Identification of British English regional accents using fusion of i-vector and multi-accent phonotactic systems

Maryam Najafian[1], *Saeid Safavi*[2], *Phil Weber*[3], *Martin Russell*[3]

[1]Center for Robust Speech Systems
University of Texas at Dallas
Richardson, TX, USA
m.najafian@utdallas.edu

[2]School of Engineering & Technology
University of Hertfordshire
Hertfordshire, UK
s.safavi@herts.ac.uk

[3]School of EESE
University of Birmingham
Birmingham, UK
[p.weber.1,m.j.russell]@bham.ac.uk

Abstract

The para-linguistic information in a speech signal includes clues to the geographical and social background of the speaker. This paper is concerned with recognition of the 14 regional accents of British English. For Accent Identification (AID), acoustic methods exploit differences between the distributions of sounds, while phonotactic approaches exploit the sequences in which these sounds occur. We demonstrate these methods are good complements for each other and use their confusion matrices for further analysis. Our relatively simple i-vector and phonotactic fused system with recognition accuracy of 84.87% outperforms the i-vector fused results reported in literature, by 4.7%. Further analysis on distribution of British English accents has been carried out by analyzing the low dimensional representation of i-vector AID feature space.

Index terms: Accent identification, I-vector, Phonotactic, British English regional accents

1. Introduction

The speech signal contains information beyond its linguistic content, including clues to the speaker's gender, age, regional accent, social background, or level of education [1, 2, 3]. In the first volume of Accents of English book [4], 'accent of English' is defined as "a pattern of pronunciation used by a speaker for whom English is the native language or, more generally, by the community or social grouping to which he or she belongs". Accent is different from dialect which also includes the use of words or phrases that are characteristic of those regions. Recent work on dialects includes varieties of English spoken as a first language in different countries (for example, US versus Australian English), geographical variations within a country [5, 6].

AID has recently emerged to be of substantial interest in the speech processing community. Recognising accent variations within a language is of importance for forensic speech scientists in speaker profiling and speaker comparison applications [7]. In addition, recognising speaker's accent can help in personalising synthetic speech of text-to-speech (TTS) systems.

Another important application of AID is in accent robust Automatic Speech Recognition (ASR) systems, where a pronunciation dictionary [8, 9, 10, 11] or an acoustic model [12, 13] can be selected based on the accent properties of the test speaker. Given a small amount of data from the test speaker, experiments on multi-accented ASR showed that using an AID based acoustic model selection will result in up to 50% Word Error Rate (WER) reduction which outperforms the supervised and unsupervised speaker adaptation alternatives in a Gaussian Mixture Model-Hidden Markov Model (GMM-HMM) baseline [12, 13].

Studies shown that for 'difficult' accents both the accuracy of ASR systems [14] and the perception of human listeners [15, 16, 17] deteriorates. Research has shown that difficult accents often lie near the extremes of the accent space and they are of high importance for training multi-accent Deep Neural Network (DNN)-HMM based ASR systems, because they expose the DNN training process to more diversity [14]. For example, Najafian [14] reported that addition of extra training material (up to 2.25 hours) from more difficult accents (e.g. Scottish accents) rather than easier alternatives (e.g. Southern English accents) leads to two times more relative WER reduction [14] across all different British English regional accents.

Much of the existing work on Accent Identification (AID) has been applied to Arabic [18, 19], English [14, 20, 21, 22, 23, 24, 25], and Chinese [26, 27], where each consists of a wide range of regional accent variations. Our work focuses on recognition of the 14 different regional accents of British English in the Accents of the British Isles (ABI) corpus [28].

After the success of acoustic and phonetic fused approaches in Language Identification (LID) performance [29], we decided to fuse the scores from i-vector based [30, 31] and phonotactic based [6] approaches in our AID system. We showed that our acoustic-phonotactic fused system outperforms that of the literature as it relies on both acoustic and phonotactic features to extract the accent-specific information. To obtain an insight on strengths and weaknesses of each AID method we analysed confusion matrices of our acoustic, phonotactic, and acoustic-phonotactic fused systems. Finally we used a two dimensional visualization map for the i-vector feature space to analyse the AID results and gain an understanding regarding the distribution of different regional accents in a low dimensional space.

2. Speech corpus description

As shown in Table 1 ABI corpus [28] represents data from 13 different regional accents, and standard Southern British English (sse). The sse speakers were selected by a phonetician. The ABI accents fall into 4 broad 'accent groups', namely Scottish (SC: shl, gla), Irish (IR: uls, roi), Northern English (NO: lan, ncl, lvp, brm, nwa, eyk) and Southern English (SO: sse, crn, ean, ilo). In ABI, regional accented speech was defined as speech of individuals who had lived in the region since birth. Each of 285 subjects read the same 20 prompt texts. The experiments in this paper focus on a subset of these texts, namely the 'short passages' (SPA, SPB and SPC), the 'short sentences' and the 'short phrases'. These are described below:

- SPA, SPB and SPC are short paragraphs, of lengths 92, 92 and 107 words, respectively, which together form the accent-diagnostic 'sailor passage' (*When a sailor in a* *small ship...*). The recordings have average durations 43.2s, 48.1s and 53.4s.

