Exploring forensic accent recognition using the Y-ACCDIST system

Georgina Brown
University of York, UK
gab514@york.ac.uk

Abstract

Forensic speech scientists may sometimes be faced with the task of extracting information about an unknown speaker in a recording. It is proposed here that accent recognition technology could assist analysts in such cases and we begin to explore the Y-ACCDIST system's potential for this purpose. Research on Y-ACCDIST so far has largely focussed on its ability to distinguish between varieties which are much more similar to one another than previous automatic accent recognition research [1]. The experiments presented here build on this and challenge Y-ACCDIST in other ways relevant to forensic applications: spontaneous speech data and degraded data.

Index Terms: forensic speech technology, accent recognition

1. Introduction

Most of the work by a forensic speech analyst is on *speaker comparison* tasks. This is the task of analysing one speech recording against another to assess whether the speech in both recordings was produced by the same speaker. Speaker recognition technology can assist with this kind of task. On a lesser scale, there is another task a forensic analyst might be faced with called *speaker profiling*. Rather than comparing two samples, this type of task involves analysing a speech sample to extract information about the speaker. One example of the type of real-life case is a ransom telephone call [2]. Useful information might include his or her geographical origin. Currently, the speaker profiling task is done manually with no technology to assist with this kind of task. This paper further investigates the possibility of testing a specific automatic accent recognition system, Y-ACCDIST (the York ACCDIST-based automatic accent recognition system) [1], for forensic applications.

The following addresses the idea of forensic accent recognition by first giving an overview of accent recognition in section 2. It first reviews forensic speaker profiling (from a manual point of view) before then exploring what has already been achieved by speech technologists in automatic accent recognition more generally. Section 3 presents experiments of the Y-ACCDIST system being applied to forensically relevant accent data. This includes a technical description of Y-ACCDIST. Section 4 summarises the paper and puts forward ways in which the research area of automatic accent recognition for forensic applications could be further developed.

2. Accent Recognition

While accent recognition and perception is of course of sociolinguistic interest, this section will initially focus on accent recognition in relation to the forensic context. The second part of this section will review past automatic accent recognition studies, which have not necessarily been motivated by forensic applications, but mostly by automatic speech recognition.

2.1. Forensic Speaker Profiling

Little literature exists on the topic of speaker profiling. As already stated, forensic speech science is largely concerned with the speaker comparison task. This imbalance of research attention is naturally reflective of a forensic analyst's workload. One study which could be loosely tied to the task of speaker profiling is [3]. They assessed the ability of different analysts to classify speakers based on accent, in the context of *Language Analysis for the Determination of Origin* (LADO). LADO is applied to a small proportion of asylum seeker applications which require additional assessment to establish information about the applicant. More specifically, we might want to know if the applicant is from where he or she claims, and an analysis of a recorded interview can assist with this. [3] had recorded stimuli of native Ghanaian English speakers and Nigerian English speakers. Different analysts (with varying degrees of expertise) were asked the question for each speaker: Do you believe this person is speaking Ghanaian English? They compared the performance of academic phoneticians, undergraduate students, native speakers of Ghanaian English and LADO professionals. They found that the native speakers were the highest-performing group at this task with an overall classification rate of 86% correct. It would be of great interest to see how an automatic system would compare in a similar task, and whether human analysts and an automatic system could combine their strengths to obtain an overall higher result. This has not yet been explored. To begin to do this, the following section covers some of the developments of automatic accent recognition systems within speech technology.

2.2. Automatic Accent Recognition

Traditionally, automatic accent recognition research has focussed on building systems which improve the overall performance of automatic speech recognition systems. By identifying the accent of a speaker before attempting to recognise what is being said, we can raise speech recognition rates [4].

