Automatic accent identification as an analytical tool for accent robust automatic speech recognition

Najafian, Maryam; Russell, Martin

DOI:
10.1016/j.specom.2020.05.003

Document Version
Peer reviewed version

Citation for published version (Harvard):
Najafian, M & Russell, M 2020, 'Automatic accent identification as an analytical tool for accent robust automatic speech recognition', *Speech Communication*, vol. 122, pp. 44-55. https://doi.org/10.1016/j.specom.2020.05.003

Automatic accent identification as an analytical tool for accent robust automatic speech recognition

Maryam Najafian[1,2], Martin Russell[3]

School of Engineering, University of Birmingham, Birmingham, UK[1]
Computer Science & Artificial Intelligence Lab, Massachusetts Institute of Technology, Cambridge, MA, USA [2]
School of Computer Science, University of Birmingham, Birmingham, UK [3]

najafian@csail.mit.edu, m.j.russell@bham.ac.uk

Abstract

We present a novel study of relationships between automatic accent identification (AID) and accent-robust automatic speech recognition (ASR), using i-vector based AID and deep neural network, hidden Markov Model (DNN-HMM) based ASR. A visualization of the AID i-vector space and a novel analysis of the accent content of the WSJCAM0 corpus are presented. Accents that occur at the periphery of AID space are referred to as "extreme". We demonstrate a negative correlation, with respect to accent, between AID and ASR accuracy, where extreme accents exhibit the highest AID and lowest ASR performance. These relationships between accents inform a set of ASR experiments in which a generic training set (WSJCAM0) is supplemented with a fixed amount of accented data from the ABI (Accents of the British Isles) corpus. The best performance across all accents, a 32% relative reduction in errors compared with the baseline ASR system, is obtained when the supplementary data comprises extreme accented speech, even though this accent accounts for just 14% of the test data. We conclude that i-vector based AID analysis provides a principled approach to the selection of training material for accent robust ASR. We speculate that this may generalize to other detection technologies and other types of variability, such as Speaker Identification (SI) and speaker variability.

Keywords: Speech recognition, accent identification, British accents, i-vector

1. Introduction

Advances in acoustic modelling combining Deep Neural Networks (DNNs) and Hidden Markov Models (HMMs) have led to significant performance improvements for Automatic Speech Recognition (ASR) [1]. The DNN enables discriminative training of HMM state posterior probabilities for a given input feature vector, replacing the Gaussian Mixture Models (GMMs) that have represented the state-of-the art since the late 1980s.

However, research [2, 3] and evidence in news media [4, 5] indicate that robust recognition of accented speech with limited resources is still an important research challenge. Accents are a primary source of speech variation [6, 7]. They are characterised by systematic changes to the realization of particular phones, and often correspond to social, educational or geographical factors. Accents corresponding to the latter are referred to as regional accents. Regional accents of the British Isles have been studied extensively (for example [8, 9]) and are a problem for ASR [10, 11]. British English accents can be divided into five broad groups, corresponding to Southern and Northern England, Ireland, Scotland and Wales, each of which can be sub-divided. For example, the accent of a woman who was born in Hull and has lived there all of her life would be categorized as northern English by most native English speakers of British English. Her vowels in words like "bath" and "strutt" are the same as in "cat", and "hood", respectively [8]. Someone from the north of England might be able to place her accent in the east of the region and, because she does not have the "geordie" accent of the area around Newcastle, in the south of the north-east. Some Hull residents would even hear that she is from the west of the city. In the lowest layer of this hierarchical description of accent she has her own unique "ideolect", influenced by physiological, social, educational and other factors.

To achieve the best ASR performance, DNN-HMM systems require large training corpora that represent potential variations in the test material. Although publicly available corpora exist for British English, for example WSJ-

CAM0 [12], only the Accents of the British Isles (ABI) corpus [13] contains recordings that are explicitly representative of regional accents. Although ABI is relatively large, with speech from 285 subjects representing 14 different accents (13 regional accents plus Standard Southern English), the amount of data per accent is limited. For example, there are just 22 subjects in the ABI-1 corpus with the same regional accent as our speaker from Hull, and 40 and 96 subjects with accents from the North-East and North of England, respectively. Thus the hierarchical description of accent in the previous paragraph reveals a familiar trade-off for acoustic modelling. A model conditioned on an accent class lower in the hierarchy will be more specific to our subject's speech, but less data will be available to train it. The challenge is to identify accent classes that are sufficiently low in the hierarchy to reduce variability, but at the same time contain sufficient data to support robust modelling. The correct model can then be selected using automatic accent identification (AID) (for example, [7, 14]). A standard alternative to model selection is to construct a single 'multiple accent' acoustic model by including all data from all accents in the training set.

