AUTOMATIC RECOGNITION OF GEOGRAPHICALLY-PROXIMATE ACCENTS USING CONTENT-CONTROLLED AND CONTENT-MISMATCHED SPEECH DATA

Georgina Brown

University of York
gab514@york.ac.uk

ABSTRACT

This contribution advances us towards developing an automatic accent recognition system with greater practical potential. Y-ACCDIST is a text-dependent accent recognition system intended for forensic applications. Given a speech sample, it aims to identify the speaker's geographical origin, which may be useful in forensic contexts. While promising accent recognition rates are reported (86.7% on a four-way classification task), these have been obtained by comparing speakers producing the same reading passage. The focus here is to observe system performance on content-mismatched (spontaneous) speech data and to speculate about ways to improve the system when it is faced with more challenging data. The ability to process content-mismatched data separates Y-ACCDIST from similar past systems.

Keywords: Automatic Accent Recognition, Forensic Phonetics, Speaker Profiling

1. INTRODUCTION

Developed for forensic applications, the York ACCDIST-based automatic accent recognition system (Y-ACCDIST) aims to identify a speaker's geographical origin given a speech sample and transcription. Based on the ACCDIST metric [4], Y-ACCDIST makes use of relative distances between phonological units to model speakers' accents.

One key feature which separates Y-ACCDIST from other ACCDIST-based accent recognition systems (e.g. [2], [3], [4], [5]) is that it is designed to process content-mismatched (spontaneous) speech data. Obviously, the ability to work on content-mismatched data has greater practical potential, particularly when considering the forensic application.

2. SYSTEM DEVELOPMENT

As noted above, Y-ACCDIST has been developed to process content-mismatched data, unlike previous ACCDIST-based systems. ACCDIST makes use of segmental units to represent a speaker's speech sample. The type of segmental units used determines whether it is possible for it to work on content-mismatched data or not. For now, we will only include vowel segments. In the case of [4], word-level vowel segments were used to compare accents. In [2] and [3], triphone vowel segments were used. Taking the first clause in the AISEB reading passage *Fern was a nurse from Harrogate*, Table 1 demonstrates these two segment types:

Table 1: Context-dependent segment types.

Segment Type	Segments found in clause
Word-context vowels	vowel in *Fern* vowel in *was* vowel in *nurse* vowel in *from* vowel in first syllable of *Harrogate* vowel in third syllable of *Harrogate*
Vowel triphones	/ɜ/ in /fɜːn/ /ɒ/ in /wɒz/ /ɜ/ in /nɜːs/ /ɒ/ in /rɒm/ /a/ in /har/ /eɪ/ in /geɪt/

These are both, to different degrees, context-dependent segment types and are detrimental to system performance and applicability in two main ways:

1. Context-dependent segments occur infrequently within a single sample and so it is difficult to find a representative number of these segments.
2. When comparing speakers, speech samples need to have a number of segmental unit types in common. We can expect to find few tokens of each highly context-dependent segment type in a single speech sample, reducing the

chances of finding enough common segmental types to be able to make comparisons with other speech samples. This lowers the probability of successfully applying a system to content-mismatched data.

The advantage of using context-dependent segments, however, is that they serve to eliminate realisational differences brought about by coarticulation.

Y-ACCDIST collapses these context-dependent segmental units into context-independent phonemes, which opens up the possibility of performing accent recognition on content-mismatched data. The following outlines the workings of Y-ACCDIST.

2.1. Data

Speech data from the AISEB (Accent and Identity on the Scottish English Border) corpus [7] were used for the experiments presented here. Speech recordings were collected from informants from four locations close to the Scottish/English border: Berwick-upon-Tweed, Eyemouth, Carlisle and Gretna. The experiments make use of the reading passage recordings and recorded answers to interview questions of 30 speakers from each of the four locations (N=120). Within each of these groups of 30, a further divide between two age groups can be made: 15 younger speakers (aged 14-27) and 15 older speakers (aged 54-93).

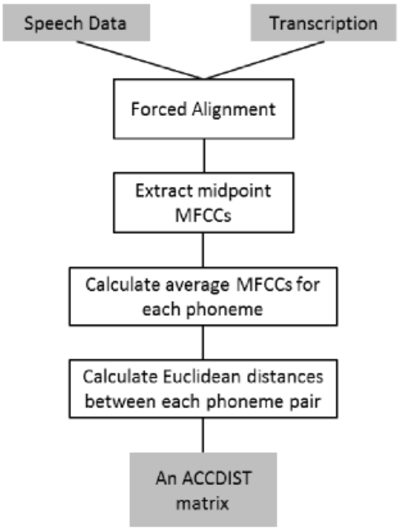
2.2. The Y-ACCDIST System

Y-ACCDIST can be described in two stages: accent modelling and classification.