- ‘Short sentences’ are 20 phonetically balanced sentences (e.g.“Kangaroo Point overlooked the ocean”). They are a subset of the 200 Pre-Scribe B sentences (a version of the TIMIT sentences for British English), chosen to avoid some of the more ‘difficult’ of those sentences, whilst maintaining coverage (146 words, average duration 85.0s).
- ’Short phrases’ are 18 phonetically rich short phrases (e.g.“while we were away”) containing English phonemes in particular contexts in as condensed form as possible (58 words, average duration 34.5s).

Table 1: *Accents represented in the ABI corpus.*

code	Location	code	Location
sse	Standard Southern English	uls	Ulster
crn	Cornwall	lan	Lancashire
ean	East Anglia	ncl	Newcastle
ilo	Inner London	lvp	Liverpool
shl	Scottish Highlands	brm	Birmingham
gla	Glasgow	nwa	North Wales
roi	Republic of Ireland	eyk	East Yorkshire

In our AID experiments we used a 3-fold cross validation so no speaker appeared simultaneously in the training and test sets. We divided the ABI data in to three subsets; two with 95 and one with 94 speakers. Each time two subsets of the data were used for training and the remaining subset was used for testing. This procedure was repeated three times with different training and test sets. SPA utterances from each ABI speaker were only used for testing and not for training. For training the AID systems we used SPB, SPC, ‘short sentences’ and ‘short phrases’.

WSJCAM0 [32] is a corpus of British English speech recorded at Cambridge University. In this work, WSJCAM0 training set was used to train the phone recognizers.

3. Related work

Similar to LID, different approaches in AID can be partitioned into acoustic based methods, such as i-vectors [30, 31, 33], and phonotactic based methods, such as Phone Recognition followed by Language Modelling (PRLM) [6, 34], and Parallel PRLM (PPRLM) [35]. Often fusion of acoustic and phonetic approaches results in a higher accuracy than each individual system [29].

Recently, low-dimensional bottleneck neural network features have been successfully used to model speech dynamics [36]. For the LID task, the use of bottleneck features as well as DNN-posterior based approaches has become a successful alternative to the i-vector based systems [37, 38, 39, 40]. In fact these DNN based approaches have outperformed the i-vector approach for the LID task [37, 41]. Conclusions in one study suggests that for a DNN based accent recognition system [42] many more hours of training data than we have might be required.

Our work will be compared with the study carried out by DeMarco et al. [43]. For recognition of the 14 regional accents of British English in ABI corpus DeMarco et al. [43] proposed an i-vector fused AID classifier with accuracy of 81%, which comprises of 630 individual i-vector subsystems. DeMarco et al.’s work analyses a standard i-vector classifier under a range of projection methods, such as Linear Discriminant Analysis (LDA) [44], Semi-supervised Discriminant Analysis (SDA) [45], Neighbourhood Component Analysis (NCA) [46], and Regularized LDA (RLDA) [47]. Each of these projection methods extract different ‘aspects’ of the accents in question. These projection methods are also tried under different i-vector factor sizes (100, 150, 200, 250, 300, 350, 400) and GMM components (16, 128, 256, 512, 1024). Furthermore, a genetic algorithm is employed over all projection methods, all GMM component sizes and all factor dimensionalities, to produce a final selection of i-vector classifiers, termed ‘weak learners’ (fusing 630 from 2520 possible systems), which, when used with a majority voting classifiers, provides a further improved AID performance of 81%.

4. Proposed AID Systems

In AID, acoustic methods exploit differences between the distributions of sounds in different accents, and phonotactic approaches exploit the accent-dependent differences in the sequences in which these sounds occur [48]. An examples of accent-dependent differences can be found in the realisation of the vowel a in the words ‘cat’, ‘$a$fter’, and ‘f$a$ther’. In this example for Northern English speakers the realisation of vowel in words ‘cat’ and ‘$a$fter’ is more similar, while for Southern English speakers the realisation of vowel in words ‘$a$fter’, and ‘f$a$ther’ are more similar. We expect the AID system to capture the influence of these accent-specific patterns on the phone N-gram frequencies, acoustic quality and phonetic realisation of phonological units.

Our proposed AID system fuses the scores from a simple i-vector AID system and a multi-accent phonotactic AID system using Brummer’s multi-class linear logistic regression (LLR) toolkit [49]. Applying the fusion of two complementary AID systems could help with capturing more accent-specific patterns and achieving a more accurate AID system.

4.1. I-vector based AID

I-vectors provide a low-dimensional representation of feature vectors that can be successfully used for classification and recognition tasks. I-vectors were initially introduced for speaker recognition [50, 51], and after their success in this area, they were applied to language [52, 31] and accent recognition [22, 43].

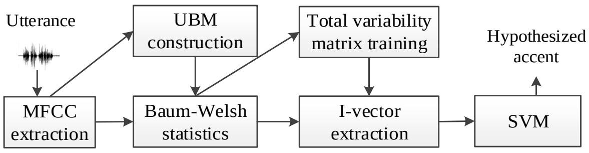

Figure 1: *I-vector extraction process*

The idea behind this text-independent (unsupervised) approach initiated from the Joint Factor Analysis (JFA) [30] technique proposed for speaker verification. In JFA speaker and channel variabilities are represented in two separate subspaces, whereas in the i-vector approach only one single space is defined for all types of variability (including both speaker and session variabilities). This is called the ‘total variability’ space. The motivation for the use of a single ‘total variability’ space is that in JFA traces of speaker dependent information can be found in the channel factors, and therefore separating speaker and channel variabilities will lead to losing some speaker dependent information [51]. The architecture of an i-vector system with a Support Sector Machine (SVM) [53] classifier is illustrated in Figure 1. Our i-vector based AID system is implemented in a similar manner to [51] using the ABI training subsets (includes SPB, SPC, shortphrase, and shortsentence), which results in 76.76% accuracy over SPA utterances (3-fold cross-validation). The process of building an i-vector system consists of the following stages.

