A segmentally-informed solution to automatic accent classification and its advantages to forensic applications

Abstract

Traditionally, work in automatic accent recognition has followed a similar research trajectory to that of language identification, dialect identification and automatic speaker recognition. The same acoustic modelling approaches that have been implemented in speaker recognition (such as GMM-UBM and i-vector-based systems) have also been applied to automatic accent recognition. These approaches form models of speakers' accents by taking acoustic features from right across the speech signal without knowledge of its phonetic content. Particularly for accent recognition however, phonetic information is expected to add substantial value to the task. The current work presents an alternative modelling approach to automatic accent recognition, which forms models of speakers' pronunciation systems using segmental information. This paper claims that such an approach to the problem makes for a more explainable method, and therefore a more appropriate method to deploy in certain settings where it is important to be able to communicate methods, such as forensic applications. We discuss the issue of explainability and show how the system operates on a large 700-speaker dataset of non-native English conversational telephone recordings.

Keywords: automatic accent recognition, explainable technologies, segmental features, forensic applications

1. Introduction

The field of automatic speaker recognition has been subject to an evolution of modelling approaches, and this has driven many of the approaches applied in Language Identification (LID) and Dialect Identification (DID). There has been a transition from Gaussian Mixture Models (Reynolds and Rose, 1995) to i-vectors (Dehak et. al., 2011) and now these fields have progressed to Deep Neural Network embeddings (Snyder et. al., 2017). Throughout this evolution, we have witnessed automatic speaker recognition error rates edge closer and closer to 0% under certain conditions. We can view automatic accent recognition as a task that is similar to LID and DID in that it is usually about assigning a class label to a speaker. However, different accents of the same language are expected to have fewer and more subtle differences discriminating them than different languages or different dialects. Within DID, we can see an Equal Error Rate (EER) as low as 6% in the work of Biadsy et. al. (2010) who aimed to automatically classify speakers into one of five broad dialect groups of Arabic[1]. In comparison, automatic accent recognition error rates are often in the region of 15-25% EER (e.g. Behravan et. al., 2013). This trend in EERs aligns with initial expectations around the relative difficulty of dialect classification tasks and accent classification tasks. Because dialects tend to vary in relation to a number of levels of linguistic analysis (e.g. lexical variation, pronunciation variation, prosodic variation), and accent relates predominantly to pronunciation variation, it is logical to expect fewer and more subtle differences among different accents. Accent also varies as a result of many factors beyond the regional background of the speaker (e.g. i.e. other social or experiential factors). It is therefore reasonable to expect accent recognition to be a more difficult task.

[1] However, we note that Boril et. al. (2012) pointed out that such a low error rate in Biadsy et. al. (2010) could have been partly down to channel characteristics that separated the classes of speakers, not just linguistic differences.

Human accent recognition performance tends to confirm the difficulty of identifying a speaker's accent. In the linguistic perception literature, Clopper and Pisoni (2004) report an overall classification rate of 30% correct on a task that involved 23 human listeners classifying speech samples into one of six North American English dialect categories. Likewise, Vieru et. al. (2011) ran experiments on human listeners to test how well they could classify non-native accented French speech samples into one of six categories. They observed an overall classification rate of 52% correct. In both studies, the chance level of performance was 16.67% correct.

The work presented in this paper builds on the existing automatic accent recognition research by demonstrating the capabilities of a segmentally-informed approach, and by justifying why this is a good option for forensic applications. A segmentally-informed approach is one where there is knowledge about the segmental content (i.e. the sequence of phone-sized units) incorporated into the system's workings and is therefore enabled to directly target the phonetic productions that are expected to be diagnostic of a speaker's accent. The specific segmentally-informed system used in this study is the York ACCDIST-based automatic accent recognition system (Y-ACCDIST) (Brown, 2016). As the prospect of using automatic speaker recognition technology for forensic casework increases, Brown (2017) entertains the idea of developing an automatic accent recognition system for forensic applications in cases where we might want to establish an accent profile of an unknown speaker. Conducting a speaker profiling task like this could assist in narrowing down the pool of potential suspects in an investigation (Watt, 2010), or it could even assist in refining a reference population for speaker comparison.

Across the forensic sciences, there are movements towards introducing technology to assist with analyses. The ability to explain how an analysis has been carried out is particularly important if the analysis feeds into evidence used in court. However, even if the analysis is not used evidentially, there are benefits to being able to explain these analysis methods. Analysts (or, indeed, other stakeholders) are likely to use methods more effectively and appropriately if they have a stronger understanding of them. Because it is segmentally-informed, rather than heavily reliant on machine learning algorithms, Y-ACCDIST is more transparent and "explainable", and these are desirable properties of methodologies and analyses that are presented to the courts. We expand on the idea of explainable methodologies in the context of forensic applications in Section 2 below, before describing the operating principles behind Y-ACCDIST. Section 3 and Section 4 contain details on the experiments we have run and the results. Section 5 discusses the results in the context of system explainability and the system's place among other automatic accent recognition systems.

2. Background

This section first discusses the increasing need for explainable solutions in the context of forensic science. It then conceptually introduces the Y-ACCDIST system and how it is different from other automatic accent recognition systems in Section 2.2. Section 2.3 then considers the nature of the data used for these experiments.

2.1 Explainable Solutions

While this paper specifically considers the importance of explainability in forensic contexts, explainability is very much a topic of interest across various disciplines that fall within artificial intelligence. There is an emerging interest in, what is called, *Explainable AI* (regularly referred to as XAI) (Adadi and Berrada, 2018). Explainable AI is concerned with improving our ability to explain the workings of different technologies that make use of increasingly complex techniques in order for us to truly trust these systems. Often, medical applications of AI tools are given as examples to illustrate the importance of explaining the inner mechanisms of these systems (e.g. Adadi and Berrada (2018) and Samek et. al. (2017)). In the context of disease diagnosis, the decision being made by a system potentially has significant consequences on people's lives, and so having a clear understanding of the method used to reach that decision can be crucial.

The work in this article targets forensic applications which is another application with significant consequences attached. The output of a forensic analysis is fed into the justice system and is possibly used to motivate life-changing decisions. In a similar way to medical purposes, it is important that we can justify the analyses used for such purposes. This is incorporated into the *Codes of Practice and Conduct* put forward by the UK Forensic Science Regulator that are expected to be followed by forensic practitioners operating in a range of forensic analytical disciplines (Tully, 2020). Within the *Codes of Practice and Conduct*, it is stated in Section 28.4.2 on the reporting competencies of forensic experts) that:

Forensic units shall ensure that all staff who provide expert evidence... are able to explain their methodology and reasoning, both in writing and orally, concisely in a way that is comprehensible to a lay person and not misleading.

This part of the *Codes of Practice and Conduct* is very much in line with a lesser-known part of the European Union's General Data Protection Regulation (GDPR) which generated a lot of public and media attention in 2018 when it was implemented. Much of the attention surrounding GDPR concerned gaining consent from individuals when organisations handled personal data. However, within the GDPR, there is also mention of the right to an explanation. This has machine learning in mind where a lot of data is typically used to yield outputs. The right to an explanation ties in with the GDPR's key aim of giving people more control of their personal data, and it is of particular importance when the stakes are high in the decisions that are made by an algorithm.

If evidence in a case is not sufficiently explained, it is quite possible for the court to discard or ignore that piece of evidence in a case. One such example was demonstrated in *R. v Clark* (EWCA Crim 1020., April, 2003). In this case, a mother was appealing her convictions for murdering her two baby sons. A number of medical experts became involved in this case, but among these the evidence given by a Home Office pathologist was called into question. The ruling commented that the pathologist appeared to be unable to communicate and explain how he arrived at the conclusion he did. Based on this inability to explain, the ruling went on to say that "this aspect of the matter called into question the competence of Dr [X]".

It would be too idealistic to expect that all methods used for forensic analysis can be explained to a point that a lay person is well versed in all the details. However, it is clear that a more explainable method, if available and fit-for-purpose, would be preferred. Alongside the case outlined above, the issue of explainability has also been raised in a case that involved forensic speech analysis. In the case of Slade and Ors ([2015] EWCA Crim 71), an automatic speaker recognition system was applied to the voice recordings in question, but this evidence was not admitted in this case. This was partly because the court was concerned about the possibility of effectively communicating the method to a jury.

In recognition of the need for clear and coherent explanations of forensic methods, the Royal Society has been involved in a project that aims to improve ways of explaining methodologies in forensic settings. Specifically, a steering group was formed to develop “primers”, which are short documents that are designed to clearly explain how a specific forensic analytical method operates, as well as laying down the limitations attached to that method. The project has been a collaborative effort between forensic scientists, judges and other legal practitioners to carefully craft the content of these primers. It is hoped that presenting primers to judges in a given case can assist in improving understanding of a specific type of evidence and therefore delivering decisions in a more informed way. So far, these primers have been developed for DNA analysis and gait analysis (i.e. analysing the way in which individuals walk), and there are intentions to develop more primers for other forensic disciplines.

