interspeech 2008 incorporating SST 08

Experiments with the ABI (Accents of the British Isles) Speech Corpus

Shona D'Arcy[1], *Martin Russell*[2]

[1]School of Electronic and Electrical Engineering, Trinity College, Dublin, Rep of Ireland.
[2]Department of Electronic, Electrical and Computer Engineering, University of Birmingham, UK
shona.darcy@tcd.ei, m.j.russell@bham.ac.uk

Abstract

The ABI (Accents of the British Isles) speech corpus contains approximately 90 hours of speech from approximately 280 speakers representing 14 different regional accents of British and Irish English. ABI includes a combination of applications-oriented and linguistically-motivated material. This paper describes experiments in which the ABI corpus is used to study the effects of these regional accents on vowel formant frequencies and automatic speech recogniser performance, and to explore inter-accent variability

Index Terms: accented speech, vowels, speech recognition

1. Introduction

Variability is the source of most of the difficulties faced by automatic speech recognition (ASR) systems. Even within a language each person speaks differently. Although this is due to many types of individual differences, the way we speak is also influenced by geographical and historical factors. The speech of an individual born in a particular area and who has lived there for a substantial part of his or her life is likely to be affected by the accent which is local to that area[1]. In the British Isles there is huge variation in the pronunciation of British English between different regions. The linguistics literature includes several studies of accent [1, 2, 3]. However, its implications for speech technology have received much less attention. In particular there has been no systematic study of the effects of British English accents on the performance of ASR systems. One explanation for this is the lack of suitable speech data.

In 2003 Aurix Limited sponsored the University of Birmingham to collect the Accents of the British Isles (ABI) speech corpus[2]. ABI contains recordings from approximately 20 subjects for each of 14 distinct accents of the British Isles (280 subjects in total). For one accent, Standard Southern English (SSE), subjects were selected by a phonetician. For all others a town or city was chosen and subjects were required to have been born in that location and lived there all of their lives (in many cases their parents were also born in that location). The towns (accents) are: Belfast (Northern Ireland (Ulster)), Birmingham (West Midlands), Burnley (Lancashire), Denbigh (Wales), Dublin (Republic of Ireland), Elgin (Scotland (Highlands)), Glasgow (Scotland (Central)), Hull (Yorkshire), Inner-London (London), Liverpool (Merseyside), Lowestoft (East Anglia), Newcastle (North-East), Truro (West Country).

Each subject contributed approximately 20 minutes of recordings, so that the whole corpus comprises over 90 hours of speech. The corpus is described in full in [4].

[1]'Accent' refers to characteristic differences in speech pronunciation, not the use of characteristic words or grammar, which is 'dialect'

[2]ABI is available from The Speech Ark Limited (http://www.thespeechark.com). Contact martin@thespeechark.com

This paper summarizes the results of three different experiments conducted on the ABI corpus. The first is an analysis of the effects of accent on vowel formant frequency based on sets of Consonant-Vowel-Consonant (CVC) syllables recorded by each subject. Of course, inter-accent differences involve much more than variations in vowel systems. However, these differences are evident to most native speakers and are well documented in the literature [1, 2, 3]. Although this analysis was performed for all of the accents in the corpus, due to space limitations this paper focuses on Standard Southern English (SSE) and the two Scottish accents, Glasgow and Elgin (Highlands).

The second experiment compares ASR performance for each of the accents using a recogniser trained on the British English WSJCAM0 corpus, with and without SSE or accent-specific MAP and/or MLLR adaptation. The third experiment explores the underlying acoustic correlates of accent variation. Each subject in the ABI corpus is represented as a 33 dimensional vector comprising the first three formant frequencies for each of the vowels from the CVC syllables. Principal Components Analysis (PCA) and Linear Discriminant Analysis (LDA) are applied to determine, in terms of vowel formant frequencies, the most significant dimensions of accent variation and for accent discrimination

2. CVC syllables in the ABI corpus

Each subject spoke 5 examples of each of 19 CVC syllables, the first and final syllables being /h/ and /d/. The 11 syllables with a monophthong are listed in table 1. The subject was prompted by the orthographic spelling of the syllable in the first column, supplemented, for nonsense words by a 'rhyming clue' (e.g: *hudd (to rhyme with bud)*). Subjects were not instructed how

Syllable	RP Transcription	To rhyme with	Wells set
had	h { d	-	trap
hard	h A d	-	palm
head	h e d	-	dress
heard	h 3 d	-	nurse
heed	h i d	-	fleece
hid	h I d	-	kit
hoard	h O d	-	thought
hod	h Q d	-	lot
hood	h U d	-	foot
hudd	h V d	bud	strut
who'd	h u d	-	goose