Feature extraction: Our system is trained with 19 MFCCs plus 49 Shifted-Delta Cepstral coefficients (SDC) with a 7-3-1-7 configuration, giving a total of 68 features per frame [54]. Feature warping [55] is applied on the feature vectors for noise normalization.

UBM construction: Speech from ABI training subsets is used to estimate the parameters of a the Universal Background Model (UBM).

Baum-Welch statistics: The UBM trained in the previous stage can now be used for extracting the zero- and first-order Baum-Welch statistics centralised over the UBM mean.

Total variability modeling: Each utterance is described in terms of a speaker and channel dependent GMM mean supervector M, where $M = m + Tw$. Suppose that K is the number of Gaussian components in the UBM and F is the dimension of the acoustic feature vectors. The speaker and channel independent statistics, m, of dimension $KF \times 1$ is constructed by concatenating means for each Gaussian component of the UBM. The aim of the total variability modelling technique is to find a low rank rectangular ‘total variability matrix’ (T-matrix), T, of dimension $KF \times H$ with $H \ll KF$, and low dimensional ‘identity vector’, w, of dimension $H \times 1$ such that the probability of the training utterances given the model defined by the supervector M is maximised. Given each utterance from corresponding ABI speaker, the value of T-matrix and i-vector are estimated iteratively using the Expectation Maximization (EM) algorithm to maximize the likelihood of the training data. In the Expectation step, T is assumed to be known, and we update w. In the Maximization step, w is assumed to be known and we try to update T.

Extracting the i-vectors: For the utterance dependent mean offset, Tw, the components of the i-vector best describe the coordinates of the utterance in the reduced total variability space. Given the utterance, u, in the Expectation step, the i-vector w which is the mean of posterior distribution is updated using the current value of the T-matrix, and the Baum-Welch statistics extracted from the UBM. Presenting the utterances in the low-dimensional total variability space, ensures that for representing a new speaker only a small number of parameters need to be estimated. To achieve this the total variability space needs to encapsulate as much as possible of the supervectors in its restricted number of dimensions.

SVM: Given a set of labeled training utterances from 14 accent groups, our multi-class SVM classifier is a collection of 2-class SVMs with linear Kernel, which is trained using the corresponding accent-specific i-vectors. Then, the test speaker’s i-vector is scored against each accent-specific SVM (using a ‘one against all’ approach). The accent which gives the maximum score determines the accent of the test utterance.

In our system, the UBM was trained on the training subset of the ABI-1 corpus using various number of UBM components and T-matrix ranks (numbers of UBM components: 512, 1024 and T-matrix ranks: 200, 400, 800). The best accuracy on the evaluation set is achieved using the i-vector system that has a UBM with 512 Gaussian mixture components, and a total variability T-matrix of rank 800.

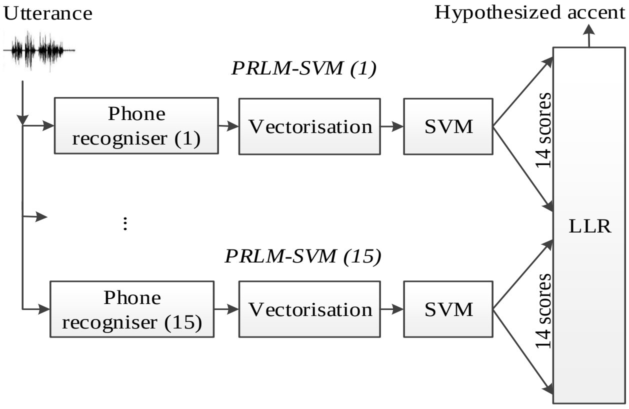

Figure 2: *Fusion of SVM scores from* 15 *parallel PRLM-SVMs*

4.2. Phonotactic based AID

In this section, we introduce the phonotactic approach to address the accent recognition problem. This approach recognises the accent of the speaker from a small amount of his or her speech, based on the frequency of use of certain phonetic sequences. Our phonotactic system comprises 15 PPRLM systems, a collection of 2-class SVM classifiers (multi-class SVM) and a LLR score fusion system [56, 21]. Our phonotactic based AID system is implemented using the ABI training subsets (includes SPB, SPC, shortphrase, and shortsentence), which results in 81% accuracy over SPA utterances (3-fold cross-validation).

The PPRLM process is carried out in four stages (Figure 2), namely phone recognition, vectorisation, SVM classification, and LLR score fusion.