The premise of this paper is that AID provides an analysis tool for a more pricipled approach to the selection of training material for accent robust ASR. AID can be used to analyse the diversity of accents in a training set and hence identify accent groups that are not represented. Furthermore, modern approaches to AID typically represent an utterance as an i-vector [15] or supervector in a high dimensional vector space. Visualisation of this space, via a two-dimensional projection, can indicate the acoustic relationships between different accents, identify accent groups at the periphery of the space, and suggests ways in which the training data could be supplemented to improve ASR performance. The paper also investigates the relationship between AID accuracy for a particular accent and ASR accuracy for the corresponding accented speech.

In this paper these ideas are applied to the WSJCAM0 and ABI corpora. We begin by visualising the AID i-vector space and the locations of WSJCAM0 and the different accents of the ABI corpus in that space. It emerges that the groupings of accent data in this space correspond, approximately, to natural

broad geographical accent groups. The term "extreme" accents is used to refer to accents that are located at the periphery of the AID accent space. Next, AID is applied to the utterances in WSJCAM0 to determine the distribution of the regional accents of its subjects, and to the ABI corpus. This analysis indicates that approximately 81% of the subjects in WSJCAM0 speak with a northern or southern English accent, with the remainder categorised as Scottish or Irish.

The second part of the paper is concerned with the application of these ideas to accent-robust ASR. A negative correlation is shown between AID accuracy and ASR accuracy for a baseline DNN-HMM ASR system trainied on WSJCAM0. Intuitively, if the AID system achieves a high accuracy for a particular accent, then it is likely that this accent is separated from the other accents in the AID i-vector space. Hence it is also likely to be separated from the ASR training data, so that ASR performance is poor. Thus, the Scottish accents **gla** and **shl**, which according to the visualisation of the AID i-vector space are extreme, have the lowest AID error rates (5% and 0%, respectively), but the highest ASR word error rates (WERs) (13.3% and 11.5%, respectively). In the remainder of the paper, the relationships between the accents in the ABI corpus inform a set of ASR experiments in which the generic training set (WSJCAM0) is supplemented with accented data from the ABI corpus. The experiments investigate the effect of the size and accent-diversity of the supplementary training data, and the impact of including data from extreme accents. The best performance across all accents, a 32% relative reduction in WER compared with the baseline ASR system, is obtained when the supplementary data comprises extreme (Scottish) accented speech, even though this accent accounts for just 14% of the test data.

The paper begins with a review of previous work on ASR for accented speech (Section 2). The ABI and WSJCAM0 corpora are described in Section 3. Section 4 describes the i-vector based AID system used in the study. Section 5 presents the results of AID experiments on the ABI-1 corpus, together with an analysis of the regional accent diversity of the WSJCAM0 corpus and a 2 dimensional visualisation of the AID feature space. The AID results inform the

design of the ASR experiments described in the remainder of the paper.

Sections 6 and 7 describe the GMM-HMM and DNN-HMM ASR systems used in the study and the ASR experiments that were conducted on the accented data in the ABI-1 corpus. Section 8 presents the results of these experiments, including a comparison of the baseline GMM-HMM and DNN-HMM systems, an exploration of the relationship between ASR and AID, and an assessment of the utility of augmenting the acoustic feature vectors with i-vectors. The results obtained using accent-dependent GMM-HMM systems with AID-based model selection are also included, because this was previously the most effective GMM-HMM method for accent robust ASR on this data. The remainder of Section 8 is concerned with the effects on DNN-HMM based ASR of augmenting the WSJCAM0 training set with different subsets of ABI-1 with varying 'accent diversity', ranging from 'multi-accent', in which the supplementary data includes examples from all of the ABI accents (maximum accent diversity), to subsets that only contain recordings from a single broad accent group. Section 9 presents our conclusions.