2.2.1. Accent Modelling

Each speaker's speech sample is processed to form a representative matrix of the individual's accent. To achieve this, the samples are passed through a forced aligner (built using the Hidden Markov Model Toolkit (HTK) [9]) and the midpoint 12-element MFCC vector is extracted from each vowel phone. For every vowel phoneme within the inventory, the corresponding MFCC vectors are brought together to produce an average MFCC vector to represent that phoneme. These phonemes form the foundations of a matrix, enabling Euclidean distances to be calculated between each phoneme pair combination. The process is illustrated in Fig. 1.

Figure 1: The process of generating an ACCDIST matrix for a single speaker.

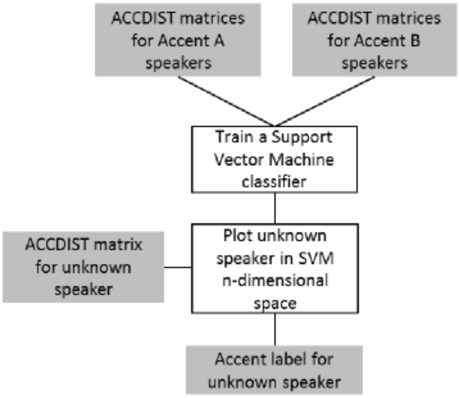

Speech Data
Transcription
Forced Alignment
Extract midpoint MFCCs
Calculate average MFCCs for each phoneme
Calculate Euclidean distances between each phoneme pair
An ACCDIST matrix

ACCDIST matrices form a normalised representation of a speaker's accent through expression of phonemic similarity. Taking the vowels in *foot* and *strut* as an example, we expect that the Euclidean distance between these two vowels would be larger for a speaker of Southern English English than it would be for a speaker of Northern English English. This is because the realisations of these vowels are the same for a typical Northern speaker, while they are different for a Southern speaker. These differences throughout the phonemic inventory are expressed through ACCDIST matrices. Intra-speaker calculations like this eliminate the effects of voice quality factors, so accents can be compared across a range of speakers.

2.2.2. Classification

The ACCDIST matrices are then fed into a Support Vector Machine (SVM) classifier [6]. One accent group at a time is taken, and every speaker in that group is effectively plotted within multidimensional space. Speaker matrices from all other groups are also plotted as one collective category, creating a 'one-against-the-rest' configuration. Between these two groups of speakers, an optimal hyperplane is formed, which acts as the decision boundary. The accent groups rotate in order for each category to form a hyperplane against 'the rest'. In effect, a SVM is created for every accent group involved (four, in the case of the AISEB data). Given an unknown speaker's speech sample and transcription, the recording can be converted into an ACCDIST matrix (in the way described in 2.2.1). It can subsequently be plotted in each of the four SVMs. The margin between the hyperplane and the unclassified ACCDIST matrix is observed in each. The SVM in which the clearest margin is formed determines the speaker's class label. The general process is outlined in Fig. 2 below:

Figure 2: Accent classification using SVMs.


ACCDIST matrices for Accent A speakers
ACCDIST matrices for Accent B speakers
Train a Support Vector Machine classifier
ACCDIST matrix for unknown speaker
Plot unknown speaker in SVM n-dimensional space
Accent label for unknown speaker

3. EXPERIMENTS

3.1. Content-Controlled Data

Previous ACCDIST-based accent recognition systems have only been tested on content-controlled speech data (i.e, all speakers produce the same read prompts). The Y-ACCDIST system is built in such a way that this does not have to be the case, and its performance on content-mismatched data is presented in the next section. However, for comparison purposes, this section presents classification results when all 120 AISEB speakers are recorded reading the same passage (approx. 3-4 mins in duration).

Table 2: Accent recognition rates for content-controlled reading passage data (25% correct expected at chance for 4-way tasks, 12.5% expected for 8-way tasks).

Data configuration	% Correct
4-way (all speakers) N=120	86.7
4-way (older speakers) N=60	83.3
4-way (younger speakers) N=60	83.3
8-way (all speakers) N=120 (locations and age groups)	69.2

3.1.1. Segmental Context-Dependency

Returning to the point made in Section 2, this subsection briefly visits the effects different segmental unit types have on the recognition task presented here. The four-way classification task on the AISEB varieties using all 120 speakers was run using word-level vowel segments and triphone vowel segments. These were compared with Y-ACCDIST's default context-independent vowel phoneme segments. To align with the previous studies where these segment types were implemented, only a relatively short portion of the reading passage was used (approx. 1 minute). Additionally, approximately the same number of segments of each type matched what was used in the previous studies (approx. 150 word-level vowels and 105 of the most frequent triphone vowels). Results comparing the three segmental conditions are displayed in Table 3.