A range of approaches have been trialled to conduct automatic accent recognition. Taking inspiration from the work of [5] in Language Identification (LID), [6] applied a variation of a *Phone Recognition followed by Language Modelling* (PRLM) system to the task of distinguishing between different Arabic varieties. By estimating the sequence of phones in the unknown utterances, using a phoneme recogniser, we can then assess this sequence in terms of how likely the sample belongs to each of the varieties in the reference set. [6] suggest that the varieties of Arabic they use are distinguishable by the different phonemic sequences, and this is reflected in the 6.0% Equal Error Rate they achieve.

Other approaches have included more acoustic modelling, rather than phonotactic modelling as above. [7] applied Gaussian Mixture Models (GMMs), combined with Mel -Frequency

Cepstral Coefficients (MFCCs), to the classification problem of distinguishing between four Mandarin Chinese accent varieties. With this approach, they achieved an error rate of 11.7% and 15.5% for females and males respectively.

The approaches discussed so far are text-independent. Text-independent systems come with the practical advantage of not needing an orthographic transcription to accompany the speech sample. As a consequence, this broadens the pool of applications these systems can be used for. However, some applications (like the forensic application) might benefit from a text-dependent option if it can bring the precision that is required. [8] introduced the ACCDIST metric for automatic accent recognition. It is a text-dependent method which calculates intra-speaker distances between vowel sounds found in a speech sample. A clearer idea of how an ACCDIST-based system works is given in the Y-ACCDIST system description further below in 3.1.

[9] compared a number of automatic accent recognition systems on the same corpus. They compared a combination of text-independent and text-dependent systems. These were GMM-based acoustic systems and ACCDIST-based systems. They were all tested on the *Accents of the British Isles* corpus [10]. This corpus contains recordings of speakers from 14 locations spanning the breadth of the British Isles. In a 14-way classification their highest-performing system (which was a text-dependent ACCDIST-based system) achieved 95.18% accuracy. This unsurprisingly outperformed their text-independent GMM-based systems (which achieved 61.13% and 76.11% on the same task). Of course, a text-dependent system outperformed the text-independent systems, but the text-independent systems still seem to perform well on a 14-way classification task.

2.3. Y-ACCDIST

Y-ACCDIST is a text-dependent accent recognition system based on the ACCDIST metric [8]. What separates Y-ACCDIST from previously developed ACCDIST-based systems in [8] and [9] is that Y-ACCDIST is able to process content-mismatched speech. Past ACCDIST-based systems have relied on testing and training speech content to be the same, as the vowel segments they analyse and compare are restricted to specific contexts. In the case of [8], word-specific vowels are analysed and compared, and in the case of [9], triphone-context vowels are used. Y-ACCDIST, on the other hand, collapses phones into phonemic categories to analyse and compare. Details of how this is done are given in section 3.1.

Y-ACCDIST has previously been tested with the forensic application in mind. [11] explored Y-ACCDIST's performance on varieties which were assumed to have a greater degree of similarity between them. Past research on automatic accent recognition (due to the focus on automatic speech recognition) has been concerned with categorising speakers into accent groups with great differences between them. This is of course useful to automatic speech recognition which suffers due to the great variation among speakers. However, for the forensic application, it is of interest to challenge an accent recognition system's sensitivity in terms of how well it can distinguish between varieties which are much more similar. Past research ([1] [11]) has demonstrated Y-ACCDIST's ability to distinguish between the varieties in the *Accent and Identity on the Scottish/English Border* (AISEB) corpus [12].

AISEB was collected for sociolinguistic purposes and contains speakers from Berwick-upon-Tweed, Carlisle, Eyemouth and Gretna. A subset of AISEB was used to test Y-ACCDIST, where the recorded reading passage was taken from 30 speakers from each of the four locations in the corpus. Testing all 120 speakers in a *leave-one-out* training and testing configuration, Y-ACCDIST was able to classify these speakers into the four accent groups with a rate of 86.7% correct.