2. Regional accents and ASR

Previous approaches to ASR for accented speech can be categorised as feature vector augmentation, acoustic modelling, pronunciation modelling, and combinations of these techniques.

2.1. Feature vector augmentation

Feature vector augmentation adds information about a speaker's accent to conventional acoustic feature vectors, incorporating accent information early in the classification process. Zheng et al. [16] append MFCCs with additional feature vectors which precisely track formant frequencies. This results in 1.4% character error rate reduction for Wu-accented Chinese speech from the Mandarin Broadcast news corpus. More recently, speaker- or utterance-level i-vectors [15] were added to acoustic feature vectors to enable DNNs to accommodate speaker

and session variability [17, 18, 19]. This method achieves a 1.53% Word Error Rate (WER) reduction on French TV broadcast data [19] and 10% [18] and 7.5% [20] relative improvements on Switchboard, compared with DNN baselines. Since i-vectors include information relevant to accent classification [14] one would expect their addition to acoustic feature vectors to provide robustness to accent variability in ASR.

2.2. Pronunciation modelling

Regional accents are characterised by systematic variations in pronunciation at the phone level. It is natural to try to compensate for this in the pronunciation dictionary. Reductions in WER for accented speech using rule-based modification of pronunciation dictionaries have been reported by a number of researchers [21, 22, 23, 24, 25, 26, 27]. More recently, [28] Polyphone Decision Trees (PDTs) have been used to model contextual acoustic variants in multi-accented Arabic speech, where PDT adaptation obtained 7% relative WER reduction compared with maximum a posteriori (MAP) [29] accent adaptation, on the Broadcast Conversations (BC) part of LDC GALE corpus. In another study [30] PDT adaptation achieved 13.9% relative improvement in WER compared with accent-specific MAP adaptation. A similar study has been conducted for variations of South African English [31].

2.3. Acoustic modelling

A number of authors use MAP and Maximum Likelihood Linear Regression (MLLR) [32] to adapt GMM-HMM based ASR systems to accented speech [33, 34, 35]. For example, a 78% reduction in WER for Korean spoken English was reported [34] from MAP and MLLR adaptation of a system trained on US English. For British English, it has been shown that adapting a baseline system to different accents and using AID to select an appropriate accent-dependent model, WER can be reduced by up to 47% [36, 10, 11]. Subspace Gaussian Mixture Models (SGMMs) [37], which enable more robust parameter estimation with limited data, are applied to accented English in [38], resulting in 8% relative

improvement in WER compared with speaker-adapted GMM-HMMs. A multi-accent DNN with an accent-specific top layer and shared hidden layers has been applied to British and Indian accented speech [39], resulting in reductions in WER of up to 30.6% on a short message dictation task, compared to a baseline system trained on 400 hours of speech. NN- and GMM-based acoustic models have also been compared in a study involving 412 hours of Chinese data with native and non-native accents [40].

2.4. Combination of multiple techniques

Additional improvements in ASR performance can be obtained by combining these approaches. Accent-specific pronunciation dictionary adaptation and accent-dependent acoustic model adaptation using MLLR are combined in [6] to achieve 36.02% relative WER reduction on Mandarin accented speech compared to a multi-accent baseline system. Chen et al. [41] combined acoustic feature vectors augmented with speaker-specific i-vectors with a DNN-HMM ASR system in which the hidden layers are common across all accents but the top softmax layers are accent-dependent. Their results showed a 11.8% relative improvement in %WER compared to a baseline DNN system.

3. Speech corpora

3.1. The ABI-1 "Accents of the British Isles" corpus

The ABI-1 corpus consists of speech from 285 speakers representing 13 British accent regions plus standard Southern British English (**sse**) [13] (Figure 1). For the 13 regions, regional accented speech was defined to be speech from individuals who had lived in the region since birth, while the **sse** speakers were selected by a phonetician. For each ABI-1 subject, his or her "true" accent is the accent (region or **sse**) that he or she represents in the corpus. The ABI accents fall into 4 broad accent groups (BAGs), namely Northern English (NE: **lan**, **ncl**, **lvp**, **brm**, **nwa**, **eyk**), Southern English (SE: **sse**, **crn**, **ean**, **ilo**), Scottish (SC: **shl**, **gla**) and Irish (IR: **uls**, **roi**). ABI contains only one example of a Welsh

