceptually motivated features', *Speech Communication* **53**, 690–706.

Barry, W. J., Hoequist, C. E. & Nolan, F. J. (1989), 'An approach to the problem of regional accent in automatic speech recognition', *Computer Speech and Language* **3**, 355–366.

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.

Broeders (2001), Forensic speech and audio analysis, forensic linguistics 1998-

2001 - a review, *in* 'Proceedings of the 13th INTERPOL Forensic Science Symposium', Lyon, France.

Brown, G. & Watt, D. (2014), 'Performance of a novel automatic accent classifier system using geographically-proximate accents', Poster. Presented at the British Association of Academic Phoneticans conference.

Brown, G. & Wormald, J. (2014), 'Speaker profiling: An automatic method?'. Presented at the International Association of Forensic Phonetics and Acoustics Conference.

Chen, N., Tam, S., Shen, W. & Campbell, J. (2014), 'Characterizing phonetic transformations and acoustic differences across English dialects', *Transactions on Audio, Speech, and Language Processing* **22**, 110–124.

D'Arcy, S., Russell, M., Browning, S. & Tomlinson, M. (2004), The accents of the British Isles (ABI), corpus, *in* 'Proceedings of Modelisations pour l'Identification des Langues', Paris, France, pp. 115–119.

DeMarco, A. & Cox, S. (2012), Iterative classification of regional british accents in i-vector space, *in* 'Proceedings of the Symposium on Machine Learning in Speech and Language Processing (SIGML 2012)', Portland, Oregon, pp. 1–4.

DeMarco, A. & Cox, S. (2013), Native accent classification via i-vectors and speaker compensation fusion, *in* 'Proceedings of INTERSPEECH'.

Dror, I., Kassin, S. & Kukucka, J. (2013), 'New application of psychology to law: Improving forensic evidence and expert witness contributions', *Journal of Applied Research in Memory and Cognition* **2**, 78–81.

Ellis, S. (1994), 'The Yorkshire Ripper enquiry: Part 1', *Forensic Linguistics* **1**, 197–206.

Eriksson, A. & Lacerda, F. (2007), 'Charlatanry in forensic speech science: a problem to be taken seriously', *International Journal of Speech, Language and the Law* **14**, 169–193.

Ferragne, E. & Pellegrino, F. (2007), Automatic dialect identification: A study of British English, 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.

Foulkes, P. & French, P. (2001), Forensic phonetics and sociolinguistics, *in* R. Mesthrie, ed., 'Concise Encyclopedia of Sociolinguistics', Pergamon, Amsterdam, pp. 329–332.

Foulkes, P. & French, P. (2012), Forensic speaker comparison: A linguistic-acoustic perspective, The Oxford Handbook of Language and Law, Oxford University Press, pp. 557–572.

Foulkes, P. & Wilson, K. (2011), Language analysis for the determination of origin: An empirical study, *in* 'Proceedings of the Seventeenth International Congress of Phonetic Sciences', Hong Kong, pp. 691–694.

French, P. & Harrison, P. (2006), Investigative and evidential application of forensic speech science, *in* A. Heaton-Armstrong, E. Shepherd, G. Gudjonsson & D. Wolchover, eds, 'Witness Testimony: Psychological, Investigative and Evidential Perspectives', Oxford University Press, Oxford, pp. 247–262.

French, P. & Harrison, P. (2007), 'Position statement concerning use of impressionistic likelihood terms in forensic speaker comparison cases, with a foreword by Peter French and Philip Harrison', *International Journal of Speech, Language and the Law* pp. 137–144.

Hanani, A., Russell, M. & Carey, M. (2011), Computer and human recognition of regional accents of British English, *in* 'Proceedings of INTERSPEECH'.

Hanani, A., Russell, M. & Carey, M. (2013), 'Human and computer recognition of regional accents and ethnice groups from British English speech', *Computer Speech and Language* **27**, 59–74.

Harrison, P. (2004), Variability of formant measurements, Master's thesis, University of York.

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, London.

Hönig, F., Stemmer, G., Hacker, C. & Brugnara, F. (2005), Revising Perceptual Linear Prediction (PLP), *in* 'Proceedings of INTERSPEECH', Lisbon, Portugal, pp. 2997–3000.

Huckvale, M. (2004), ACCDIST: A metric for comparing speakers' accents, *in* 'Proceedings of the International Conference on Spoken Language Processing', Korea, pp. 29–32.

Huckvale, M. (2007a), 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.

Huckvale, M. (2007b), Hierarchical clustering of speakers into accents with the ACCDIST metric, *in* 'Proceedings of the International Congress of Phonetic Science', Saarbrucken, Germany, pp. 1821–1824.

Hughes, A., Trudgill, P. & Watt, D. (2005), *English Accents and Dialects: An Introduction to Social and Regional Varieties of English in the British Isles*, 4th edn, Hodder, London.

Jessen, M. (2007), Speaker classification in forensic phonetics and acoustics, *in* C. Müller, ed., 'Speaker Classification', Vol. 2 of *Lecture Notes in Computer Science*, Springer-Verlag, Berlin Heidelberg, pp. 180–204.

Kassin, S., Dror, I. & Kukucka, J. (2013), 'The forensic confirmation biase: Problems, perspectives, and proposed solutions', *Journal of Applied Research in Memory and Cognition* **2**, 42–52.

Koller, O., Abad, A. & Trancoso, I. (2010), Exploiting variety-dependent phones in portuguese variety identification, *in* 'Proceedings of INTERSPEECH', Makuhari, Chiba, Japan, pp. 748–752.