Within the phonetics literature, there have been attempts to increase the explainability of machine learning techniques when they are applied to voice data. Ferragne et. al. (2019) carried out a very focussed study where they extracted tokens of a single French vowel phoneme from 45 speakers and passed the spectrograms of these vowels through Convolutional Neural Networks (CNN) in speaker classification trials. In different sets of trials, the authors manipulated and varied which frequencies were included in the training and test data that were used by the CNN to attribute the vowel spectrograms to individual speakers. By manipulating the data in this way, the authors aimed to uncover which features the CNNs were learning, and which frequencies were important, to perform the speaker classification task. Manipulating the data like this presents one way in which we can learn more about the inner workings of machine learning techniques.

As part of the effort to achieve a firmer grasp of the underlying mechanisms of a method, we can also look towards developing alternative methods that are inherently easier to explain and to understand. The Y-ACCDIST system that is presented in this paper is less dependent on machine learning mechanisms that do not make it easy to recover the features it based its overall decisions on. The Y-ACCDIST system is restricted to focussing on how the different vowels and consonant phonemes are phonetically realised by individual speakers, which is expected to reveal the accent of a given speaker. By focussing the system in this way, we can be more sure that the features key to the outcome of an accent recognition decision are taken into account by the system. Increasing certainty like this is a way that explainability can be integrated into an automatic accent recognition system. Other types of system rely much more heavily on machine learning techniques to estimate which combination of acoustic features is likely to lead to a correct outcome. It is then difficult to determine whether this latter method has definitely taken into account those features which are likely to contribute most to a successful accent recognition task, and therefore more difficult to explain to the relevant parties involved in the legal domain how the method has delivered its outcome. The Y-ACCDIST system is a segmentally-informed alternative, and a method where we can be more sure that it is the phonetic realisations of vowels and consonants that count towards its decision.

2.2 The Y-ACCDIST System

The York ACCDIST-based automatic accent recognition system (Y-ACCDIST) is inspired by the ACCDIST metric that was devised by Huckvale (2004). Unlike other accent recognition system architectures, ACCDIST-based approaches explicitly implement segmental information in the modelling process to take advantage of differences we primarily associate with accents in the linguistic literature. Specifically, these are phonetic realisational differences, such as the pronunciation of the BATH vowel that distinguishes typical speakers from the North and South of England. Northern speakers would typically produce an [æ] while Southern speakers typically produce [ɑ]. ACCDIST-based systems are therefore quite selective in the information they include. However, we acknowledge that they ignore other potentially informative layers of analysis that could be indicative of a speaker’s accent. For example, we know that there are prosodic cues that are characteristic of specific varieties (Grabe, 2004). Additionally, voice quality parameters have been studied as characteristic features of accents. Voice quality is concerned with the underlying settings of the vocal tract which determine longer-term characteristics of a speaker’s voice (for example, a speaker might sound particularly nasal throughout his or her speech productions because of their tendency to position the velum in a certain way). Voice quality is usually associated with individual speaker differences because it is closely linked with a speaker’s physiology (and therefore perhaps a more appropriate level of analysis for speaker recognition). There are also certain behavioural aspects of voice quality which lead to speakers acquiring certain vocal settings. This results in speakers within a speech community sharing voice quality features. As one case study, Stuart-Smith (1999) found that working-class Glasgow speakers typically have a more whispery voice quality than middle-class Glasgow speakers. It is important to acknowledge that these other potentially relevant parameters of speech are overlooked by the approach we take in this work. These parameters are perhaps captured by other types of accent recognition system like GMM-UBM (e.g. Chen et. al. (2001)), i-vector-based systems (e.g. Najafian et. al. (2016)) and neural network-based systems (e.g. Shon et. al. (2018)) because they are expected to capture a more global acoustic representation of speech samples. Such types of acoustic system have also been compared with ACCDIST-based approaches on the same datasets in past studies (Brown, 2016; Hanani et. al., 2013) confirming our expectations that the text-dependent ACCDIST-based systems outperform the text-independent acoustic-based systems in each case. This suggests that a more targeted approach that is based on the features, from the outset, that are expected to differentiate among classes (in this case, accent groups) can also improve overall performance.

What separates the Y-ACCDIST system from other ACCDIST-based systems seen in Huckvale (2004, 2007) and Hanani et. al. (2013) is that it has been adapted to be able to process spontaneous content-mismatched speech (Brown 2015, 2016). Specific details of how Y-ACCDIST does this are provided further below in the technical description of the system in Section 3.3. Other studies that have tested ACCDIST-based systems have relied on highly controlled data where the spoken content of the training data has matched that of the test data (the same reading passage or read prompts). Much of the previous work on the Y-ACCDIST system has been motivated by forensic applications, and so removing this restriction was a central focus in its development.

It is possible to be selective and vary the phone segments that are included in the Y-ACCDIST models. Other ACCDIST-based studies (Huckvale, 2004; Hanani et. al., 2013) only included vowel-based segments. In the case of Huckvale (2004, 2007), word-specific vowels were used, meaning that the vowel in *kit* and the vowel in *trip* were treated as separate segmental variables (or features) in the models. In the case of Hanani et. al. (2013), triphone-specific vowels were used as segments for modelling. This meant that the vowel in *kit* and *trip* were, again, treated as separate vowel segments, but the vowels in *trip* and *rip* were collapsed into the same segmental variable. In these previous studies, it is likely that selecting only vowel-based units was a way to reduce the computational cost, because the types of segments they used resulted in making a lot of feature types available to use in their models. The use of highly context-specific segments restricts these systems to being implemented in tasks that only involve reading passage data, where the spoken content of the training speakers matches that of the test speakers. In contrast, the segments used for modelling in Brown (2015) and Brown (2016) were not context-specific and, instead, phoneme-based segments were implemented in the modelling. As a consequence, the vowels in *kit*, *trip* and *rip* were all collapsed into one segmental variable in the modelling, making it possible to apply the system to spontaneous content-mismatched speech data. We took advantage of the fact that there are therefore fewer segments available for modelling – this makes it possible to show the benefits of including consonant segments as well. Although the sociophonetics research literature often focusses on differences in vowel productions between different accent varieties, of course consonant productions can also be used as diagnostics in accent classification. By only focussing on vowel segments, we would possibly ignore many useful cues in an accent recognition task. In previous testing of the Y-ACCDIST system, on two different corpora of native English accents, it was found that in one case a vowels-only setting outperformed an all-phonemes setting, whereas the reverse outcome came about for the other corpus (Brown, 2017). We can gather that whether a vowels-only or an all-phonemes setting is the more successful one depends on the specific set of accents that are being distinguished between[2].

The data in the present work provide us with an additional opportunity to include filled pauses (“er” and “erm”) as unique segments within the accent models. Filled pauses have accumulated some attention in the forensic phonetics literature, specifically in the context of forensic speaker comparison (Hughes et. al., 2016; McDougall and Duckworth, 2017). We also found in Franco-Pedroso and González-Rodríguez (2016) that filled pauses were among the most discriminative units for automatic speaker recognition. Because of their frequency in spontaneous speech, we would assume that they form a strong representation in the speaker models. The fact that we can think of them as unconscious events suggests that they are more difficult for speakers to disguise and they may also transfer more straightforwardly from a speaker’s first language. Consequently, they are expected to be strong variables to include in a speaker model. For these same reasons, we could view filled pauses as variables worth pursuing for accent recognition.

Until the present study, another impracticality of Y-ACCDIST has remained, making it less appealing as a tool for real-life applications. Even though it has been demonstrated that Y-ACCDIST can process content-mismatched (spontaneous) speech data, it still requires a transcription as input. So far, these transcriptions have been manually generated, but the experiments in this paper make use of estimated transcriptions based on the output of a speech recognition system. Incorporating this step still comes with its pitfalls, but it removes a large amount of labour from the overall process.

2.3 Non-native accents

All past work on ACCDIST-based approaches to accent recognition have been concerned with native varieties of English. The work of Huckvale (2004, 2007), Hanani et. al. (2013) and Ferragne and Pellegrino (2010) made use of the Accents of the British Isles (ABI) corpus (D’Arcy et. al., 2004). The ABI corpus enabled researchers to classify speech samples into one of 14 accent categories that were dispersed across the breadth of the British Isles. Work with the Y-ACCDIST system has homed in on accent varieties that are expected to be much

[2] There is also some intervention from the SVM, in that during training it estimates weights for the different input features that are included, and so some features (or segments) will contribute more to the classification decision than others.

more similar to one another, in that they are geographically much closer together than the accents offered by the ABI corpus (Brown 2016, 2017).

The work presented here moves away from native accents and towards categorising non-native accent data. This work makes classifications of the speakers' native language based on their production of English. The non-native task introduces new challenges and factors to consider. We can expect that the data used to train a model for a single accent is more variable for a non-native accent than it is for a native accent. Piske et. al. (2001) offer a list of factors that affect a speaker's foreign accent. Among these are factors like the age of the speakers when they learn a language, the length of time speakers have resided in a place where the second language is spoken, the motivation of the speaker and the aptitude of a speaker. There are also aspects that link to a speaker's identity. Drummond (2012) points out that some second-language speakers may intentionally avoid adopting pronunciation features of native speakers to reinforce their native identity.