accent (**nwa**), which was recorded in Denbigh in North Wales. Because of the proximity of Denbigh to the north of England, and Liverpool in particular, it was decided to include the **nwa** accent in the NE group rather than to include a Welsh set with one member. The boundaries between the BAGs in Figure 1 are very approximate and loosely based on [8] and [9]. ABI is used for training and testing our AID and ASR systems. Each of the subjects read the same 20 prompt texts. The experiments reported 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 comprising 92, 92 and 107 words, respectively, which together form the accent-diagnostic 'sailor passage' (*"When a sailor in a small craft ..."*)[1]. The recordings have average durations 43.2 s, 48.1 s and 53.4 s.
- '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 85.0 s).
- '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 34.5 s).

Table 1 complements Figure 1. It details the recording locations ("accents") for each BAG, the numbers of subjects and total hours of data for each accent and BAG, and the numbers of hours of data for each of the three subsets listed above for each BAG, namely the Short paragraphs SPA, SPB and SPC, Short sentences and Short phrases.

The ABI-1 corpus is publicly available[2].

[1] http://www.phon.ucl.ac.uk/resource/scribe/scribe-manual.htm

[2] http://www.thespeechark.com/abi-1-page.html

Table 1: *Details of the ABI-1 corpus. Horizontal blocks correspond to Broad Accent Groups (BAGs). Columns indicate BAG, accent, accent code, total subjects per accent and per BAG, total hours per accent and per BAG, and hours per BAG for the Short Passages A, B & C (SPass), Short Sentences (SSent) and Short Phrases (SPhr). The final row gives totals for the whole corpus.*

BAG	Location	Code	Subjects		Hours		Hours (per BAG)		
			accent	BAG	accent	BAG	SPass	SSent	SPhr
Northern England (NE)	Birmingham	brm	20	127	4.07	26.08	5.24	3.06	1.23
	East Yorkshire	eyk	25		6.24				
	Lancashire	lan	21		3.66				
	Liverpool	lvp	20		4.38				
	Newcastle	ncl	20		3.82				
	North Wales	nwa	21		3.90				
Southern England (SE)	Cornwall	crn	20	76	3.56	15.14	3.00	1.81	0.73
	East Anglia	ean	20		4.74				
	Inner London	ilo	21		3.59				
	Std. S. English	sse	16		3.26				
Scotland (SC)	Glasgow	gla	20	42	3.95	8.89	1.77	1.03	0.43
	Scottish Highlands	shl	22		4.94				
Ireland (IR)	Dublin	roi	20	40	3.51	7.57	1.44	0.83	0.35
	Belfast	uls	20		4.06				
Totals	14		285		57.7		11.5	6.7	2.73

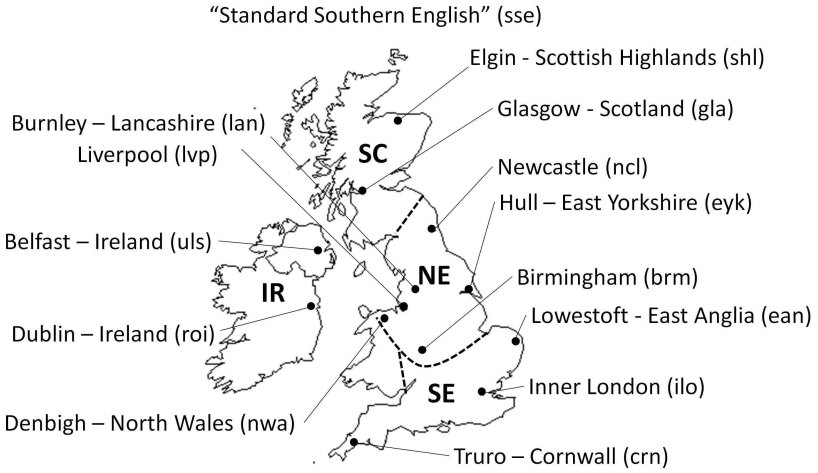

Figure 1: *Accents of the British Isles, their codes and their broad accent groups (***IR***,* **NE***,* **SC** *and* **SE***),represented in the ABI corpus*

3.2. The WSJCAM0 Corpus

The WSJCAM0 corpus is a British English version of the US American English WSJ0 corpus [12]. In the work reported in this paper, the WSJCAM0 training set (referred to in this paper as WSJT) is 15.5 hours of speech, comprising 90 utterances from each of 92 speakers, selected randomly in paragraph units from the WSJCAM0 training set. The development set is 2.25 hours of speech, made up of 90 utterances from each of 18 speakers, taken from the WSJCAM0 development set. The test set is part of the WSJCAM0 test set. It comprises recordings from 48 speakers each reading 40 sentences contained within a 5,000 word vocabulary (referred to as SI-dt-o5). Detailed transcriptions of all the utterances are available.