Köster, O., Kehrein, R., Masthoff, K. & Boubaker, Y. H. (2013), 'The telltale accent: identification of regionally marked speech in German telephone conversations by forensic phoneticians', *The International Journal of Speech, Language and the Law* **19**, 51–71.

Labov, W. (1966), *The Social Stratification in New York City*, Center for Applied Linguistics, Washington D.C.

Ladefoged, P. (2003), *Phonetic Data Analysis: An Introduction to Fieldwork and Instrumental Techniques*, Blackwell, Oxford.

Lawson, E., Scobbie, J. M. & Stuart-Smith, J. (2014), A socio-articulatory study of Scottish rhoticity, *in* R. Lawson, ed., 'Sociolinguistics of Scotland', Palgrave Macmillan, London.

Lincoln, M., Cox, S. & Ringland, S. (1998), A comparison of two unsupervised approaces to accent identification, *in* 'Proceedings of the Fifth International Conference on Spoken Language Processing (ICSLP 98)'.

Llamas, C. (2010), Convergence and divergence across a national border, *in* C. Llamas & D. Watt, eds, 'Language and Identities', Edinburgh Univeristy Press, Edinburgh, pp. 227–236.

McDougall, K. (2004), 'Speaker-specific formant dynamics: An experiment on Australian English /ai/', *The International Journal of Speech, Language and the Law* **11**, 103–130.

Metze, F., Ajmera, J., Englert, R., Bub, U., Burkhardt, F., Stegmann, J., Müller, C., Huber, R., Andrassy, B., Bauer, J. G. & Littel, B. (2007), Comparison of four approaches to age and gender recognition for telephone applications, *in* 'Proceedings of ICASSP', Honolulu, Hawaii, pp. 1089–1092.

Meyer, B. & Kollmeier, B. (2009), Complementarity of MFCC, PLP and Gabor features in the presence of speech-intrinsic variabilities, *in* 'proceedings of INTERSPEECH'.

Morrison, G. S. (2012), 'The likelihood-ratio framework and forensic evidence in court: a response to R v T', *International Journal of Evidence and Proof* **16**, 1 – 29.

Najafian, M., DeMarco, A., Cox, S. & Russell, M. (2014a), Unsupervised model selection for recognition of regional accented speech, *in* 'Proceedings of INTERSPEECH', Singapore.

Najafian, M. & Russell, M. (2014), 'Unsupervised model selection fo recognition of regional accented speech', Poster. Presented at the UK Speech Conference.

Najafian, M., Safavi, S., Hanani, A. & Russell, M. (2014b), Acoustic model selection using limited data for accent robust speech recognition, *in* 'Proceedings of European Signal Processing Conference', Lisbon, Portugal.

Reynolds, D., Quatieri, T. & Dunn, R. (2000), 'Speaker verification using adapted Gaussian Mixture Models', *Digital Signal Processing* **10**, 19–41.

Rhodes, R. (2014), 'Cognitive bias in forensic speech science', Poster. Presented at the International Association of Forensic Phonetics and Acoustics Conference.

Rose, P., Warren, P. & Watson, C. (2006), The intrinsic forensic discriminatory power of diphthongs, *in* 'Proceedings of the Eleventh Australasian International Conference on Speech Science and Technology', Auckland, New Zealand, pp. 64–69.

Shötz, S. & Müller, C. (2007), A study of acoustic correlates of speaker age, *in* C. Müller, ed., 'Speaker Classification', Vol. 2 of *Lecture Notes in Computer Science*, Springer-Verlag, Berlin Heidelberg, pp. 1–9.

Stuart-Smith, J. (1999), Glasgow: accent and voice quality, *in* P. Foulkes & G. Docherty, eds, 'Urban Voices: Accent Studies in the British Isles', Arnold, London, pp. 203–222.

Teixeira, C., Trancoso, I. & Serralheiro, A. (1997), Recognition of non-native accents, *in* 'Proceedings of European Conference on Speech Communication and Technology (Eurospeech)', Rhodes, Greece, pp. 2375–2378.

Torres-Carrasquillo, P., Singer, E., Kohler, M., Greene, R., Reynolds, D. & Deller, J. R. (2002), Approaches to language identification using Gaussian Mixture Models and Shifted Delta Cepstral features, *in* 'Proceedings of the International Conference on Spoken Language Processing', Denver, Colorado, US.

Vieru, B., de Mareuil, P. B. & Adda-Decker, M. (2011), 'Characterisation and identification of non-native French accents', *Speech Communication* **53**, 292–310.

Watt, D., Llamas, C. & Johnson, D. E. (2014), Sociolinguistic variation on the Scottish-English border, *in* R. Lawson, ed., 'Sociolinguistics of Scotland', Palgrave Macmillan, London.

Wells, J. (1982), *Accents of English 2*, Cambridge University Press, Cambridge.

Woehrling, C., de Mareuil, P. B. & Adda-Decker, M. (2009), Linguistically-motivated automatic classification of regional French varieties, *in* 'Proceedings of INTERSPEECH', pp. 2183–2186.

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.

Zeng, Y.-M., Wu, Z.-Y., Falk, T. & Chan, W.-Y. (n.d.), Robust GMM based gender classification using pitch and RASTA-PLP parameters of speech.

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.

Zissman, M. (1996), 'Comparison of four approaches to automatic language identification of telephone speech', *IEEE Transactions on Speech and Audio Processing* **4**, 31–44.