Linguistic proficiency is not relevant to native accents, and so introduces a different kind or extent of variability to the data in this paper's experiments. Behravan et. al. (2015) looked at the effect of language proficiency on i-vector-based automatic accent recognition performance. Using a database of second-language speakers of Finnish, they found that more proficient speakers are less likely to be correctly labelled in an accent recognition task. We might draw from this that a higher level of proficiency removes features that are indicative of a speaker's native language. This is a factor that will no doubt have an effect on the experiments and results presented here. The chosen dataset is one that consists of hundreds of speakers for whom English is their second language. There will be a great range of proficiencies among the speakers and perhaps an imbalance of proficiencies across the different accent groups. Metadata is not available on the proficiencies of these different speakers. This is therefore an aspect to keep in mind when reviewing the results, rather than a factor that will be analysed in detail.

3. Experiments

This section first introduces the dataset that was used to train and test the Y-ACCDIST system, before describing the system's inner workings. This section also provides explanations of the different ways we can observe the system's outputs and performance evaluation.

3.1 NIST SRE data

The data used to test the performance of Y-ACCDIST on non-native varieties are from the *National Institute of Standards and Technology Speaker Recognition Evaluation* (NIST SRE) 2004, 2005 and 2006 datasets (Przybocki et. al., 2004). These datasets are primarily intended for automatic speaker recognition experiments, but metadata are available that make it possible to conduct other kinds of task. These databases are largely made up of telephonic speech, and the subset used for the experiments in this paper is just made up of telephonic speech (as this is a data condition that is reflective of a large proportion of forensic speech casework). Not all speakers in the combined dataset are relevant to the research objectives of this paper. Not all the recorded conversations in the database are in English, and these recordings are not relevant to these experiments. Additionally, of the recordings where speakers are speaking English, there are great imbalances between the number of speakers within different accent groups. By far, the group with the largest number of speakers is made up of native English speakers, and there are large numbers of native speakers of Russian who are speaking in English. At the other end of the spectrum, there are very few Vietnamese speakers, for example. From the different speaker groups available, seven were selected, based on the fact that these categories allowed for accent groups of 100 speakers each. The resulting non-native accents for these classification experiments were therefore:

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

More detailed information is not available on the speakers' language background. For instance, the Arabic speakers could be native speakers of any Arabic dialect, but this information is not documented in the metadata. Also, having listened to a small number of the recordings in the native English category, it seems that most of the speakers are speaking a North American variety of English (as expected), but there are also recordings we strongly suspect are of British English speakers. These factors may bring about additional variability in the accent models, on top of those that come with non-native accents.

Each speaker's speech sample is part of a two-way casual conversation with another speaker which lasts for approximately 5 minutes in total. We can therefore expect that each individual speaker's speech sample contains around 2-2.5 minutes of speech (although we should acknowledge that there is likely to be variation in duration per speaker). Having listened to a small number of these recordings, it seems that it is not uncommon for recordings in this database to come with some background noise (e.g. other people talking in the background).

3.2 Phone estimation

As stated in Section 2.2 above, Y-ACCDIST requires a transcription of the recordings for the modelling process. In past work, manually prepared transcriptions have been used and the initial stage of the system has been forced aligned. In this work, the outputs of an automatic speech recognition are used instead as the dataset of spontaneous speech is much larger than those used previously. The resulting estimated labels are those that were used for the experiments in Franco-Pedroso and González-Rodríguez (2016). These labels were produced by the *Decipher* automatic speech recognition system by researchers at SRI (Kajarekar et. al., 2009)[3]. This system achieves a Word Error Rate of 23.0% for native English speech and 36.1% for non-native English speech. These error rates suggest that the resulting transcriptions we are using in these experiments are riddled with errors. We can therefore expect numerous errors that feed into the Y-ACCDIST models.

Another detail to point out is that the speech recognition system is trained on North American English. This means that the estimated transcriptions make use of North American phonology. For the Y-ACCDIST system's processes, we need a single reference phonology for the differences in phonetic realisations between the accent varieties to be expressed.

[3] These transcriptions were not generated specifically for the purpose of these experiments, but had been used for past experiments for applications such as speaker recognition. Using the ASR system transcriptions was a practical decision based on the size of the dataset. Other Y-ACCDIST studies have not used automatic methods to produce the transcriptions.

However, we can see that some of the non-native speakers may be targeting a British phonology (depending on the speakers' previous linguistic exposure or education).

3.3 Y-ACCDIST system description

Using the phone labels and estimated segmentations produced by the speech recognition system, we aim to form a model of accent for each speaker. Taking each training speaker's segmented speech sample, a single MFCC vector is extracted at the midpoint of each estimated phone segment. The MFCCs that are extracted consist of 12 cepstral coefficients, plus energy, amounting to a vector of 13 elements. The MFCCs had a standard frame duration of 25ms. From these features, the average MFCC vector is calculated for each phoneme in the phoneme inventory, gathering a sense of a speaker's production of different segments. Using these representations, it is then possible to form a model of each speaker's accent by compiling a matrix for a single speaker with the cosine distances of all the phoneme pair distances that are possible within the phoneme inventory, using the average midpoint MFCC representations[4]. Figure 1 demonstrates part of a Y-ACCDIST matrix with just three phonemes.

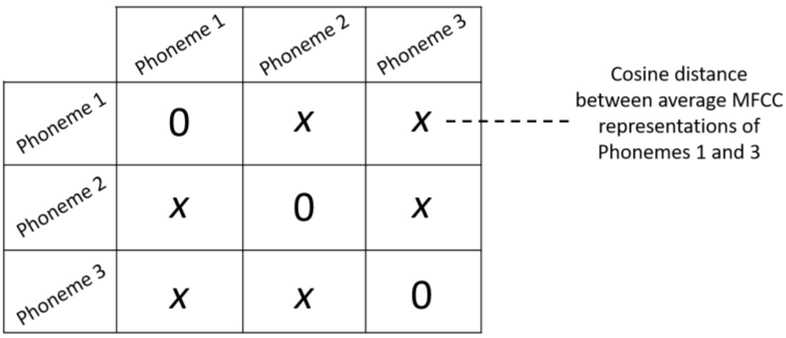

Figure 1: Illustration of part of a Y-ACCDIST matrix with only the first three phonemes of the inventory.

Forming a model of accent at this stage of the process is key to opening up the explainability of the system. The particular modelling method draws from our understanding of accent variation to access the accent-specific information and organises the acoustic features before passing them through machine learning classification mechanism. This is instead of using a machine learning or statistical method to estimate the information or combination of features that is useful to the accent classification task without drawing on existing understanding of how accents vary.

Taking the speaker for each accent group in our training set, we can then train a Support Vector Machine (SVM) classifier using the Y-ACCDIST matrices as input features. In this classifier, each speaker is effectively plotted in multidimensional space, positioned by the Y-

[4] In past work using the Y-ACCDIST system, Euclidean distance has been used. In the case of the specific dataset in this work, we saw that cosine distance brings about the optimal performance. We will note this performance difference between the distance metrics in Section 4.1 and further discuss the possible reasons for the difference in Section 5.

ACCDIST matrix elements. For each accent category in our dataset, we can form a *one-against-the-rest* configuration, where, on rotation, each accent is the ‘one’, while the speaker matrices from the other accent categories become one collective group forming ‘the rest’. A hyperplane and optimal margin that best separate these two groups of speakers is computed with a linear kernel. To classify an unknown speaker’s accent, we model their speech sample as a Y-ACCDIST matrix. On each rotation for each accent category in the SVM, the unknown speaker’s matrix is fed into the SVM. The clearest margin it forms with the hyperplane on one of these rotations indicates its accent category. Figure 2 below illustrates the whole series of processes involved in the variant of the Y-ACCDIST system used in this study.