Table 1: *The 11 monophthong CVC syllables recorded by ABI subjects, with 'rhyming clues' where appropriate. The SAMPA phonemic transcriptions indicate the SSE pronunciation*

10.21437/Interspeech.2008-137

Accepted after peer review of full paper

to pronounce these syllables, either verbally or by a phonemic transcription. If an individual did not know how to pronounce a syllable the operator would indicate the rhyme clue. For example, in Lowestoft many of the subjects pronounced the diphthong syllable 'hured' as /h 3 d/, to rhyme with 'heard', since they normally pronounce 'cured', its rhyming clue, in this way.

The recordings were transcribed at the phoneme level using forced Viterbi alignment with a set of single state, multiple Gaussian mixture, monophone HMMs trained on the ABI corpus. Single rather than three state HMMs were used to avoid ambiguity in the location of the vowel centre in each syllable, variability being accommodated using the large number of mixture components.

The results of ESPS formant analysis were combined with the timing information to obtain vowel centre estimates of F1 and F2 for each syllable.

3. Vowel diagrams

The following figures show plots of F1 against F1-F2, estimated at vowel centres, for the 11 syllables *had, hard, head, heard, heed, hid, hoard, hod, hood, hudd, who'd* spoken by the 10 male subjects for the 'Standard Southern English' (SSE), Glasgow and Elgin (Highlands) accents. Similar figures for all of the ABI accents are presented in [5]. The choice of axes facilitates interpretation in terms of tongue position, with the vertical axis corresponding to the height of tongue placement in the mouth ('open' (top) versus 'close' (bottom)) and the horizontal axis corresponding to the placement of the highest point of the tongue ('front'(left) versus 'back' (right)).

3.1. Standard Southern English

Figure 1 is for Standard Southern English. Although there is some overlap (in particular for the vowels in *had, hudd* and *hard*) the figure shows 11 distinct clusters in the F1 - (F1-F2) plane, as one would predict.

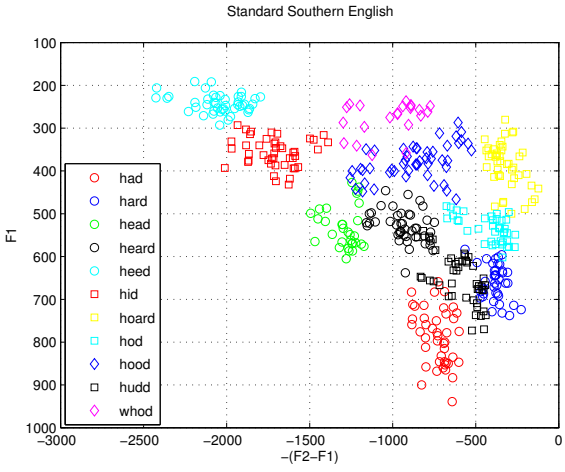

Figure 1: *Vowel plot for 'Standard Southern English' (SSE)*

3.2. Scotland

Figures 2 and 3 are the vowel diagrams for Glasgow and Elgin (Highlands) respectively, the two Scottish accents in the corpus. As one would expect, Glasgow and Highland speakers do not distinguish between *who'd* and *hood*, and this is evident in the diagrams. The formant values for the vowels in these two syllables overlap completely. In the Glasgow diagram the combined *who'd/hood* cluster is shifted towards the overlapping *hid* and *head* clusters. By contrast, in the 'Elgin' plot, although there is still no distinction between *who'd* and *hood*, and *hid* and *head* again overlap, there is much greater separation between these two clusters. Both figures show some overlap of the clusters corresponding to *had* and *hard*, though this is more pronounced in the Elgin (Highlands) plot.

Together with the vowel diagram corresponding to the other accents in the ABI corpus, these diagrams are consistent with the more subjective analyses in the accent literature. However, from the perspective of ASR they demonstrate that important acoustic differences occur within a given broad accent, such as 'Scottish English', and that accommodation of differences within such broad accent categories may require sophisticated phone-specific adaptation.