Phone recognition: Each SPA utterance utterance from ABI is passed through 14 accent-specific phone recognisers and one standard Southern English recogniser trained on WSJCAM0 to generate a phone level transcription of the utterances. Each accent-specific phone recogniser is trained on WSJCAM0 training set and adapted to one accent from the ABI corpus using the Maximum Likelihood Linear Regression (MLLR) [57] approach. For adaptation the ABI training subset (includes SPB, SPC, shortphrase, and shortsentence) was used to create 14 phone recognisers with accent-specific acoustic models. All phone recognisers use 39 dimensional MFCCs and during the phone recognition process they use bigram triphone grammars built based on WSJCAM0 and ABI training subsets (3-fold cross-validation). A triphone dictionary with 8875 triphone entries from a phoneme set of size 44, is constructed using the WSJCAM0 and ABI training subsets and using the British English Example Pronunciation dictionary (BEEP) [58]. All phone HMMs have an 8 component GMM per state. The output of the phone recogniser is a sequence of phones from which an N-gram phone-language model is estimated.

Vectorisation: The sequence of phones corresponding to an utterance is represented as a D dimensional vector, whose i-th entry is the relative frequency of the i-th phone N-gram in the set and denoted by p_i (Equation 1).

$$p_i = \frac{Count(C_i)}{\sum_{j=1}^{D} Count(C_j)} \tag{1}$$

$Count(C_i)$ is the number of times the N-gram C_i occurs in the utterance. The outcome of this stage is a D-dimensional vector that represents N-gram frequencies per utterance, which is referred as the phonotactic system's supervector. Our best phonotactic AID result is obtained by applying the LLR fusion to the outputs of 15 individual phonotactic systems with 4-grams, using LLR (chosen empirically from 2-,3-,4-, and 5-grams on evaluation set). For our 4-gram language model supervectors are of dimension $D = 21,696$.

SVM: Given a set of labeled training utterances from 14 accent groups, there are in total 15 accent-specific PRLM systems trying to classify the test utterances into 14 classes. The test speaker's D dimensional supervector is then evaluated with a multi-class SVM containing a collection of two class SVMs to obtain a classification score (using a one against all approach). Each PRLM system produces 14 scores per test utterance. These scores determine to what extent each supervector belongs to each accent group.

Score fusion using LLR: To determine the test speaker's accent, the SVM scores generated by 14 parallel accent-specific PRLM systems and one general English PRLM system are fused using the LLR approach [59]. Figure 3 shows the process in which the scores from 15 PRLM-SVM systems are fused to recognise the test speaker's accent.

5. Experimental Results and Discussion

Our accent identification system fuses the scores from an i-vector and a multi-accent phonotactic based AID systems. As shown in Figure 3 this system consists of one i-vector system, and 15 phonotactic systems. DeMarco's i-vector acoustic system uses a combination of 630 subsystems with much higher T-matrix rank and the UBM size compared to our simple i-vector system [43]. Our system relies on complementary information exploited from the i-vector and phonotactic AID approaches, and it avoids genetic algorithm trials to obtain a quasi-optimal solution based on 630 learners, which reduces training time significantly. Our acoustic and phonotactic fused AID system outperforms the i-vector fused AID system proposed by DeMarco et al. [43], by 4.7%. Table 2 summarises the results for both AID systems.

Table 2: *Accuracy of coustic (Ac.) and phonotactic (Phon.) AID*

AID systems	Acoustic systems	Phonotactic systems	Final fused AID performance
DeMarco et. al. [43]	AID (630 Ac.) 81.05%	—	Fused (630 Ac.) 81.05%
Our proposed AID Systems	AID (1 Ac.) 76.76%	AID (15 Phon.) 80.65%	Fused (1 Ac. & 15 Phon.) 84.87%

5.1. Confusion matrices of AID systems

In this section we start by analysing the confusion matrices of phonotactic AID and i-vector AID systems. Then we present the confusion matrix for their fusion and describe how this fusion combines the strengths from both systems.

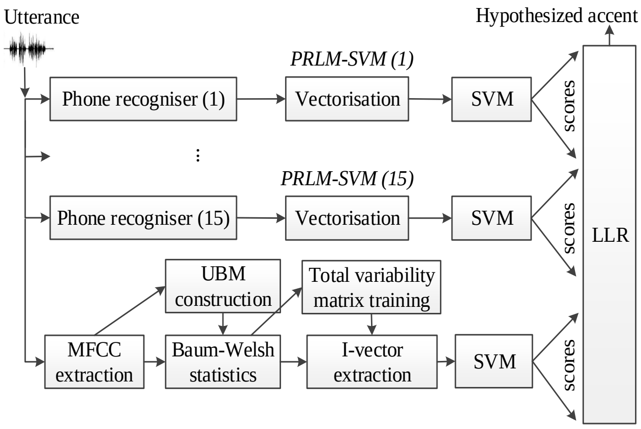

Figure 3: *AID with PRLM-SVM and i-vector system fusion*

Table 3 shows the confusion matrix corresponding to our multi-accent phonotactic AID system. This system assigns 81% of utterances to the correct regional accent and over 92% of utterances to the correct broad accent group.