In a similar way to [9], [13] compared Y-ACCDIST against three different GMM-based accent recognisers. However, these comparative experiments were run on the AISEB corpus of geographically-proximate accents, rather than the ABI corpus. As already discussed, [9] found the text-dependent ACCDIST-based systems to outperform the text-independent GMM-based systems. [13] found that this was also the case when testing these kinds of systems on the AISEB corpus, but the ACCDIST-based systems appeared to be much more robust to more similar accent varieties. The text-independent GMM-based systems seemed to suffer much more due to the more challenging distinctions to make between more similar varieties. On the four-way AISEB accent classification task the GMM-UBM system achieved 37.5% correct, which is well below the 86.7% the Y-ACCDIST system achieved.

While researching system performance on geographically-proximate accents does move towards more forensically relevant experiments, there are of course still a number of aspects that are forensically relevant, which remain uncovered. The experiments presented below address other areas which are of interest to the forensic application: spontaneous speech data and degraded data.

3. Experiments

First, this section gives a technical description of the inner workings of Y-ACCDIST. Following this, the experiments and results of testing Y-ACCDIST on conversational spontaneous speech data, and degraded data are given.

3.1. Y-ACCDIST Development Details

The following steps take place to train Y-ACCDIST and classify an unknown speech sample:

1. For each speaker in the training data, a speech sample and orthographic transcription are passed through a forced aligner. The forced aligner was built using the Hidden Markov Model Toolkit (HTK) [14] and a British English phoneset.
2. A midpoint MFCC (12 coefficients) was extracted for each phone in each speaker's sample.
3. An average MFCC is calculated for each phoneme in the phoneme inventory.
4. In a matrix (a Y-ACCDIST matrix), the Euclidean distance between every pair of phonemes is calculated using the average MFCC representations. One of these matrices is computed for each speaker. These distances are expected to capture accent-specific information. To illustrate, we can look at the *foot-strut split* in British English. A typical speaker of Northern English English will produce the vowels in *foot* and *strut* very similarly. A typical speaker of Southern English English, however, will produce the vowels of *foot* and *strut* differently. The Euclidean distance between these two vowels in a Northern speaker's matrix will therefore be smaller than the one calculated for a Southern speaker. An illustration of a matrix is given in the figure below:

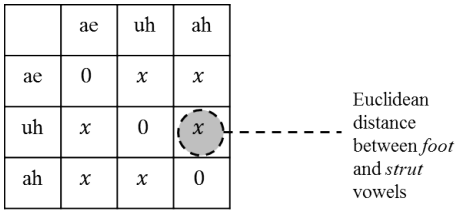

Figure 1: Illustration of part of a Y-ACCDIST matrix.

For the experiments in this paper, all phonemes in the phoneset (vowels and consonants) were included in the Y-ACCDIST matrices

5. Using these Y-ACCDIST matrices to represent each of the training speakers, they are fed into a Support Vector Machine (SVM) with a linear kernel. For each accent group on rotation, the training speaker Y-ACCDIST matrices are plotted to form a *one-against-the-rest* in the SVM, which is effectively multi-dimensional space, forming an optimal 'hyperplane' each time. When classifying an unknown speaker's speech sample, it is first converted into a Y-ACCDIST matrix in the way described in steps 1-3. On each of the rotations for each accent, the unknown speaker's matrix is fed into the SVM. The accent group the unknown matrix forms the clearest margin with is the accent group the unknown speaker is assigned.

3.2. Spontaneous Speech

So far, Y-ACCDIST has largely been tested on tightly controlled experimental data. The speech data in past experiments on the AISEB corpus were done where each speaker was recorded reading the same passage. A change in dataset allows for more forensically relevant experiments to take place, where a large amount of spontaneous speech is produced by speakers of different accents. A description of the selected corpus is given below:

3.2.1. The Corpus

The data used for the experiments presented here were taken from the *Language Change in Northern English* corpus [15]. A subset of speakers' conversational speech recordings (with a sampling rate of 44.1kHz), along with their orthographic transcriptions were taken for these experiments. The speech recordings of 45 adult speakers (male and female) from Manchester, Newcastle and York (15 in each group), along with their orthographic transcriptions, were manually pre-processed. 10 minutes of net speech per speaker (and the accompanying orthographic transcriptions) were prepared for the experiments shown below.