4. Automatic accent identification (AID)

This section describes the AID system and the experimental method that was used.

4.1. Automatic accent identification (AID) system

In this paper i-vector based AID is used for accent-specific acoustic model selection (Section 7.2), analysing the accent properties of the WSJCAM0 corpus (Section 5.2), and augmenting the acoustic features that are input to a DNN-HMM ASR system (Section 7.1).

An i-vector for a speech segment u is a representation of u in a relatively low dimensional 'total variability' vector space that is designed to capture salient information about u. The i-vector concept was proposed for speaker verification in [15] based on the work of Kenny et al. [42]. The GMM supervector s_u corresponding to u is represented as $s_u = s_0 + Tw_u + \epsilon$, where s_0 is the supervector corresponding to a 'universal' GMM estimated using all available training material, T is a linear transform from the total variability vector space V into the supervector space S, $w_u \in V$ is the i-vector corresponding to u and ϵ is the residual error. The transform T and the i-vectors w_u are determined by an iterative process to maximise, over the training set, the probability of the utterance u given the GMM corresponding to the supervector $s_0 + Tw_u$, where w_u is drawn from a zero mean unit covariance Gaussian distribution. GMM-based i-vectors have subsequently been replaced by "DNN i-vectors" where sufficient statistics are computed from senone posterior probabilities (obtained at the output of HMM/DNN acoustic model). These statistics are used to compute T-matrix [43].

Our system uses GMM i-vectors estimated using the Microsoft Research i-vector toolkit [44]. Acoustic feature vectors are 68 dimensional, comprising 19 MFCCs (with C0) plus Shifted-Delta Cepstral coefficients (7-3-1-7 configuration) [45]. All GMMs have 512 components. The dimension of the total variability vector space V is 200 for accent identification (Sections 5.1 and 7.2) and 100 for acoustic feature augmentation (Section 7.1). Further, the back-end classifier is represented by a multi-class Support Vector Machine (SVM) [46]. The SVM is trained to classify the i-vectors into the 14 accent classes.

4.1.1. AID training and test data

Our AID experiment used a 3-fold cross validation. The ABI corpus was divided into three subsets; two with 95 and one with 94 speakers. For each experiment, two subsets were used for training and the remaining subset was used for testing. The SPA utterances from each ABI speaker were used for testing and the SPB and SPC utterances plus 'short sentences' and 'short phrases' were used for training (Section 3.1).

5. Accent identification experiment results

This section reports the results of applying the i-vector based AID system (Section 4.1) to identify accents in the ABI-1 corpus (Section 5.1), analyse the distribution of regional accents of British English in the WSJCAM0 corpus (Section 5.2), and visualise the relationships between the accents in the ABI-1 corpus (Section 5.3). The results are report at this stage because they inform the ASR experiments in Section 6.

5.1. Accent identification results on ABI-1

The i-vector based AID system (Section 4.1) identifies the correct ABI-1 accent and correct BAG (NE, SE, SC, IR) with accuracies of 76.8% and 89.8%, respectively [14]. Table 2 shows the confusion matrix from this experiment. The blocks of the matrix correspond to the BAGs. The percentages of speakers from the NE, SE, IR and SC BAGs that are assigned to the incorrect BAGs are 9.4%, 21%, 0% and 2.4%, respectively. The poorest performance is for SE, with 18% of SE speakers assigned to the NE BAG. The poorest performance for an individual accent is for **nwa**, the North Wales accent. This is confused mainly with the northern English accents (though not with the Liverpool accent, to which it is closest, subjectively). The best performance is achieved for the two Scottish accents, **shl** and **gla**. Note that the 3-fold cross validation ensures that no test speaker appears in the training set for his or her experiment and excludes the possibility that the system is doing speaker, rather than accent, identification.