Table 3: *Confusion matrix for the multi-accent phonotactic system*

Accent code	Accent group	Acc.	brm	eyk	lan	lvp	ncl	nwa	ilo	sse	ean	crn	roi	uls	shl	gla
brm	Northern	75%	15	0	0	0	0	0	1	1	2	0	1	0	0	0
eyk		76%	0	19	2	0	1	1	0	1	0	1	0	0	0	0
lan		76%	0	3	16	0	0	1	0	0	0	0	0	0	0	1
lvp		85%	0	1	0	17	0	1	1	0	0	0	0	0	0	0
ncl		90%	1	0	0	1	18	0	0	0	0	0	0	0	0	0
nwa		81%	0	0	0	1	1	17	0	1	1	0	0	0	0	0
ilo	Southern	71%	2	0	0	0	0	0	15	0	3	0	1	0	0	0
sse		69%	0	0	0	0	0	2	0	11	2	1	0	0	0	0
ean		68%	0	0	0	0	0	0	0	6	13	0	0	0	0	0
crn		85%	0	0	0	0	0	0	1	1	0	17	1	0	0	0
roi	Irish	84%	0	0	0	0	0	1	0	0	0	0	16	2	0	0
uls		90%	0	0	0	0	0	0	0	0	0	0	2	18	0	0
shl	Scottish	86%	0	0	0	0	0	0	0	1	0	0	1	0	19	1
gla		90%	0	0	0	0	0	0	0	0	0	0	0	0	2	18

The highest accuracy of 90% belongs to gla (Glaswegian), uls (Ulster), and ncl (Newcastle) accents which could be due to the fact that these accents are very different from other accents in terms of their phonotactic and linguistic properties. Some of the recognition errors in this table can be justified by linguistic, geographical or historical explanations.

There are a number of regional accents within the same broad accent group that are mis-classified as each other. For instance, East Anglia (ean) is geographically in the South East, so in a sense it's part of Southern English, and this may account for the 6 ean (East Anglia) accent speakers being classified as sse, and the two speakers of sse being classified as ean. In fact, the lowest accuracy of 68% belongs to the ean accent. Among the Northern English accents the geographical proximity of lan (Lancashire), and eyk (East Yorkshire) may account for the three lan accent speakers being classified as eyk, and the two speakers of lan being classified as eyk.

Focusing on the errors which fall outside the accent groups,

Table 4: *Confusion matrix for the proposed i-vector system*

Accent code	Accent group	Acc.	brm	eyk	lan	lvp	ncl	nwa	ilo	sse	ean	crn	roi	uls	shl	gla
brm	*Northern*	80%	16	0	0	0	0	1	0	1	1	1	0	0	0	0
eyk		84%	1	21	2	0	1	0	0	0	0	0	0	0	0	0
lan		76%	1	0	16	0	1	1	1	0	1	0	0	0	0	0
lvp		85%	0	0	1	17	0	2	0	0	0	0	0	0	0	0
ncl		65%	0	0	2	1	13	0	0	0	1	0	0	0	2	1
nwa		52%	1	4	1	0	1	11	0	0	0	2	0	0	1	0
ilo	*Southern*	57%	2	1	3	0	0	0	12	0	2	0	0	0	0	1
sse		69%	0	2	0	0	0	1	2	11	0	0	0	0	0	0
ean		84%	1	1	0	0	0	0	1	0	16	0	0	0	0	0
crn		55%	0	1	0	0	1	1	3	1	1	11	0	0	1	0
roi	*Irish*	78%	0	0	0	0	0	0	0	0	0	0	15	4	0	0
uls		90%	0	0	0	0	0	0	0	0	0	0	2	18	0	0
shl	*Scottish*	100%	0	0	0	0	0	0	0	0	0	0	0	0	22	0
gla		95%	0	0	0	0	1	0	0	0	0	0	0	0	0	19

Table 5: *Acoustic-phonotactic fused AID confusion matrix*

Accent code	Accent group	Acc.	brm	eyk	lan	lvp	ncl	nwa	ilo	sse	ean	crn	roi	uls	shl	gla
brm	*Northern*	80%	16	0	0	0	0	0	0	2	2	0	0	0	0	0
eyk		92%	0	22	1	0	0	0	0	1	0	1	0	0	0	0
lan		95%	0	2	19	0	0	0	0	0	0	0	0	0	0	0
lvp		85%	0	1	0	17	0	1	1	0	0	0	0	0	0	0
ncl		85%	0	1	0	1	17	1	0	0	0	0	0	0	0	1
nwa		71%	1	0	0	2	1	15	0	0	1	0	0	0	0	0
ilo	*Southern*	77%	1	1	0	1	0	0	17	1	2	0	0	0	0	0
sse		75%	0	0	1	0	0	0	1	12	1	0	0	0	0	0
ean		75%	0	1	0	0	0	0	0	4	15	0	0	0	0	0
crn		85%	0	1	0	0	0	0	3	0	0	17	0	0	0	0
roi	*Irish*	84%	0	0	0	0	1	0	0	0	0	0	16	2	0	0
uls		90%	0	0	0	0	0	0	0	0	0	0	2	18	0	0
shl	*Scottish*	95%	0	0	0	0	0	0	0	1	0	0	0	0	21	0
gla		95%	0	0	0	0	0	0	0	0	0	0	0	0	1	19

the most interesting case is Birmingham, where 29% of speakers are assigned to Southern English accents. However, Birmingham is on the boundary between the linguistic Northern and Southern English regional accents, and this may account for the result.

Table 4 shows the confusion matrix corresponding to our i-vector system. The i-vector system assigns 76.76% of utterances to the correct accent region and 89.9% to the correct broad accent group.

The i-vector system achieves a slightly better accuracy on the Scottish accents (gla, shl). However, for the phonotactic system only 10% of speech from Southern English group is classified out of group, but this rises to 27% for the i-vector system. A considerable number of Southern English regional accents are recognised as Northern English regional accents. This might be due to the fact that the sequential information captured by the phonotactic system is so much better than the static spectral information captured by the i-vector system for some accents. Phonotactics approach relies on accent-dependent differences in the sequences in which different sounds occur and that might be the reason that the phonotactic approach has managed to provide a higher accuracy in distinguishing between Northern and Southern English accents compared to the i-vector approach that only relies on the distribution of sounds in an utterance.