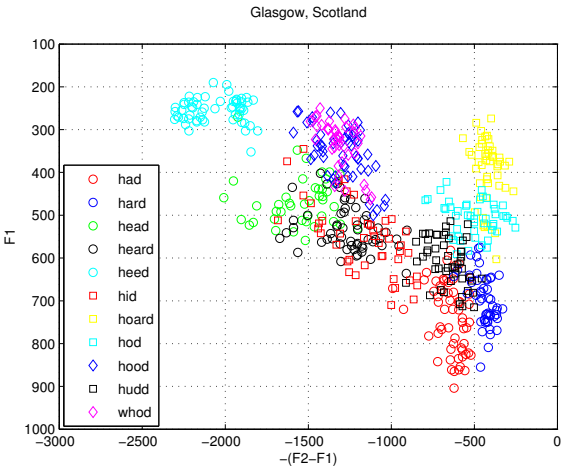

Figure 2: *Vowel plot for Glasgow*

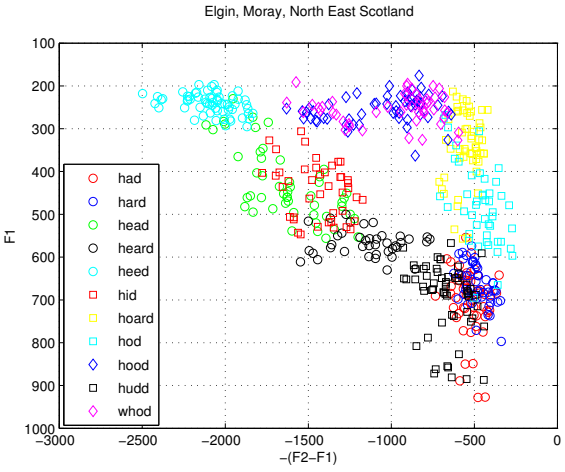

Figure 3: *Vowel plot for Elgin (Highlands)*

4. Speech Recognition Experiments

4.1. ASR System

In the nest set of experiments we investigate the effects of accent on ASR performance.

The ASR system used in these experiments is based on decision-tree clustered, tied-state, three-state hidden Markov models (HMMs) with 4-mixture component Gaussian observation distribution per state. Audio is parameterised using static and dynamic Mel-Frequency Cepstral Coefficients (MFCC) with Cepstral Mean Normalisation [6]. The system was developed using HTK [7] and the training subset of the Cambridge WSJCAM0 corpus of British English Speech (over 8000 utterances from 92 males and females aged between 18 and 40 reading from the Wall Street Text Corpus). A WER of 23.1% was achieved on the WSJCAM0 test set using these models.

4.2. The ABI test set

All of the speech from every speaker for each accent was available for testing. Each recording session in the ABI corpus resulted in 15 audio files containing a range of material including lists of digit quadruples, and phonetic letter quadruples, application- and phonetically-motivated phrases, SCRIBE [3] sentences, and a long passage of approximately fifteen sentences. These audio files were segmented according to their annotations to produce smaller, error free utterances. The annotation of the recordings of each speaker included non-speech, speech errors and insertions as well as silence, and these were used as break points indicating where to segment the audio file. A 800 word unigram language model was generated from the corpus. The final ABI test set contained approximately 8000 utterances from all 297 speakers.

4.3. Experiment Results

4.3.1. No Adaptation

The WSJCAM0 models were initially used without adaptation to establish baseline recognition results for each accent (figure 4, solid line). The error rates are very high compared with the performance on the WSJCAM0 test set, ranging from 64% to 92%, with an average WER of 78.7%. These high error rates compared with the performance on WSJCAM0 are likely to arise because the WSJCAM0 experiment utilises a more powerful language model, because of the accented speech and becasue of differences in the recording conditions between the two corpora. However, it is evident that those accents which are subjectively difficult, such as Glasgow and Belfast (Ulster), generally correspond to the highest error rates.

4.3.2. Adaptation

In the next sets of experiments the generic WSJCAM0 British English models were adapted to improve their performance on the ABI data using two standard adaptation techniques, namely Maximum Likelihood Linear Regression (MLLR) [8] and Maximum a Posteriori (MAP) adaptation [9]. Our study of vowel spaces demonstrates that accent differences include phone-specific, and not just global, shifts. Hence, given the modest amount of training data which is available, one might expect MAP to perform better than MLLR. To increase the amount of adaptation data a leave-one-out, adapt-on-all-the-rest approach was taken. For each accent, all files corresponding to one speaker were taken as the test set. The adaptation set is then the rest of the accent set (usually 19 other speakers). This process is repeated for each speaker and the average of all tests taken as the result for that accent group.