Table 2: *Confusion matrix for the i-vector accent identification system (NE: Northern English, SE: Southern English, SC: Scottish English, IR: Irish English)*

Accent code	Accent group	Acc.	brm	eyk	lan	lvp	ncl	nwa	ilo	sse	ean	crn	roi	uls	shl	gla
brm	NE	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	SE	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	IR	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	SC	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

5.2. Accent properties of WSJCAM0

According to [12], speakers in the WSJCAM0 corpus came from the Cambridge area, but effort was made to exploit Cambridge University's diverse population to include a wide range of regional accents. Our i-vector based AID system (Section 4.1) was applied to WSJCAM0 to investigate its distribution of regional accents of British English. The results (Figure 2), show that 32%, 47%, 13% and 8% of the subjects in WSJCAM0 are categorised as having Southern English (SE), Northern English (NE), Scottish (SC) and Irish (IR) accents, respectively, suggesting that the objective of finding speakers with a range of regional accents was achieved. Despite Cambridge's location in East Anglia, only 0.3% of the speakers were assigned to the **ean** class.

5.3. Visualisation of the ABI-1 i-vector accent space

Figure 3 is a 2D projection of the i-vector space for the ABI-1 corpus. Each of the 14 regional accents is represented by a 0.7-standard-deviation contour in

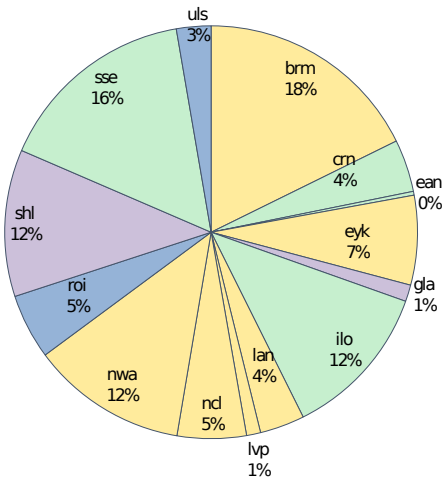

Figure 2: *Accent properties of WSJCAM0 according to i-vector based AID*

the space (hence the true overlap of the data is much greater than illustrated). A contour representing WSJCAM0 has been added. Principal Components Analysis (PCA) [47] followed by Linear Discriminant Analysis (LDA) [48] was used to map the i-vector space onto 2 dimensions [11]. The figure reflects, to some extent, natural relationships between the ABI accents. The clusters in the top right of the figure are Scottish, those in the upper left quarter are Northern English, and those in the bottom left quarter are mainly Southern English. The two Irish accents appear together near the bottom of the figure. Less intuitively, Standard Southern English (**sse**) appears at the bottom of the figure, and the Lancashire (**lan**) and East Yorkshire (**eyk**) accents intersect with North Wales (**nwa**), Inner London (**ilo**) and Cornwall (**crn**). The location of the WSJCAM0 contour in the centre of the figure, and the relationship between this and the other contours, are consistent with the distribution of accents in WSJCAM0 (Section 5.2). Finally, the AID results from Section 5.1 are consistent with the topology of Figure 3. For example, the lowest AID accuracies for accents in

the SE BAG are for **crn** (55%) and **ilo** (57%), which, as noted above, overlap significantly with the NE accents in the figure.

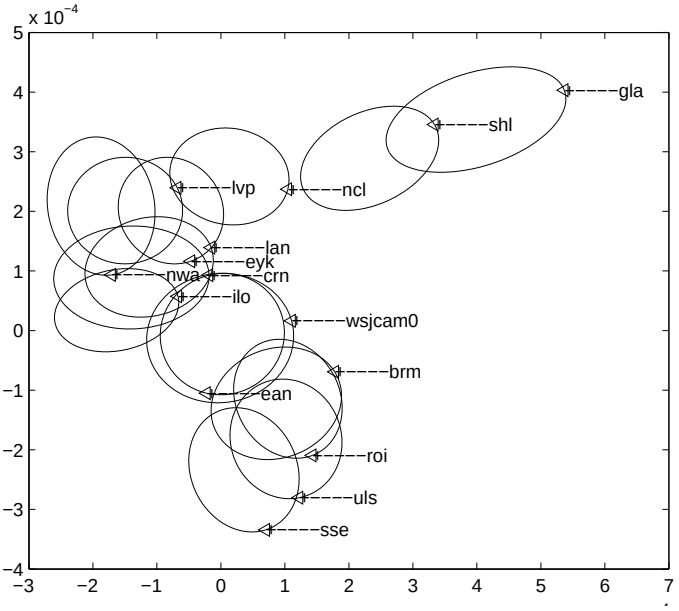

Figure 3: *Visualization of the i-vector accent space*

6. Automatic speech recognition systems

This section describes the ASR systems that were tested and the experimental method that was used.