The nwa (North Wales) accent is recorded at Denbigh, where the recordings were made, is close to lvp (Liverpool) and this may account for the 2 lvp accent speakers being classified as nwa. Also a large number of people from neighbouring areas are traveling and retiring in Wales which might have affected the locals accent, and this might be the reason that a number of speakers with nwa (North Wales) accent are being mis-classified as other Northern English accents. For nwa and crn accents recognition accuracy falls by approximately 35% when the i-vector method is applied rather than the phonotactic method, and this leaves them as the most mis-recognised accents by the i-vector AID system.

Table 5 shows the confusion matrix corresponding to fusion of the i-vector and phonotactic based AID systems. This acoustic and phonotactic fused AID system assigns 84.87% of utterances to the correct accent region and 94% to the correct broad accent group.

This fusion had a very positive impact on recognition of nwa and crn accents as they were previously largely misclassified as other neighbouring accents by the i-vector system. The best identification rate of 95% is achieved for lan, gla and shl due to possibly their linguistically more distinguishable accent-specific pattern. The lowest identification rate of 71% belongs to the nwa and it can be seen that speakers from this region are mainly mis-recognised as speakers with other Northern English accent, possibly due to the influence of neighbouring travelers to this area.

5.2. Visualisation of the AID feature space

For the purposes of visualisation the i-vectors from our proposed i-vector system are projected onto 666 and 2 dimensions using the EM algorithm for Principal Components Analysis (EM-PCA) [60] and LDA algorithms respectively. For each accent region, 0.7-standard-deviation contours from the mean value represent utterances of that accent in the accent space.

Table 6: *ABI regions and corresponding broad accent groups*

Northern English	Southern English	Irish	Scottish
brm eyk lan lvp ncl nwa	ilo sse ean crn	roi uls	shl gla

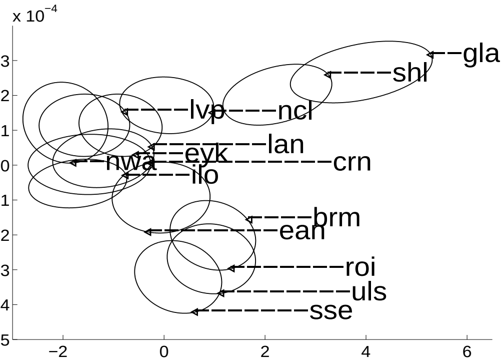

Figure 4: *Visualization of the i-vector feature space*

Table 6 summarises different accent regions and their corresponding broad accent groups and Figure 4 represents the distribution of different accent groups for the ABI corpus in a two dimensional AID i-vector space.

In the top right, a cluster can be seen for the two Scottish accents gla and shl and these two accents have the biggest relative distance from the rest of accent groups. The clusters for the two Irish accents roi and uls are situated on the bottom left side of the visualisation map with large overlap as expected. Looking at the confusion matrix (Table 4) and the visualisation map for the i-vector system (Figure 4) suggests that these features are quite successful in separating Irish and Scottish accents. Interestingly, even in two dimensions the major accent properties of the i-vectors can be observed. However, due to the similarity between the two Irish accents which is also evident from the confusion matrix, four out of five times the mis-recognition of these accents was due to one Irish accent being mis-recognised as the other Irish accent.

Southern English accents are scattered around the centre of the figure and overlap with Northern and Irish accent groups. One reason might be the fact that in each accent region we expect to see members who exhibit hints of Southern accent in their speech which could be influenced by social factors. Looking at the visualisation map for the i-vector system suggests that these features are quite successful in separating different regional accents, and even in two dimensions the major accent properties can be observed.

Three accent clusters corresponding to Northern, Irish and Scottish accents are present in this visualisation map, however, there is no separate cluster for the Southern English accents. Both the i-vector's confusion matrix and visualisation map suggest that the i-vector features are not very strong in capturing accent-specific differences between Northern and Southern English accents. The social or educational factors for some of the Northern speakers could be the reason for them being incorrectly identified as Southern accents.

6. Summary and future work

We shown that fusing two complementary AID systems, namely i-vector and phonotactic results in a higher accuracy (84.87%) compared with DeMarco et al.'s i-vector fused system (81%). This work resulted in a more computationally efficient system which is more suitable for real-time applications.

Looking at the confusion matrix of i-vector system, shows these features are quite successful in recognising Irish and Scottish accents, and not very strong in recognising Northern and Southern accents. Fusing the result of this system with that of the phonotactic system has significantly improved the AID accuracy.

The results shown in the confusion matrix of i-vector confirm the conclusion made from the two dimensional visualisation plot of the i-vector feature space. Both the two dimensional visualisation and the confusion matrix can be used to analyse similarities and differences among different regional accents. Direct benefit of these analyses can be seen in the ASR systems when there is a mis-match between the accent properties of the training and test data [14]. In such cases, one could determine difficult accents and add them as an extra training material to the training set. Another successful approach is to use the knowledge regarding the user's accent for the accent specific acoustic model [12, 13] or pronunciation dictionary selection [9, 8].