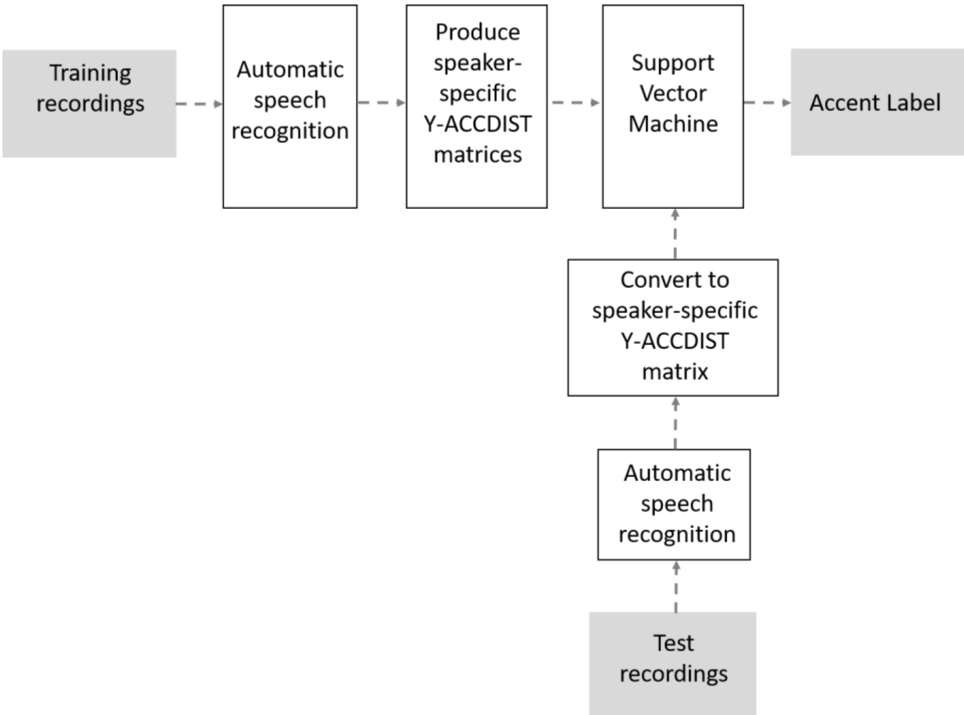

Figure 2: System diagram of the Y-ACCDIST-SVM system used in these experiments.

3.4 Performance Evaluation: accent recognition as classification and verification problems

This paper presents results in two different forms. One form is a ‘hard’ decision for each accent recognition trial. This looks at accent recognition as a classification problem. Taking the best-performing segmental configuration from these results (i.e. vowels-only, all-segments, with/without filled pauses), this paper then moves on to analysing the performance through observing the accent trials as ‘soft’ decisions (i.e. deriving likelihoods of speech samples assuming an accent category or not that accent category). For the soft decisions, we use the scores produced by the SVM, not just the label that led to the clearest margin. These scores are obtained by using a test speech sample’s distance from the hyperplane in the SVM. These soft decision outputs allow us to consider accent recognition as a verification problem. It is important to consider these likelihoods in the context of applications for accent recognition technology. It is plausible to have speech samples that are computed to form fairly convincing margins for all the candidate accents in a system, and likewise, we could have speech samples that are not assessed to be particularly similar to any of the candidate accents. A hard decision would simply assign a label which would carry equal weight in each case. Looking at soft decisions gives us some gradience to an outcome. In the context of forensic applications, a soft decision allows us to take the weight of the evidence into account.

To investigate the use of soft decisions, likelihood ratio framework is implemented. We take inspiration from how it is implemented in speaker recognition and apply a version of this to accent recognition. The likelihood ratio can be expressed through the simple formula below:

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

The numerator is the likelihood of the evidence (the speech recording from unknown source) when assuming that the prosecution hypothesis, H_p, is true. In the context of speaker verification, this hypothesis is represented by a model of the suspect's voice, while in the case of accent recognition it will be a given target accent. On the other hand, the denominator is likelihood of the evidence when assuming that the defence hypothesis, H_d, is true. That is, the speaker in the recording is a speaker other than the suspect, or in the accent recognition scenario, the speech sample belongs to a different accent than the given one. In effect, we have a ratio that puts the degree of similarity between the unknown and target samples (suspect or given accent category) against the estimated degree of typicality that the evidence presents.

3.5 Calibration

Calibration is a means of measuring reliability or "confidence" of a system's outputs (González-Rodríguez et. al., 2007). Rather than simply outputting a likelihood ratio from a system and assessing whether the value outputs a probability that is in favour of the true label (based on a test set of data), we can gather a much more proportionate indication of how accurate outputs are. For example, if a system outputs a score that is in strong support of a given hypothesis and it turns out that the true value lies with the alternative hypothesis, it is important to know the extent to which a system does this. In these kinds of instances, we would want to 'penalise' the system heavily and for this to be reflected in an overall measure of the system's reliability. Calibration can help us to capture this kind of information (see Ramos-Castro et. al. (2006)).

It has been suggested that in a classification task, rather than a recognition task like speaker recognition (a sort of binary setup), we might want to conduct calibration to account for multiple hypotheses (i.e. the likelihoods of a speech sample belonging to each of the classes in our set). Brümmer and Van Leeuwen (2006) offer solutions to the automatic Language Identification research community, enabling researchers to calibrate scores for a classification task like this, and indeed, Bahari et. al. (2013) implement such a method on an accent classification task similar to the one presented in this paper. It is proposed here that we continue to use the binary setup that is associated with speaker recognition problems. This is because we perceive this configuration as more applicable to forensic applications, the original intended application for the Y-ACCDIST system, and allows us to explore the more "open-set" type of question, rather than just a "closed-set" question.

The experiments presented in this paper make use of pool adjacent violators (PAV) calibration (as outlined by Brümmer (2007))[5]. Other calibration methods exist. A more common option would be to use logistic regression (e.g. Morrison (2013)), but the distribution of scores from the Y-ACCDIST-SVM system is more suited to a non-parametric method of calibration. As a result, PAV calibration has been selected to produce the likelihood ratios in these experiments. PAV calibration essentially "bins" the scores into ordered score categories, allowing us to work with proportions within a set of probabilities, rather than the raw values that do not necessarily reflect a linear scale. To conduct calibration, an additional partition of data is needed to develop the system in this way. This is described further below.

To observe system performance using hard decisions, the simple measure of % correct is used in the current work. To observe system performance using soft decisions, measures that are standardly used in speaker recognition research are used: Equal Error Rate (EER) and Log-likelihood-ratio Cost Function (Cllr).

3.6 Experimental Setup

A *leave-one-out* cross-validation configuration has been implemented for these initial experiments, where one speaker is removed from the dataset and reserved as a test speaker, while the rest of the speakers are used to train the system. This setup allows us to maximise the amount of data available to train the system. First, we will present different segmental settings, looking at different combinations of vowels-only and all-phonemes settings with and without filled pause segments.

4. Results and Analysis

The results are presented in two parts to first reflect the closed-set type of question (hard decision) and then the open-set type of question (soft decision).

4.1 Accent recognition as a classification problem

The following results demonstrated Y-ACCDIST's performance on the seven-way task in terms of a 'forced-choice' classification through % correct as the performance measure. Results are presented where different segmental settings have been in place. In previous ACCDIST-based research, vowel segments have often been favoured, while some other work has also included consonants (as discussed in Section 2). Table 1 presents results where only vowel segments have been used to model accents, as well as results where the whole phoneme inventory has been used. We also take advantage of the automatic speech recognition system's outputs and present results that include filled pauses as a separate segmental variable. Tables 2-5 show the accompanying confusion matrices for each segmental setting in turn.

Segmental setting	% Correct
Vowels-only	45.0
All-phonemes	60.4
Vowels-only plus filled pauses	48.1
All-phonemes plus filled pauses	62.6

Table 1: Y-ACCDIST classification rates on the seven-way non-native accent recognition task using different segmental settings (chance rate = 14.3% correct).

[5] PAV calibration in these experiments were implemented through a MATLAB script written by Daniel Ramos-Castro.

Accent	Ara.	Ben.	Chn.	Eng.	Far.	Spa.	Rus.
Ara.	**40**	11	6	9	6	17	11
Ben.	5	**61**	10	3	10	5	6
Chn.	6	14	**41**	7	12	12	8
Eng.	8	1	4	**53**	10	14	10
Far.	9	16	11	11	**37**	9	7
Spa.	12	9	7	11	9	**45**	17
Rus.	13	4	15	13	5	12	**38**

Table 2: Confusion matrix of the Y-ACCDIST non-native classification experiment using the vowels-only setting.

Accent	Ara.	Ben.	Chn.	Eng.	Far.	Spa.	Rus.
Ara.	**58**	7	9	1	8	5	12
Ben.	7	**73**	4	1	8	4	3
Chn.	5	3	**60**	8	6	10	8
Eng.	7	3	8	**54**	7	12	9
Far.	5	4	4	11	**62**	6	8
Spa.	5	4	6	14	4	**60**	7
Rus.	8	1	5	15	6	9	**56**

Table 3: Confusion matrix of the Y-ACCDIST non-native classification experiment using the all-phonemes setting.

Accent	Ara.	Ben.	Chn.	Eng.	Far.	Spa.	Rus.
Ara.	**46**	5	9	3	9	13	15
Ben.	4	**72**	7	2	7	1	7
Chn.	6	10	**42**	7	16	7	12
Eng.	5	3	6	**49**	13	11	13
Far.	10	15	13	13	**35**	12	2
Spa.	6	5	6	15	8	**54**	6
Rus.	13	5	13	17	5	8	**39**

Table 4: Confusion matrix of the Y-ACCDIST non-native classification experiment using the vowels-only plus filled pauses setting.