[3]The SCRIBE sentences are an anglicised version of the TIMIT sentences

4.3.3. Standard Southern English (SSE) adaptation

An important issue with mis-matched training and test data is whether an adaptation scheme is merely accommodating the recording characteristics of the test data. To address this issue the models were first adapted using just the entire SSE data set, using MAP and MLLR. These adapted models were then tested on the speech from the individual accents. The results are shown in Figure 4. As anticipated above, MAP adaptation yields the best results, with error rates between 43.5% and 73%, and an average of 56.8%. This is a 10% reduction in error rate relative to the results for MLLR. It is interesting that the ordering of the accents in figure 4, based on error rates for the unadapted models, remains approximately the same after SSE adaptation, suggesting that the accents of the WSJCAM0 speakers are close to SSE. Also, putting aside the Scottish and Irish accents, those accents with the highest error rates before and after SSE adaptation, namely Newcastle, Liverpool, Ulster, Lancashire and Yorkshire, correspond approximately to 'northern' accents, while the remainder correspond to 'southern' English.

4.3.4. Accent-specific adaptation

Next, using the same leave-one-out adaptation technique the original WSJCAM0 models were adapted individually to each accent in turn. There were either 19 or 20 speakers for each accent group, equating to approximately 5000 utterances for adaptation. The average WERs after accent specific adaptation are 57% and 45% for MLLR and MAP respectively. Figure 5 plots the recognition performance for each accent for MAP, MLLR and a combination of MAP and MLLR. The latter gives the best performance with an average error rate of 42.5%. Again, in general those accents which are subjectively most difficult correspond to the highest error rates. However, there are clear exceptions such as the relatively small error rate for Belfast (Ulster).

In conclusion, adapting to SSE data improves recognition performance, indicating that the extremely high baseline error rates are due in part to mismatch between the WSJCAM0 and ABI recording conditions (and in part to a weak language model), rather than just accent differences. However the benefit of accent specific adaptation can also be seen. This is observed in the lower error rates for accent specific adaptation, but also in the consistency of improvement achieved by this method. A significant improvement in WER for accents which the unadapted recogniser has particular difficulty recognising, for example Glasgow, is also seen.

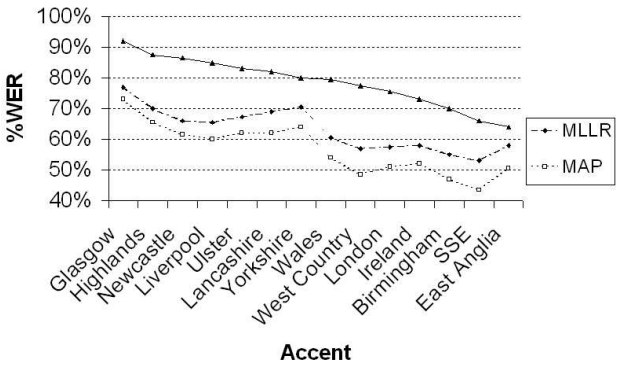

Figure 4: *Word error rates for the full test set for each of the accents in the ABI corpus after no adaptation, and MLLR and MAP adaptation using SSE data.*

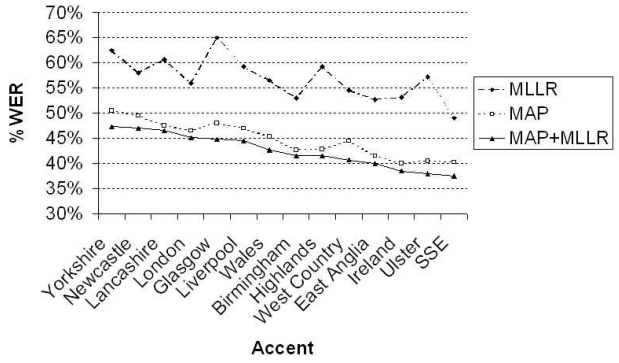

Figure 5: *Word error rates for the full test set for each of the accents in the ABI corpus after accent-dependent MLLR, MAP and MLLR-plus-MAP adaptation.*

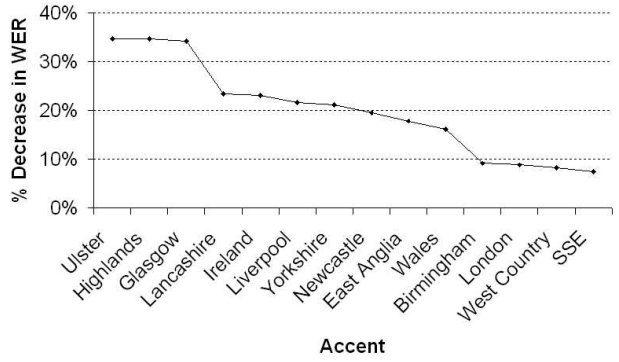

Figure 6: *Decrease in word error rate for accent-specific compared with SSE MAP adaptation. Accents are ordered according to the size of this decrease.*