The next stage of this research will focus on the use of bottleneck features as well as the DNN-posterior based approaches for the AID task. Given our limited accented resources it is interesting to see whether use of DNN based features can outperform our current system.

7. References

[1] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. MüLler, and S. Narayanan, "Paralinguistics in speech and language state-of-the-art and the challenge," *CSL*, vol. 27, no. 1, pp. 4–39, 2013.

[2] S. Safavi, M. J. Russell, and P. Jancovic, "Identification of age-group from children's speech by computers and humans." in *INTERSPEECH*, 2014, pp. 243–247.

[3] S. Safavi, A. Hanani, M. Russell, P. Jancovic, and M. J. Carey, "Contrasting the effects of different frequency bands on speaker and accent identification," *Signal Processing Letters, IEEE*, vol. 19, no. 12, pp. 829–832, 2012.

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

[5] J. H. Hansen and G. Liu, "Unsupervised accent classification for deep data fusion of acoustic and language information," *Speech Communication*, 2016.

[6] M. Zissman, T. P. Gleason, D. M. Rekart, B. L. Losiewicz *et al.*, "Automatic dialect identification of extemporaneous conversational, Latin American Spanish speech," in *ICASSP*, vol. 2. IEEE, 1996, pp. 777–780.

[7] G. Brown and J. Wormald, "Speaker profiling: An automatic method?" in *IAFPA*, vol. 31, 2014.

[8] J. J. Humphries and P. C. Woodland, "Using accent-specific pronunciation modelling for improved large vocabulary continuous speech recognition." in *EUROSPEECH*, 1997.

[9] M. Tjalve and M. Huckvale, "Pronunciation variation modelling using accent features," in *INTERSPEECH*, 2005, pp. 1341–1344.

[10] Y. Liu and P. Fung, "Multi-accent Chinese speech recognition," in *INTERSPEECH*, vol. 1, 2006, p. 133.

[11] M. Liu, B. Xu, T. Hunng, Y. Deng, and C. Li, "Mandarin accent adaptation based on context-independent/context-dependent pronunciation modeling," in *ICASSP*, vol. 2. IEEE, 2000, pp. II1025–II1028.

[12] M. Najafian, A. DeMarco, S. Cox, and M. Russell, "Unsupervised model selection for recognition of regional accented speech," *INTERSPEECH*, 2014.

[13] M. Najafian, S. Safavi, A. Hanani, and M. Russell, "Acoustic model selection for recognition if regional accented speech," *EUSIPCO*, 2014.

[14] M. Najafian, "Acoustic model selection for recognition of regional accented speech," Ph.D. dissertation, University of Birmingham, 2016.

[15] A. Hanani, M. Russell, and M. J. Carey, "Computer and human recognition of regional accents of british english," in *ISCA*, 2011.

[16] M. A. Kadam, A. J. Orena, R. M. Theodore, and L. Polka, "Reading ability influences native and non-native voice recognition, even for unimpaired readers," *JASA*, vol. 139, no. 1, pp. EL6–EL12, 2016.

[17] M. M. Faris, C. T. Best, and M. D. Tyler, "An examination of the different ways that non-native phones may be perceptually assimilated as uncategorized," *JASA*, vol. 139, no. 1, pp. EL1–EL5, 2016.

[18] A. Hanani, H. Basha, Y. Sharaf, and S. Taylor, "Palestinian Arabic regional accent recognition," in *SPeD*. IEEE, 2015, pp. 1–6.

[19] F. Biadsy, "Automatic dialect and accent recognition and its application to speech recognition," Ph.D. dissertation, Columbia University, 2011.

[20] P. Angkititrakul and J. H. Hansen, "Advances in phone-based modeling for automatic accent classification," *T-ASLP*, vol. 14, no. 2, pp. 634–646, 2006.

[21] A. Hanani, M. Russell, and M. Carey, "Human and computer recognition of regional accents and ethnic groups from British English speech," *ICSLP*, vol. 27, no. 1, pp. 59–74, 2013.

[22] M. H. Bahari, R. Saeidi, D. Van Leeuwen *et al.*, "Accent recognition using i-vector, Gaussian mean supervector and Gaussian posterior probability supervector for spontaneous telephone speech," in *ICASSP*. IEEE, 2013, pp. 7344–7348.

[23] M. Huckvale, "ACCDIST: a metric for comparing speakers' accents," *ICSLP*, 2004.

[24] N. F. Chen, W. Shen, J. P. Campbell, and P. A. Torres-Carrasquillo, "Informative dialect recognition using context-dependent pronunciation modeling," in *ICASSP*. IEEE, 2011, pp. 4396–4399.

[25] N. F. Chen, S. W. Tam, W. Shen, and J. P. Campbell, "Characterizing phonetic transformations and acoustic differences across English dialects," *ASLP*, vol. 22, no. 1, pp. 110–124, 2014.

[26] J. Hou, Y. Liu, T. F. Zheng, J. Olsen, and J. Tian, "Multi-layered features with SVM for Chinese accent identification," in *Audio Language and Image Processing (ICALIP), 2010 International Conference on*. IEEE, 2010, pp. 25–30.

[27] Q. Zhang, H. Boril, and J. H. Hansen, "Supervector pre-processing for PRSVM-based Chinese and Arabic dialect identification," in *ICASSP*. IEEE, 2013, pp. 7363–7367.