Accent	Ara.	Ben.	Chn.	Eng.	Far.	Spa.	Rus.
Ara.	**60**	8	6	3	9	2	12
Ben.	7	**76**	5	1	6	3	2
Chn.	5	5	**62**	6	3	11	8
Eng.	7	3	7	**57**	6	12	8
Far.	5	6	4	11	**63**	5	6
Spa.	9	3	4	10	7	**59**	8
Rus.	5	3	5	10	7	8	**61**

Table 5: Confusion matrix of the Y-ACCDIST non-native classification experiment using the all-phonemes plus filled pauses setting.

The results above uncover two main findings and some initial interesting observations with regards to the contents of the confusion matrices. Looking at such details can promote the system's explainability.

The first main finding is that the all-phonemes setting outperforms the vowels-only setting. In past works that have tested variants of the Y-ACCDIST system, this has not always been the case (Brown, 2017). When testing it on a smaller corpus of geographically-proximal accents (where there is an expected increase in similarity between the varieties in the dataset), the vowels-only setting yielded a higher recognition rate than the all-phonemes setting. When applying the same segmental settings to another corpus where the accent varieties are expected to be more different from one another, the all-phonemes setting outperforms the vowels-only setting (i.e. a larger Y-ACCDIST matrix seems to be more effective than a smaller Y-ACCDIST matrix in this particular task). In an accent recognition task that involves more similar accents, we can expect that fewer segments will contribute to the task, and so the vowels-only setting might exclude more features that do not help with accent discrimination. In the case of the NIST dataset used in the experiments in this paper, it seems that there are a larger number of segments that help to distinguish between these particular varieties. There are also issues to consider in relation to how the inclusion of consonants may have an impact on the explainability of the system. For example, we do not necessarily know what the distance between /t/ and /æ/ might contribute, unlike the distance between /ɑ/ and /æ/. Such points are considered further below within a wider discussion of the Y-ACCDIST system's explainability.

The second key finding is that including the filled pause segments in the accent models seems to yield a slight improvement on performance. The overall best performance is a combination of all phonemes, plus filled pauses with a result of 62.6% correct[6]. As already pointed out above, among the forensic phonetics research there is some interest in establishing whether filled pauses carry acoustic information that can discriminate speakers. There has also been some work that looks at how filled pause production can be characteristic of whole languages (de Leeuw, 2007). Even within bilingual speakers, Lo (2020) found that there were distinctive language-specific pausing behaviour between two different languages. Given that

[6] Recall from Section 3.3 that different distance metrics were trialled in the modelling of speakers' accents. Under this same segmental configuration, but replacing cosine distance with Euclidean distance, the system achieves 54.4% correct. This could be because of the cosine distance's apparent suitability to imbalanced data samples (in this case, spontaneous speech samples), compared with Euclidean distance. This would explain why the Euclidean distance has been a better metric on more controlled data (read speech samples) in past experiments.

the classification task in this paper is to identify the first language of the speakers, it is unsurprising to find that filled pauses seem to make some contribution towards the increased success rate of this system.

It is possible to make some remarks in relation to the classifications and misclassifications of the specific accent varieties included in these experiments to speculate about whether the system is performing in an expected way. Taking Table 5, which is the confusion matrix attached to the highest-performing system configuration, it is possible to make some tentative observations with regards to some of the higher rates of misclassification. Among English, Spanish and Russian, it seems that there is some degree of consistency in the higher number of misclassifications among these non-native accents. For example, English speakers are misclassified as speakers with Spanish as a first language on 12 occasions, while Spanish speakers are misclassified as native English speakers on 10 occasions. Similarly high misclassifications (in both directions) occur among English, Spanish and Russian in all permutations. Although, with these numbers, it is right to be cautious, the level of confusion among these three particular accent groups could reflect their relative similarity among this particular accent dataset. English, Russian and Spanish fall towards the European side of the Indo-European language family. While this does not necessarily point towards greater linguistic similarity in their modern forms, it may have some influence on their overall similarity as a group of languages, and therefore an increase in their overall similarity in the spoken production of English.

Although potentially interesting patterning has emerged in the confusion matrix, it is right to be cautious in these particular experiments. As automatically-derived transcriptions have been used, it is quite possible that some of the accents are more susceptible to speech recognition errors than others. Such a trend in accent-specific speech recognition errors is well-documented (e.g. Vergyri et. al., 2010). These errors could then contribute to overall differences in performance between the different accent categories. With the information and data available, it is not possible to derive the exact effect this has on the confusion matrix above, but it is certainly worth keeping in mind during the interpretation of these results.

It is possible to take a closer look at individual test speakers to shed light on the detail of what information the system may be using to carry out its classifications. Doing so can enable us to go further to address this paper's key theme of system explainability, reinforcing the advantages of using a Y-ACCDIST-based modelling method over others. In a similar way to Ferragne et. al. (2019), we can focus on the classification of an individual test speaker in order to speculate on the inner workings of this particular classification system, gathering an indication of which features (pairwise distances) are contributing to a classification. To do this, we have taken one speaker that was incorrectly classified as a test case. We have then plotted that speaker's Y-ACCDIST matrix distances against the average Y-ACCDIST matrix distances that can represent the category the speaker should have been classified as. A plot like this should reveal the segmental pairs that are most dissimilar from its category's average Y-ACCDIST matrix, therefore exposing which segmental pairwise distances are likely to have been responsible for the incorrect classification. In the example demonstrated in Figure 3, the speaker's L1, in truth, is English, but was incorrectly classified as a speaker whose L1 is Mandarin Chinese.

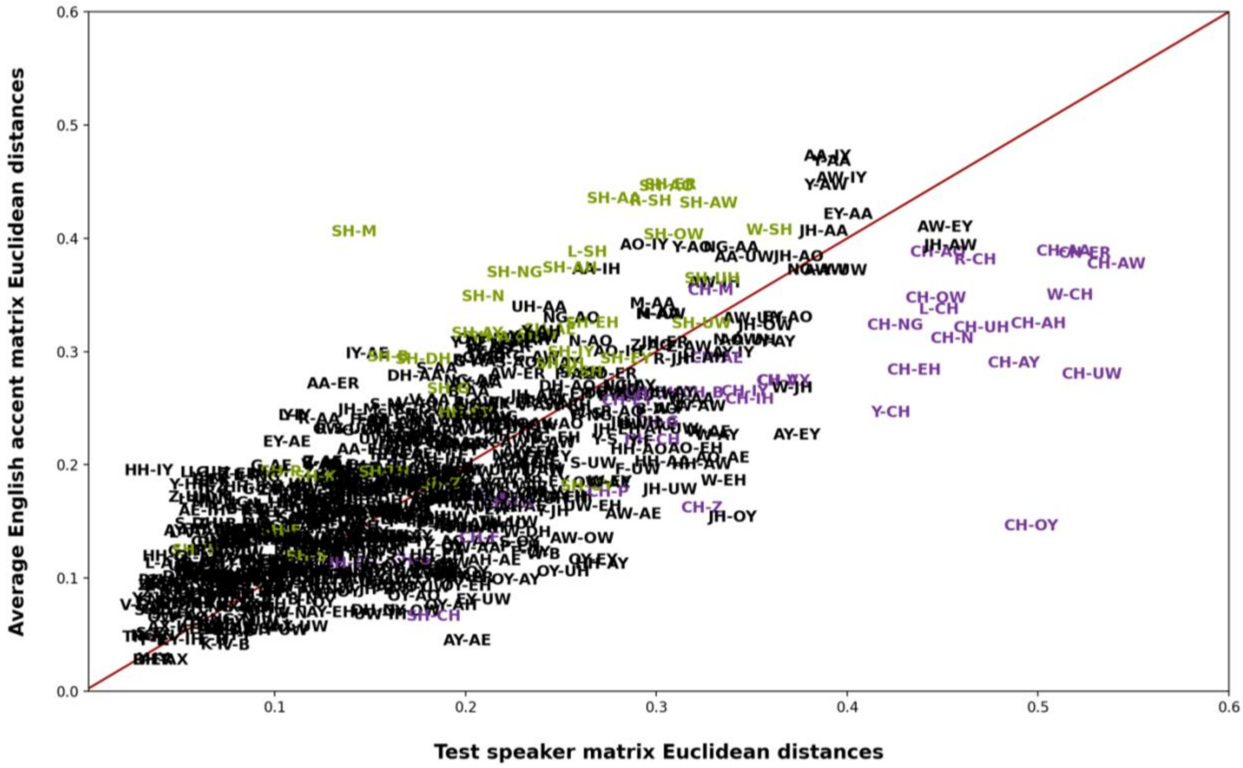

Figure 3: An incorrectly classified test speaker matrix (with an English L1), that was incorrectly classified as a speaker with Chinese as their L1, plotted against the corresponding pairwise segmental distances of the average English Y-ACCDIST matrix. The green and purple colours have been included to assist with the accompanying discussion.