Figure 6 shows the percentage decrease in word error rate for accent-dependent relative to SSE adaptation. This can be interpreted as a measure of the dissimilarity of a particular accent from SSE. Interestingly, the figure partitions the accents into three groups: (1) Ulster, Highlands and Glasgow, (2) Lancashire, Ireland, Liverpool, Yorkshire, Newcastle, East Anglia and Wales, and (3) Birmingham, London, West Country and SSE. Although the inclusion of Ulster with the two Scottish accents may seem anomalous from a geographic perspective, there are strong historic links, and these links are also evident in similarities between the Ulster and Glasgow vowel diagrams.

5. Analysis of accent variation

The purpose of our final experiment is to try to identify acoustic correlates of accent variation.

Each of the male subjects in the ABI corpus was represented as a vector in 33 dimensional space whose components are the average values of the first three formants for the vowels in the syllables from table 1. Figure 7 shows the first four eigenvectors which result from Principal Components Analysis (PCA) of this data. Inspection of the first eigenvector shows that the direction of maximum variation is dominated by a component which lies in the subspace spanned by F3 for vowels in the syllables hudd, hoard, hod, hid, heed, head, heard, hard, had. However, Linear Discriminant Analysis (LDA) indicates that this is due to inter-speaker rather than inter-accent variation, and that inter-accent variation is more closely aligned with variations in the values of F1 in several vowels, in particular those in who'd, hood, hudd and hod. More details of this analysis are presented in [5].

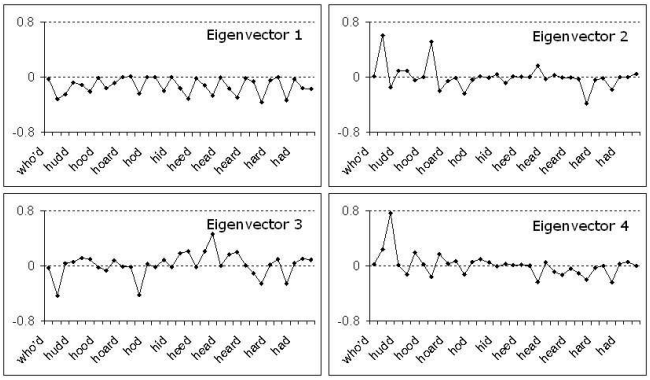

Figure 7: *First 4 principal components from 33 dimensional formant data.*

6. Conclusions

This papers describes aspects of some initial, data-driven experiments to investigate the implication of regional accents of British English for speech technology, and ASR in particular. Our study of vowel diagrams illustrates the extent of accent-related, acoustic variation even within a broad accent group (in this case 'Scottish English'). The study confirms that regional accents do present difficulties for ASR, and that in general those accents which native British English speakers would identify as 'difficult' are the same ones that cause problems for ASR. The final study has identified some of the acoustic correlates of accent variation in British English

7. References

[1] J. C. Wells, *Accents of English*, vol. 1 and 2, Cambridge University Press, 1982.

[2] A. Hughes, P. Trudgill, and D. Watt, *English Accents and Dialects*, Hodder Education, Great Britain, 4th edition, 2005.

[3] S. Elmes, *Talking for Britain: a Journey Through the Nation's Dialects*, Penguin Books, 2005.

[4] S.M. D'Arcy, M.J. Russell, S.R. Browning, and M.J. Tomlinson, "The accents of the british isles (abi) corpus," in *Proc. Modelisations pour l'Identification des Langues, MIDL, Paris*, 2004, pp. 115–119.

[5] S. M. D'Arcy, *The Effect of Age and Accent on Automatic Speech Recognition Performance*, Ph.D. thesis, University of Birmingham, UK, 2007.

[6] N. Cooke and M.J. Russell, "Using the focus of visual attention to improve spontaneous speech recognition," in *Proc. Interspeech 2005, Lisbon, Portugal*, 2005, pp. 1213–1216.

[7] S. J. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*, Entropic Camb. Res. Lab., Cambridge, UK, v2.1 edition, 1997.

[8] C.J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," vol. 9, no. 2, pp. 171–185, 1995.

[9] J-L. Gauvain and C. Lee, "Maximum a-posteriori estimation for multivariate gaussian mixture observations of markov chains," vol. 2, pp. 291–298, 1994.