3.2.2. Results

On the three-way accent classification task, distinguishing between speakers from Manchester, Newcastle and York, Y-ACCDIST achieved a recognition rate of 80.0% correct. Displayed below is the resulting confusion matrix for this task.

Table 1: Confusion matrix of results generated using spontaneous speech recordings.

Accent	Manc.	Newc.	York.
Manc.	**12**	0	3
Newc.	2	**12**	1
York.	1	2	**12**

3.3. Degraded Data

While testing Y-ACCDIST on spontaneous speech is more forensically relevant than previous experiments, the quality of the recordings is still unrealistic to the application. Forensic speech scientists mostly work with telephonic-quality or other degraded recordings. To begin to explore Y-ACCDIST's potential as an analytical tool on degraded data, the data used in the experiments above were degraded to a quality which resembles telephony. The recordings were downsampled to 8kHz, bandpass-filtered 250-3500Hz, and a-law compression was applied. The same experiments were re-run to achieve a recognition rate of 64.4% correct. The confusion matrix attached to this result is given in the table below:

Table 2: Confusion matrix of results generated using spontaneous speech recordings degraded down to telephonic quality.

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

Between the results generated from the good-quality data and those from degraded data, we can see an expected drop in performance. When we compare the two confusion matrices from each of the quality conditions, it is interesting to observe which particular accent groups appear to suffer more in the degraded condition. While the number of correctly classified Newcastle speakers seem to remain approximately the same under the degraded condition, the number of correctly classified Manchester and York speakers fall. This may be indicating that out of the three accent groups, Newcastle is the most distinct variety and so can withstand a more challenging data condition. Research on a larger data pool would be required to investigate this further.

3.4. Quantity of Data

One key criticism of the experiments presented above is the quantity of speech used to train and test the system. A 10-minute speech sample per speaker does not realistically align with what would normally be available to a forensic analyst. We can diminish the quantity of speech per speaker and run the system at different speech sample lengths. In increments of 30 seconds, the graph below demonstrates the effect of speech sample length on classification rate under the good-quality data condition.

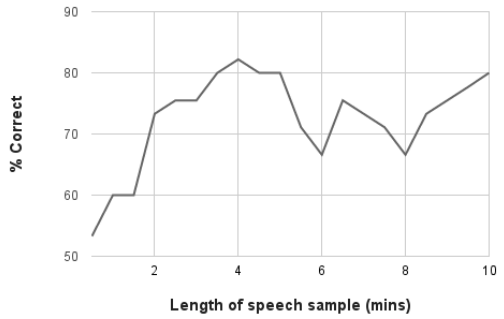

Figure 2: Classification rates with varying length of speech sample.

The graph shows a general improvement in performance between 30-second samples and 10-minute samples. However, it does not appear to be a stable increase in classification rate. While we reach a good recognition rate of 75.5% correct with 3 minutes of speech (a length much more reflective of what might be available to a forensic analyst), performance fluctuates beyond this. Due to the relatively small dataset, a larger dataset would be required to investigate this further. However, it might be that the change in segmental distribution that occurs in different lengths of speech samples has an unpredictable effect on performance.

4. Summary and Discussion

This paper has further entertained the idea of applying automatic accent recognition technology to forensic casework. The Y-ACCDIST system was applied to the *Language Change in Northern English* corpus to assess its performance on spontaneous speech data (compared to reading passage data which Y-ACCDIST has already been tested on). On a three-way accent classification task, 80.0% correct was achieved. Degrading this data down to a quality that resembled telephony generated a result of 64.4% correct.