While inspecting this scatterplot, we are interested in the segmental pairs that deviate most from the straight line, as these are expected to have contributed most to the incorrect classification. Among the observations, two specific segments seem to stand out as repeatedly deviating from the straight line as part of a number of different segmental pairwise combinations. In this particular case, these segments are symbolised by ‘CH’ (/tʃ/) and ‘SH’ (/ʃ/). For a clearer viewing of the pairwise distances concerned, those that include the CH segment are highlighted in purple, and those that include the SH segment are highlighted in green.

Because these two segments are repeatedly deviating from the straight line, we can reasonably expect that these two segments are at least partly responsible for the speaker’s misclassification. It could be that the speaker's CH and SH segments are not particularly typical of a speaker with English as an L1. There could be other factors that are feeding in to the performance, for example the frequencies that that the different segments occur in a speech sample. The interaction of these sorts of factors are further discussed, while continuing to refer to Figure 3, in the Discussion in Section 5.1.

4.2 Accent recognition as a verification problem

As stated further above, for calibration to take place an additional partition of data is required, not just one training set and a test set, which is effectively what was in place for the hard-decision results. We have therefore altered the training setup from the relatively straightforward *leave-one-out* cross-validation configuration implemented above. For the soft-decision experiments, 80 randomly-selected speakers per accent group have been used to train the Y-ACCDIST system. The remaining 20 speakers per accent (totalling 140 speakers) are used as calibration data. To generate the test trials, the 80 speakers per accent group are used in a *leave-one-out* cross-validation setup to test the system. Scores could then be computed based on the distance between the test speaker and the hyperplane/optimal margin formed for each accent class in the dataset – these scores are expected to express degree of similarity between a speaker and an accent class. As only 80 speakers have been used in the training dataset for these experiments, a lower number of speakers have therefore been used to train this verification system, compared to the classification system. This naturally might mean a reduced overall recognition rate compared to the hard-decision experiments presented above, because this data reduction might lead to weaker accent representations in the SVM. We therefore also present a hard-decision result where we only use 80 speakers per accent group for training to compare with the performance reported above. The following results in Table 6 are based on the segmental setting that yielded the best performance in the hard decision tasks, all-phonemes plus filled pauses.

Performance measure	**Result**
% Correct	61.4
Equal Error Rate (EER)	19.0
Cllr	0.6316
Cmin	0.6053
Ccal	0.0263

Table 6: Performance evaluation of the Y-ACCDIST-SVM non-native accent recognition task, having incorporated calibration.

It can be seen that Cllr is in agreement with the EER metric, be it far from 0 cost due to the rather high error rate, but indicating that informative logLRs are obtained (Cllr < 1). Furthermore, those logLRs are reliable as the cost is close to its minimum possible for that discriminating capability (low Ccal = Cllr - Cmin).

4.2.1 Tippett plot

Below in Figure 3 is a Tippett plot that allows us to inspect the distribution of likelihood ratios outputted by the system[7]. The plot consists of two lines to represent the performance of a single system: one line to represent same-accent likelihood ratios and the other to represent different-accent likelihood ratios. The same-accent line in the plot (the blue line) represents the cumulative proportion of the likelihood ratios that are smaller than the likelihood ratio marked on the x-axis, whereas the different-accent line (the red line) represents the cumulative proportion of likelihood ratios that are greater than the likelihood ratio marked on the x-axis.

[7] This Tippett plot was generated using a MATLAB script developed and made available by Geoffrey Morrison. See: http://geoff-morrison.net.

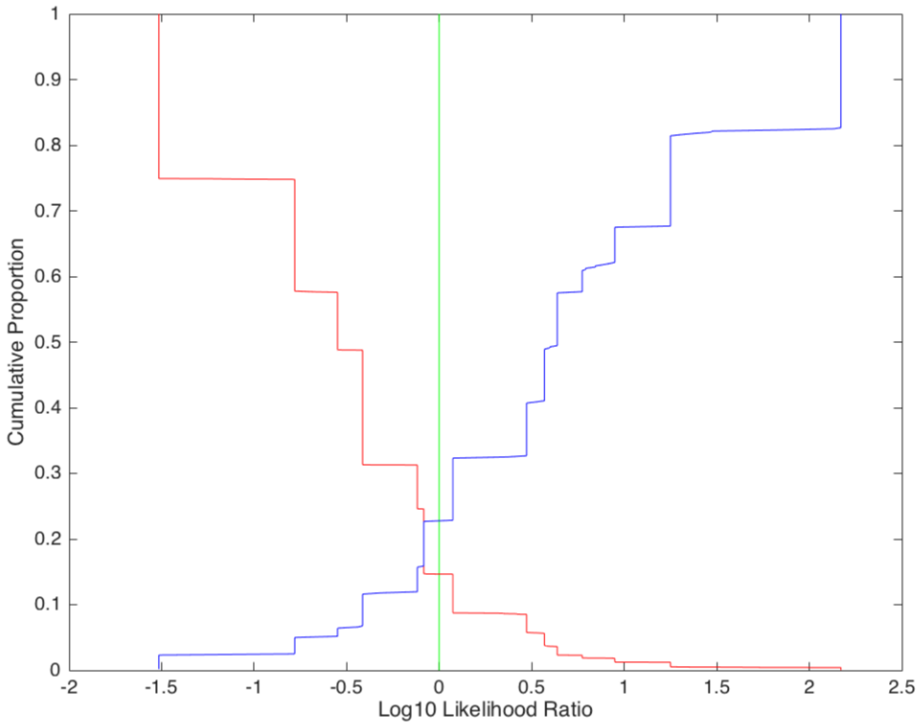

Figure 4: Tippett plot of the log-likelihood ratios generated from the Y-ACCDIST-SVM system after PAV calibration. The red line represents the different-accent proportion of log-LRs, while the blue line represents the same-accent proportion.

The stepwise nature of these lines is not typically seen so vividly in Tippett plots. This is a result of the non-parametric PAV calibration method that we used, in combination with a relatively small test set (compared to speaker recognition experiments, where we usually find Tippett plots). The points at which these two lines cross the vertical 0 line are of some interest. This shows us the proportion of these scores that reflect a value that supports the alternative hypothesis (i.e. the errors). Decomposing these, the point at which the same-accent line crosses 0 into the negative values shows the proportion of "false misses" the system makes, whereas the point at which the different-accent line crosses 0 indicates the proportion of these scores that we classify as "false hits".

5. Discussion

This paper has presented the performance of an automatic accent recognition system, that is proposed to be a more explainable alternative to other accent recognition and classification technologies. There are therefore two key aspects of the system to discuss in this section. This discussion will first start by further discussing system explainability, this time in light of the findings above. It will then move on to talk about the system's performance, as this is of course a vital factor when considering a system for any context.

5.1 Further explaining the inner workings of Y-ACCDIST

Explainability makes a system a more attractive prospect for applications like forensic ones. The end of Section 4.1 took a closer look at the features that were likely to contribute to the misclassification of a misclassified speaker, allowing a more detailed explanation of that specific error. In Section 4.1's example, a speaker with English as their L1 was mistaken for a speaker whose L1 is Mandarin Chinese. An inspection of Figure 3 suggested that the 'CH' and ‘SH’ segments may be partly responsible. Of course, this does provide a detailed explanation (i.e. about the precise phonetic nature of these segments that has led to this specific error), but identifying these particular segments as likely contributors to the misclassification can allow us to make further suggestions as to how the system may be operating, or how the system responds to the data. Having observed which segments were ‘misaligned’ in Figure 3, we can now put forward three different suggestions: the first suggestion relates to the phonetic realisation of the segments in question, the second suggestion relates to the recording conditions, and the third suggestion refers to the frequency of these segments. Starting with the first suggestion, it could be that this English speaker may produce retracted realisations of these particular segments, with the tongue tip positioned further back in the oral cavity. In a perceptual study, Lan (2020) showed how native Mandarin Chinese speakers are most likely to map the English /tʃ/ and /ʃ/ to the Chinese sounds, /tɕ/ and /ʂ/, respectively. As Lan (2020) points out, perceptual similarity can map on to phonetic similarity and so it could be that more retracted fricative and affricate realisations are characteristic of the Chinese accent models in our system. Slight realisational differences in this particular speaker’s productions of the affricate and fricative could have resulted in the contents of this speaker’s Y-ACCDIST matrix sharing a greater degree of similarity with the matrices of native Mandarin Chinese speakers than the matrices of native English speakers.

As an alternative (second) suggestion, it is possible that the ‘CH’ and ‘SH’ segments might be particularly problematic in these particular data as they are telephone-transmitted recordings, meaning that the higher frequencies of the spectrum are lost. Fricatives and affricates tend to be characterised in the higher frequencies of the spectrum. As a result, these segments may not be acoustically well represented and may subsequently present instability to a speaker's Y-ACCDIST matrix. Having inspected the equivalent scatterplots for other speakers, however, this does not appear to be a pattern that consistently occurs (and we would expect this to be the case if this explanation applied), therefore this is not necessarily the reason for these segments to arise as features responsible for the misclassification of this speaker.