6.1. Baseline automatic speech recognition (ASR) systems

All of our ASR systems are built using the Kaldi toolkit [49]. The baseline GMM-HMM and DNN-HMM ASR systems are trained on WSJT (section 3.2). These baseline systems are referred to as *G_0* and *D_0*, respectively. For the GMM-HMM system, speech is transformed into a sequence of 39 dimensional feature vectors, comprising mel frequency cepstral coefficients (MFCCs) 0 to 12 plus the corresponding Δ and Δ^2 parameters (25 ms analysis window, 100

vectors per second, mean and variance normalisation). A conventional tied-state GMM-HMM triphone system with 6 GMMs per state is trained on WSJT. The resulting system has 1619 physical states. This is the baseline GMM-HMM system *G_0*. Each utterance in WSJT is forced-aligned with the appropriate sequence of HMM states, associating each feature vector in WSJT with a unique physical HMM state (senone). We use the BEEP dictionary [50], extended to include all of the words in the ABI corpus.

A DNN was created with 195 input units, 5 hidden layers (each with 1024 neurons), and 1619 output units. The inputs to the DNN are 195 dimensional vectors comprising 13 dimensional mean and variance normalized filterbank features, spliced with a context of ± 7. During pre-training the input and hidden layers are treated as a Deep Belief Network (DBN) consisting of a stack of restricted Boltzman machines (RBMs) such that the hidden layer of each RBM acts as the visible layer for the next. The DBN weights are pretrained using 5 epochs of Contrastive Divergence training [51]. A 1619 dimensional "softmax" output layer is added. The resulting DNN is trained using minimum cross entropy driven stochastic gradient descent with a mini-batch size of 256, a learning rate of 0.008 and 1619 dimensional posterior probability targets determined by the mapping of feature vectors onto physical states. The resulting baseline DNN-HMM system is *D_0*.

The ASR system uses a bigram language model obtained as a weighted combination of the 5k WSJ0 bigram language model and a bigram language model built from the ABI corpus. The weight is chosen empirically to give the same WER on the ABI and WSJCAM0 development sets [36].

6.1.1. ASR training and test data

The Short Passage A (SPA) recordings from each speaker in the ABI-1 corpus are used as test data (average 43.2 s per speaker, total 3.42 hours). Accent-dependent training sets are taken from Short Passages B and C (SPB, SPC), the Short Sentences and the Short Phrases (3.1). In all cases care is taken to ensure that the test speaker is excluded from the training set. For example, in

unsupervised accent-dependent model selection experiments (Section (7.2)) with a given test speaker, the accent-specific model for the speaker's "true accent" was created using training material excluding data from that speaker (obviously, data for that speaker will not occur in accent-dependent training sets for the other 13 accents).

The experiments described in Section 7 use different subsets of the ABI-1 corpus to investigate accent- and BAG-specific training, the effect of training set size, and the utility of specific accents as training data for accent-robust ASR. The structure of the ABI-1 corpus (Section 3.1) was exploited to ensure that the distributions of words and speaker gender within the different sets was consistent. It was assumed that sets with similar distributions would give similar recognition results. However, the effect of choosing particular training sets within these constraints was not explicitly investigated.

7. ASR experiment conditions

For clarity each ASR experiment is given a code *X_Y[_Z]*. In this code *X* is either *G* (GMM-HMM) or *D* (DNN-HMM), *Y* is *0* (baseline (Section 6.1)), *S* (supervised accent adaptation (test speaker's "true" accent given)), or *U* (unsupervised accent adaptation (test speaker's accent identified using AID)). *Z* is an optional more detailed description of the experiment.