[28] S. M. D'Arcy, M. J. Russell, S. R. Browning, and M. J. Tomlinson, "The Accents of the British Isles (ABI) corpus," *Proceedings Modélisations pour lIdentification des Langues*, pp. 115–119, 2004.

[29] W. Campbell, T. Gleason, J. Navratil, D. Reynolds, W. Shen, E. Singer, and P. Torres-Carrasquillo, "Advanced language recognition using cepstra and phonotactics: MITLL system performance on the NIST 2005 language recognition evaluation," in *ODYSSEY*. IEEE, 2006, pp. 1–8.

[30] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," *CRIM*, 2005.

[31] N. Dehak, P. A. Torres-Carrasquillo, D. A. Reynolds, and R. Dehak, "Language recognition via i-vectors and dimensionality reduction," in *INTERSPEECH*, 2011, pp. 857–860.

[32] J. Fransen, D. Pye, T. Robinson, P. Woodland, and S. Young, "WSJCAM0 corpus and recording description," Tech. Rep., 1994.

[33] Q. Zhang, G. Liu, and J. H. Hansen, "Robust language recognition based on diverse features," in *ODYSSEY*, 2014, pp. 152–157.

[34] M. Zissman, E. Singer *et al.*, "Automatic language identification of telephone speech messages using phoneme recognition and n-gram modeling," in *ICASSP*, vol. 1. IEEE, 1994, pp. I–305.

[35] W. Shen, W. Campbell, T. Gleason, D. Reynolds, and E. Singer, "Experiments with lattice-based PPRLM language identification," in *ODYSSEY*. IEEE, 2006, pp. 1–6.

[36] L. Bai, P. Jancovic, M. Russell, and P. Weber, "Analysis of a low-dimensional bottleneck neural network representation of speech for modelling speech dynamics," *INTERSPEECH*, 2015.

[37] F. Richardson, D. Reynolds, and N. Dehak, "A unified deep neural network for speaker and language recognition," *INTERSPEECH*, 2015.

[38] Y. Zhang, E. Chuangsuwanich, and J. Glass, "Language ID-based training of multilingual stacked bottleneck features," in *INTERSPEECH*, 2014, pp. 1–5.

[39] B. Jiang, Y. Song, and S. Wei, "Task-aware deep bottleneck features for spoken language identification," *INTERSPEECH*, vol. 9, no. 7, 2014.

[40] P. Matejka, L. Zhang, T. Ng, H. Mallidi, O. Glembek, J. Ma, and B. Zhang, "Neural network bottleneck features for language identification," *ODYSSEY*, pp. 299–304, 2014.

[41] F. Grezl, E. Egorova, and M. Karafiát, "Further investigation into multilingual training and adaptation of stacked bottle-neck neural network structure," in *SLT*. IEEE, 2014, pp. 48–53.

[42] I. Lopez-Moreno, J. Gonzalez-Dominguez, O. Plchot, D. Martinez, J. Gonzalez-Rodriguez, and P. Moreno, "Automatic language identification using deep neural networks," in *ICASSP*. IEEE, 2014, pp. 5337–5341.

[43] A. DeMarco and S. J. Cox, "Native accent classification via i-vectors and speaker compensation fusion," in *INTERSPEECH*, 2013, pp. 1472–1476.

[44] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[45] D. Cai, X. He, and J. Han, "Semi-supervised discriminant analysis," in *ICCV*, 2007.

[46] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov, "Neighbourhood components analysis," in *NIPS*. MIT Press, 2004, pp. 513–520.

[47] J. H. Friedman, "Regularized Discriminant Analysis," *JASA*, vol. 84, no. 405, pp. 165–175, 1989.

[48] M. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *ICASSP*, vol. 4, no. 1, pp. 31–, 1996.

[49] N. Brummer and D. A. van Leeuwen, "On calibration of language recognition scores," in *ODYSSEY*. IEEE, 2006, pp. 1–8.

[50] M. Senoussaoui, P. Kenny, N. Dehak, and P. Dumouchel, "An i-vector extractor suitable for speaker recognition with both microphone and telephone speech." in *ODYSSEY*, 2010, p. 6.

[51] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *INTERSPEECH*, vol. 19, no. 4, pp. 788–798, 2011.

[52] M. Soufifar, M. Kockmann, L. Burget, O. Plchot, O. Glembek, and T. Svendsen, "I-vector approach to phonotactic language recognition." in *INTERSPEECH*, 2011, pp. 2913–2916.

[53] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[54] B. Bielefeld, "Language identification using shifted delta cepstrum," in *Fourteenth Annual Speech Research Symposium*, 1994.

[55] J. W. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *ODYSSEY*, 2001, pp. 213–218.

[56] F. S. Richardson, W. M. Campbell, and P. A. Torres-Carrasquillo, "Discriminative n-gram selection for dialect recognition." in *INTERSPEECH*, 2009, pp. 192–195.

[57] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *JSTOR*, pp. 1–38, 1977.

[58] A. Robinson, "The British English Example Pronunciation (BEEP) dictionary," 1996.

[59] L.-F. Zhai, M.-H. Siu, X. Yang, and H. Gish, "Discriminatively trained language models using support vector machines for language identification," in *ODYSSEY*, 2006, pp. 1–6.

[60] S. Roweis, "EM algorithms for PCA and SPCA," *NIPS*, pp. 626–632, 1998.