We also acknowledge that there is also the interaction of a segment’s frequency (i.e. how often tokens of a segment occur in speech). This factor's influence on Y-ACCDIST accent classification was more comprehensively explored in Brown (2017) and Brown (2018). /tʃ/ is a very infrequent phoneme in English. /ʃ/ is not as infrequent but is still a relatively infrequent phoneme in English. As discussed in previous work, we might expect that more infrequent segments would naturally incur less stable average representations in the Y-ACCDIST matrix, and therefore present a greater misclassification risk.

It is plausible that a combination of the above reasons (differences in phonetic realisation, the consequences of the recording conditions and the segmental frequencies) are all playing a role in the incorrect classification of this particular speaker. As we present a range of reasons, it could be argued that we are no closer to explaining the system’s inner mechanisms. Ultimately, however, Figure 3 points towards parts of the speech signal that are very likely to be contributing towards correct and incorrect classifications. This indication has enabled us to offer further detailed reasons for these patterns. Without being able to observe the contribution of individual segments in the way that the Y-ACCDIST system allows, we would not be able to put forward any of these suggestions for a further detailed investigation. We would not be able to make such suggestions in relation to systems that are not segmentally informed.

Figure 3 may be simultaneously offering an argument to exclude consonants from our Y-ACCDIST matrices. It seems that in the above test case, certain consonants are likely to have been responsible for an error. This is not to say that these consonants are responsible for all of the system's errors, and when we compare the results generated from including and excluding consonants, they seem to bring about a performance benefit overall. It could also be argued that including the consonants could be diminishing the overall explainability of the system as we perhaps do not understand in as much detail what the distance between 'CH' and 'AE' might contribute, compared to the distance between 'AA' and 'AE'. The fact that 'CH' repeatedly emerged within segmental pairs that seemed to be distancing the particular test speaker from its correct category suggests that it is something to do with 'CH' as an individual segment, rather than the individual pair combinations, that is contributing to the overall classification outcomes. It is also plausible that individual consonant segments can repeatedly emerge in pairs that contribute to successful classifications. Although the individual segments form pairs in our Y-ACCDIST models, it seems that they also make individual contributions to the classifications which make for achievable explanations.

5.2 The performance of Y-ACCDIST

This paper has devoted a lot of attention to a system's explainability, but performance is also obviously a key factor in implementing systems and indeed in deciding whether a system should be implemented at all. The remainder of the Discussion section therefore places the performance results reported in this paper in the broader context of automatic accent recognition, and points towards other relevant performance considerations.

A valid criticism of the current work is that these experiments have not allowed for Y-ACCDIST's performance to be directly compared with the performance of other systems because a dataset has been used that is not typically used for accent classification purposes[8]. There are indeed corpora in existence that have been more widely tested by other accent classification researchers, and this is perhaps a research design decision that the authors would change if they were to carry out the study again. Having said that, drawing on the relevant literature that reports on other attempts to automatically classify speakers according to their first language, the system's performance generally seems to be in line with performances reported elsewhere. The INTERSPEECH 2016 Computational Paralinguistics Challenge (Schüller et. al., 2016) provided developers with the opportunity to showcase their automatic classification techniques on a large dataset of speech recordings from non-native speakers of English. The dataset provided researchers to run classification experiments on a dataset consisting of 11 classes of non-native accents of English, and hundreds of speakers per class. Huckvale (2016) presents his results on this dataset using a number of different systems based on Gaussian Mixture Models and Support Vector Machines. The highest performance reported was an accuracy of 69.8%.

While there are similarities between the data used in Huckvale's (2016) experiments and those presented in the current work, there are also differences. One difference is that there were more accent categories available for classification which makes the problem more difficult (i.e. it creates more opportunities for misclassification). Another difference is the amount of data available for training. There were many more speech samples per accent group that were used to train the system in Huckvale (2016) than there were in the experiments presented here. This is an additional consideration to keep in mind.

[8] Such a comparison on a different dataset was carried out in Brown (2016), however.

We also cannot assume that the quality of the dataset used in the above work was the same as the NIST subset used for the current work. We have already indicated further above that the NIST data used in this work is not necessarily the most reliable, which could also impact on a difference in performance between similar works. If one dataset is labelled more accurately than another, less noise is present in the data used to train the accent recognition models. In attempt to set the performance result in the context of an accent classification study that used a similar subset that was also extracted from the NIST collections, we can compare the results generated in this work with that of Bahari et. al. (2013). In their experiments, they also used NIST SRE datasets (speech samples taken from SRE challenges that took place between 2004 and 2008). In a similar way to the experiments presented here, they aimed to classify speakers according to their first language, based on their production of English. However, they had a different and smaller set of accent groups (and a smaller number of speakers per accent group, ranging from 32 to 84). They used five accent groups which were Russian, Hindi, US English, Thai and Vietnamese-Cantonese. The best-performing system configuration used the i-vector modelling technique, and they reported a result of 58% correct on this five-way classification task. While this result is not directly comparable, it suggests that the result produced by the Y-ACCDIST system is generally in line with other types of accent classification system on a similar task.

One thing that we have looked at that has moved the work here away from other accent recognition and language identification work is to extract an open-set style of decision output from the system, rather than bind the system to a closed-set type of task. To do this, we used the remaining accents in the database to collectively represent the “different-source” scores. This connects with the idea behind “relevant population” that is used to produce different-source scores in speaker recognition research. Within forensic speaker comparison research, there has been some work to look at what is meant by a “relevant population” Hughes and Foulkes (2015). They investigated how selecting different datasets that form an estimation of typicality (i.e. the denominator of the likelihood ratio formula) can affect the resulting likelihood ratio from an analysis. Ideally, we need to use reference data that are similar to the speech samples being analysed (for example, to match them in terms of speaker sex, accent, etc.). Among their findings, Hughes and Foulkes found that when a mismatched dataset is used to make speaker comparisons, we tend to produce weaker results in support of the defence hypothesis. There are also questions surrounding the size of the reference database and what number of speakers is sufficient to conduct a comparison. These kinds of problems also transfer to the problem of accent recognition when applying the likelihood ratio framework. Perhaps using the other accents in the NIST dataset was not a suitable choice. We should look further into the nature of the “relevant population” we use to compute the LRs for accent recognition (i.e. how many different accents we should include, etc.).

Another aspect to test that would be of interest to forensic applications is to test the Y-ACCDIST system’s performance on recordings produced in mismatched conditions. All the recordings used in these experiments were telephone conversations. It is expected that this matrix forms a simple, but robust, model of an individual’s accent. Because these are intra-recording calculations, without any reliance on external models, it is expected that the model logs only segmental differences with minimal interference from characteristics of voice quality or the channel. It is reasonable to hypothesise that a modelling method of this sort could provide a good way of overcoming minor channel mismatches. In this regard, it would be of interest to conduct experiments to assess the effects of channel mismatch on this modelling method. It could, of course, be the case that the Y-ACCDIST system is less robust to other kinds of mismatch (in instances of speech accommodation or disguise, for example). It would be of interest to test different types of recording mismatch on the Y-ACCDIST system in this way.

One minor aspect of performance that has arisen as a result of this work lies in the choice of distance metric in the construction of the Y-ACCDIST matrices. In Section 4.1, we pointed out that in one segmental configuration using cosine distance to compute the Y-ACCDIST matrices, the system produced a result of 62.6% correct. However, when we exchanged cosine distance for Euclidean distance, a lower result of 54.4% correct was achieved. The contrast between Euclidean distance and cosine distance is often referred to in the context of text classification, where texts are represented as vectors and are then broadly compared with one another using a metric. The advantage of cosine similarity is that it does not take into account the magnitude of a vector and is a way of normalising the length of documents Kataria and Singh (2013). Cosine distance takes into account the relative orientation of vectors, rather than the magnitude, unlike Euclidean distance. This may explain the difference in performance that we witnessed between the two metrics in these experiments.

Finally, there is an additional advantage to a segmentally-informed system of the sort presented here that was beyond the scope of this particular paper. Not only does the segmental information bring about more explainability, but it is also a lower-resource solution to the accent recognition problem. In the context of forensic applications where the set of accents we wish to distinguish between or recognise is expected to change, a solution that requires a smaller dataset for training is more desirable. This is because accessing enough data to train a system that relies much more heavily on machine learning can be extremely challenging. It has been implied in past work Brown (2016) that segmentally-informed approaches require a lot less data. Another study would be required to explore the precise detail of this hypothesis.

6. Conclusion

This paper has demonstrated the performance of a segmentally-informed, linguistically-motivated and more explainable automatic accent recognition system. It seems that the Y-ACCDIST system is able to perform generally in line with the performances demonstrated by other systems that have been used to classify speakers according to non-native accent groups. In addition, however, we were able take a close-up look at specific errors that the system made, and in turn demonstrate the detail of that error. This opened up further explanations around how this particular system is operating and how it is making its classifications. We make the point that the ability to do this is important to forensic applications. Overall, the advantage of using an alternative segmentally-informed approach is that the inner workings are more explainable and accountable, which are key traits when it comes to the administration of justice.