These findings only just begin to explore a specific automatic accent recognition system's performance in a forensic context. Numerous avenues for further research exist. Below covers just three of these:

- As already stated above, much of the focus in forensic speech science is the task of speaker comparison. It would be interesting to see whether Y-ACCDIST could assist in some speaker comparison cases (i.e, can a Y-ACCDIST matrix represent a speaker's pronunciation pattern to specifically distinguish him or her from a wider speaker population?).
- Y-ACCDIST is highly reliant on segmental information to be able to work. In a given accent recognition task, certain segments are more telling than others (and this will vary depending on the particular accent varieties we are dealing with). It is therefore reasonable to assume that an unknown speech sample might need to contain a certain distribution of segments to obtain a reliable result. Another line of inquiry is to establish the segmental criteria an unknown speech sample needs to meet in order to be reliably processed and assessed by Y-ACCDIST.
- One important topic in forensic speech science is to do with the conclusion outputs of a system. In the experiments in this paper, only a closed-set classification task has been conducted. In reality, it might be more useful to determine the likelihood of a speaker belonging to a certain accent group with a more open-set approach. Integrating likelihood ratios is therefore considered an important development to the system.

Future research will target these directions.

5. References

[1] G. Brown, "Y-ACCDIST: An automatic accent recognition system for forensic applications," Master's thesis, University of York, UK, 2014.

[2] P. Foulkes and P. French, "Forensic speaker comparison: A linguistic-acoustic perspective," ser. The Oxford Handbook of Language and Law, P. Tiersma and L. Solan, Eds. Oxford University Press, 2012, pp. 557–572.

[3] P. Foulkes and K. Wilson, "Language analysis for the determination of origin," in *Proceedings of the 17th International Congress of Phonetic Sciences*, Hong Kong, 2011, pp. 692–694.

[4] M. Najafian, A. DeMarco, S. Cox, and M. Russell, "Unsupervised model selection for recognition of regional accented speech," in *Proceedings of Interspeech*, Singapore, 2014, pp. 2967–2971.

[5] M. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Transactions on Speech and Audio Processing*, vol. 4, pp. 31–44, 1996.

[6] F. Biadsy, H. Soltau, L. Mangu, J. Navratil, and J. Hirschberg, "Discriminative phonotactics for dialect recognition using context-dependent phone classifiers," in *Proceedings of Odyssey: The Speaker and Language Recognition Workshop*, Brno, Czech Republic, 2010, pp. 263–270.

[7] T. Chen, C. Huang, E. Chang, and J. Wang, "Automatic accent identification using Gaussian Mixture Models," in *IEEE workshop on ASRU*, Italy, 2001.

[8] M. Huckvale, "ACCDIST: a metric for comparing speakers' accents," in *Proc. International Conference on Spoken Language Processing*, Jeju, Korea, 2004, pp. 29–32.

[9] A. Hanani, M. Russell, and M. Carey, "Human and computer recognition of regional accents and ethnic groups from British English speech," *Computer Speech and Language*, vol. 27, pp. 59–74, 2013.

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

[11] G. Brown, "Automatic recognition of geographically-proximate accents using content-controlled and content-mismatched speech data," in *Proceedings of the 18th International Congress of Phonetic Sciences*, 2015.

[12] D. Watt, C. Llamas, and D. E. Johnson, "Sociolinguistic variation on the Scottish-English border," in *Sociolinguistics in Scotland*, R. Lawson, Ed. London: Palgrave Macmillan, 2014, pp. 79–102.

[13] G. Brown, "Automatic accent recognition systems and the effects of data on performance," in *Proceedings of Odyssey: The Speaker and Language Recognition Workshop*, Bilbau, Spain, 2016.

[14] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book for HTK Version 3.4*. Cambridge: Cambridge University Engineering Department, 2009.

[15] W. Haddican, P. Foulkes, V. Hughes, and H. Richards, "Interaction of social and linguistic constraints of two vowel changes in Northern England," *Language, Variation and Change*, vol. 25.