Table 3: Accent recognition rates comparing Y-ACCDIST's performance, varying segmental type.

Segmental Unit Type	%Correct
Word-level vowels	74.2
Triphone vowels	75.0
Context-independent phonemes	76.7

Only marginal differences exist between the recognition rates for each segmental type. However, it appears that the context-independent phoneme is the preferred type here. For this particular classification task, aiming to eliminate coarticulation effects does not seem to improve performance. This might be down to the particular accents in question. [2], [3], [4] and [5] all used the Accents of the British Isles (ABI) corpus [1] which contains speakers from 13 locations spanning Britain. A more geographically-proximate task like this is expected to involve greater overlap between varieties. When using much larger ACCDIST matrices of the context-dependent segment types, it is likely that a greater proportion of elements may serve only to create 'noise' in the model.

3.2. Content-Mismatched Data

Using the same 120 speakers, a spontaneous speech sample (approx. 3 mins per speaker) was orthographically transcribed and processed in Y-ACCDIST. The classification task distinguishing between the four speaker groups is compared in Table 4 using content-controlled and content-mismatched data.

Table 4: Accent recognition rates for Y-ACCDIST's performance for content-controlled and content-mismatched data.

Content-controlled	Content-mismatched
86.7%	52.5%

The results show a substantial difference between the system's performance on the two data types. This is expected because different phonemic distributions exist in each speech sample in the case of the spontaneous speech data. It is expected that there are particular phonemes more useful to the task of accent recognition than others and it is likely that a number of occurrences are required to generate a stable average MFCC phoneme representation. When using content-mismatched data, then, it is likely that a larger quantity of data is required. [8] indicate this in their accent recognition study of French accents, which obtains similar accent recognition results for content-controlled and content-mismatched data, but use a substantially larger quantity of content-mismatched data (approximately 3 minutes of content-controlled data and 13 minutes of content-mismatched data).

4. SEGMENTAL SELECTION

Y-ACCDIST is highly dependent on phonemic classifications. Depending on the particular accents involved, some segments are expected to be of more value than others in distinguishing between varieties. Previous ACCDIST-based systems have solely focussed on vowel segments. Not all vowels, of course, carry distinctive weight when distinguishing between accents. Equally, some consonants may have distinctive power to offer. To demonstrate, the results in Table 5 compare Y-ACCDIST's accent recognition rates when /r/ is included in the ACCDIST matrix. /r/ was chosen for this particular task as Scottish English and English English varieties are involved. Rhoticity is argued to be a key distinguishing feature between these two major accent groups.

Table 5: Accent recognition rates comparing Y-ACCDIST's performance using content-controlled and content-mismatched data, including and excluding /r/.

	Content-controlled	Content-mismatched
- /r/	86.7%	52.5%
+ /r/	89.2%	59.3%

These results demonstrate the positive effects segmental selection can have on accent recognition performance. Research into discovering the optimum phoneme combination could boost recognition rates still further.

5. FURTHER DIRECTIONS

Given the above findings, a clear research direction would be to explore feature selection methods concerning the phonemes included in the representative ACCDIST matrices. It is possible to adopt and trial methods used in pattern recognition to automatically identify the most valuable phoneme segments in any given accent recognition task. In light of this, another potential direction is to focus on the effects of a speech sample's phonetic content on accent recognition. The results above show the positive effect a single phoneme can have on accent recognition. It is therefore likely that some speech samples will consist of richer phonetic content for accent recognition than others. Devising a set of criteria regarding the content of a speech sample could assist in some applications.

6. REFERENCES

[1] D'Arcy, A., Russell, M., Browning, S., Tomlinson, M. 2004. The accents of the British Isles. *Proc. of Modelisations pour l'Identification des Langues* Paris, France. 115–119.

[2] Hanani, A., Russell, M., Carey, M. 2011. Computer and human recognition of regional accents of British English. *Proc. Interspeech* Florence, Italy. 729–732.

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

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

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

[6] Vapnik, V. 1998. *Statistical Learning Theory*. New York: Wiley.

[7] Watt, D., Llamas, D., Johnson, D. 2014. Sociolinguistic variation on the Scottish-English border. In: Lawson, R., (ed), *Sociolinguistics in Scotland*. London: Palgrave Macmillan 79–102.

[8] Woehrling, C., de Mareuil, P., Adda-Decker, M. 2009. Linguistically-motivated automatic classification of regional French varieties. *Proc.of Interspeech* Brighton, UK. 2183–2186.

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