Acknowledgements

We are grateful to Daniel Ramos-Castro for his generous advice on calibration methods. This work was supported by the Economics and Social Research Council (ESRC).

References

Adadi, A. and Berrada, M. Peeking inside the black-box: a survey on Explainable Artifical Intelligence (XAI). *IEEE Access*. 6. 52138-52160.

D'Arcy, S., Russell, M., Browning, S. and Tomlinson, M. (2004). The Accents of the British Isles (ABI) corpus. In *Proceedings of Modélisations pour l'Identification des Langues*. Paris, France. 115-119.

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

Behravan, H. Hautamäki, V. and Kinnunen, T. (2013). Foreign accent detection from spoken Finnish using i-vectors. In *Proceedings of Interspeech*. Lyon, France. 79-82.

Behravan, H., Hautamäki, V. and Kinnunen, T. (2015). Factors affecting i-vector based foreign accent recognition: A case study in spoken Finnish. *Speech Communication*. 66. 118-129.

Biadsy, F., Soltau, H., Mangu, L, Navratil, J. and 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. 263-270.

Boril, H., Sangwan, A. and Hansen, J. (2012). Arabic Dialect Identification – 'Is the secret in the slence?' and other observations. In *Proceedings of Interspeech*. Portland, Oregon. 30-33.

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

Brown, G. (2016). Automatic accent recognition systems and the effects of data on performance. In *Proceedings of Odyssey: The Speaker and Language Recognition Workshop*. Bilbao, Spain.

Brown, G. (2017). Considering automatic accent recognition technology for forensic applications. Ph.D. thesis, University of York, UK.

Brown, G. (2018). Segmental content effects on text-dependent automatic accent recognition. In *Proceedings of Odyssey: The Speaker and Language Recognition Workshop*. Les Sables d'Olonne, France. 9-15.

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

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

Brümmer, N. and Van Leeuwen, D. (2006). On calibration of language recognition scores. In Proceedings of Odyssey: The Speaker and Language Workshop. San Juan, Puerto Rico.

Chen, T., Huang, C., Chang, E. and Wang, J. (2001). Automatic accent identification using Gaussian Mixture Models. In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding.* Italy.

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

Dehak, N., Kenny P., Dehak, R., Dumouchel, P. and Ouellet, P. (2011). Front-End Factor Analysis for Speaker Verfification. *IEEE Transactions on Audio, Speech and Language Processing*. 19. 788-798.

Drummond, R. (2012). Aspects of identity in a second language. ING variation in speech of Polish migrants living in Manchester, UK. *Language Variation and Change*. 24. 107-133.

Ferragne, E., Gendrot, C. and Pellegrini, T. (2019). Towards phonetic interpretability in deep learning applied to voice comparison. In *Proceedings of the International Congress of Phonetic Sciences*. Melbourne, Australia.

Ferragne, E. and Pellegrino, F. (2010). Vowel systems and accent similarity in the British Isles: Exploiting multidimensional acoustic distances in phonetics. *Journal of Phonetics*. 38. 526-539.

Franco-Pedroso, J. and González-Rodríguez (2016). Linguistically-constrained formant-based i-vectors for automatic speaker recognition. *Speech Communication*. 76. 61-81.

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

Grabe, E. (2004). Intonational variation in urban dialects of English spoken in the British Isles. In P. Gilles and J. Peters (Eds.) *Regional Variation in Intonation*. Linguistische Arbeiten, Tuebingen. 9-31.

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

Huckvale, M. (2004). ACCDIST: a metric for comparing speakers' accents. In *Proceedings of the International Conference on Spoken Language Processing*. Jeju, Korea. 29-32.

Huckvale, M. (2007). ACCDIST: An accent similarity metric for accent recognition and diagnosis. In C Müller (Ed.) *Speaker Classification, Volume 2 of Lecture Notes in Computer Science*. Springer-Verlag, Berlin Heidelberg. 258-274.

Huckvale, M. (2016). Within-speaker features for native language recognition in the Interspeech 2016 Computational Paralinguistics Challenge. In *Proceedings of Interspeech*. San Francisco, USA. 2403-2407.

Hughes, V. and Foulkes, P. (2015). The relevant population in forensic voice comparison: Effects of varying delimitations of social class and age. *Speech Communication*. 66. 218-230.

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

Kajarekar, S. S., Scheffer, N. Graciarena, M., Shriberg, E., Stolcke, A., Ferrer, L. and Bocklet, T. (2009). The SRI NIST 2008 Speaker Recognition Evaluation System. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*. Taipei, Taiwan. 4205-4208.

Kataria, A. and Singh, M. D. (2013). Review of data classification using k-nearest neighbour algorithm. *International Journal of Emerging Technology and Advancing Engineering*. 3. 354-360.

Lan, Y. (2020). Perception of English fricatives and affricates by advanced Chinese learners of English. In *Proceedings of Interspeech*. Shanghai, China. 4467-4470.

de Leeuw, E. (2007). Hesitation markers in English, German and Dutch. *Journal of Germanic Linguistics*. 19. 85-114.

Lo, J. (2020). Between *Äh(m)* and *Euh(m)*: The Distribution and Realization of Filled Pauses in the Speech of German-French Simultaneous Bilinguals. *Language and Speech*. 63. 746-768.

McDougall, K. and Duckworth, M. (2017). Profiling Fluency: An analysis of individual variation in disfluencies in adult males. *Speech Communication*. 95. 16-27.

Morrison, G. S. (2013). Tutorial on logistic-regression calibration and fusion: converting a score to a likelihood ratio. *Australian Journal of Forensic Sciences*. 45. 173-197.

Najafian, M., Safavi, S., Weber, P. and Russell, M. Identification of British English regional accent using fusion of i-vector and multi-accent phonotactic systems. In *Proceedings of Odyssey: The Speaker and Language Recognition Workshop*. Bilbao, Spain.

Pryzybocki, M. and Martin, A. (2004). NIST Speaker Recognition Evaluation chronicles. In *Proceedings of Odyssey: The Speaker and Language Recognition Workshop.* Toledo, Spain.

Ramos-Castro, D. González-Rodríguez, J. and Ortega-Garcia, J. (2006). Likelihood ratio calibration in a transparent and testable forensic speaker recognition framework. In Proceedings of Odyssey: The Speaker and Language Recognition Workshop. San Juan, Puerto Rico.

Reynolds, D. and Rose, R. (1995). Robust Text-Independent Speaker Identification using Gaussian Mixture Speaker Models. *IEEE Transactions on Speech and Audio Processing*. 3. 72-83.

Samek, W., Wiegand, T., Müller, K-R. (2017). Explainable Artificial Intelligence: understanding visualizing and interpreting deel learning models. URL: https://arxiv.org/pdf/1708.08296.pdf

Schüller, B., Steidl, S., Batliner, A., Hirschberg, J., Burgoon, J., Baird, A., Elkins, A., Zhang, Y., Coutinho, E. and Evanini, K. Computational Paralinguistics Challenge. Deception, sincerity and native language. In *Proceedings of Interspeech*. San Francisco, USA. 2001-2005.

Shon, S., Ali, A., & Glass, J. (2018). Convolutional neural network and language embeddings for end-to-end dialect recognition. In Proceedings of Odyssey: the speaker and language recognition workshop. Les Sables d'Olonne, France. 98-104.

Snyder, D., Garcia-Romero, D., Povey, D. and Khudanpur, S. (2017). Deep Neural Network Embeddings for Text-Independent Speaker Verification. In *Proceedings of Interspeech*. Stockholm, Sweden. 999-1003.

Stuart-Smith, J. (1999). Glasgow: Accent and Voice Quality. In P. Foulkes and G. Docherty (Eds.) *Urban Voices: Accent Studies in the British Isles*. Routledge, London. 203-222.

Tully, G. (2020). Codes of Practice and Conduct for forensic science providers and practitioners in the Criminal Justice Sysytem. FSR-C-100. Issue 5. The UK Government. URL: https://www.gov.uk/government/publications/forensic-science-providers-codes-of-practice-and-conduct-2020.

Vergyri., D., Lamel, L. and Gauvain, J-L. (2010). Automatic Speech Recognition of Multiple Accented English Data. *Proceedings of Interspeech*. Makuhari, Chiba, Japan. pp 1652-1655.

Vieru, B., de Mareüil. and Adda-Decker, M. (2011). Characterisation and Identification of Non-native French Accents. *Speech Communication*. 53. 292-310.

Watt, D. (2010). The identification of the individual through speech. In C. Llamas and D. Watt (Eds.) Language and Identities. Edinburgh University Press, Edinburgh. 76-85.

Watt, D., Harrison, P., Cabot-King, L. (2020). Who owns your voice? Linguistic and legal perspectives on the relationship between vocal distinctiveness and the rights of the individual speaker. *The International Journal of Speech, Language and the Law*. 26. 137-180.