*7.1. DNN-HMM system with i-vector augmentation of feature vectors (*D_U_iV*)*

Augmenting each acoustic feature vector in a test utterance with a 100 dimensional i-vector derived from that utterance (Section 4.1) enables the DNN to learn the senone-level posterior probabilities for an acoustic feature vector given information contained in the i-vector about systematic variation in the utterance. Since i-vectors contain information about accent in addition to other types of variability [14], this approach to utterance-level speaker adaptation includes an element of accent adaptation. For this reason the method is included in the current study. The resulting system is *D_U_iV*. Similar approaches to speaker and accent adaptation are described in [18, 41].

*7.2. GMM-HMM system with unsupervised AID-based model selection (*G_U_MSel*)*

For each test speaker, 14 accent-specific GMM-HMM systems were created by MAP adaptation [29] of the baseline GMM-HMM system using approximately 40 minutes of accent-specific training data from the ABI-1 corpus per accent. Each such set equates to approximately 60% of the available accent-specific training data. As indicated in Section 6.1.1, the effect of using different 40 minute sets was not explicitly investigated. For the speaker's "true" accent, the accent-specific GMM-HMM system was created using training material excluding that speaker (Section (6.1.1)). An accent-specific model was chosen using the result of i-vector based AID (Section 4.1) applied to the test speaker's Short Passage A data (Sections 3.1, 6.1.1). Speaker-adaptation was then applied using MLLR [32] with 43 s of speaker-specific data. This system is referred to as *G_U_MSel*. Due to differences between the HTK [52] and Kaldi toolkits, the GMM-HMM system error rates reported in this study are lower than in [10, 11, 36]. All subsequent experiments use DNN-HMMs. This experiment enables the model-selection approach to accent-robust ASR from [10] to be compared with the baseline DNN system (*D_0*, Section 6.1) and i-vector augmentation (*D_U_iV*, Section 7.1).

*7.3. Supervised accent-dependent data augmentation (*D_S_**)*

These are also referred to as "oracle" systems because they know the test speaker's true accent. The experiments compare the performances of DNN-HMM systems in which the WSJT training set is augmented with data from the speaker's specific accent or data from the speaker's BAG (SE, NE, SC or IR). For each ABI accent, the WSJT training set is supplemented with 40 minutes of accent-specific data, resulting in 14 accent-specific (AS) DNN-HMM systems. Separately, WSJT is supplemented with 2.25 hours of data from one of the 4 'broad accent' groups (BAGs), namely Southern English (SE), Northern English (NE), Scottish (SC) or Irish (IR) (Section 3.1), resulting in 4 BAG dependent systems. The BAG-specific training sets are restricted to 2.25 hours because this

is the largest possible training set for the smallest BAG (IR). As in previous experiments, the effect of choosing different 2.25 hour sets for the bigger BAGs was not investigated, but the different training sets share the same distributions of gender and vocabulary. For a given test speaker, accent-dependent model selection is supervised, according to the speaker's true accent. As always, the current test speaker is excluded from the training set for his or her accent-specific system (Section (6.1.1)). These experiments are denoted by (*D_S_AS*) and (*D_S_BAG*) (where *D* indicates DNN-HMM, *S* indicates supervised model selection (the test speaker's true accent is know) and *_AS* and *_BAG* indicates accent-specific and broad accent group dependent systems, respectively.

Since these experiments were conducted a number of methods have been proposed for DNN-HMM adaptation using restricted domain-specific data, for example by training a small set of new [53] or existing DNN parameters (typically those in the *softmax* layer) [54, 55]. DNNs have also been adapted to compensate for factors including far-field microphones and room acoustics [56].

*7.4. Accent-independent, unsupervised DNN-HMM systems (*D_U_**)*

An alternative to accent-dependent training followed by model selection is to use the same multiple accent DNN-HMM system for all speakers. Two properties of the data that is added to WSJT to train an accent-independent model are potentially important, namely its quantity and its 'accent diversity' (the number and types of different accents in the data). The experiments in this section investigate these two factors.

*7.4.1. Multiple-accent DNN-HMMs (*D_U_MA**)*

This experiment investigates the effect of the quantity of multi-accent data that is added to the WSJT training set. Accent-independent multi-accent (MA) DNN-HMMs are created by adding (a) 2.25 hours and (b) 8.96 hours of data, distributed uniformly across all 14 ABI-1 regional accents, to WSJT. These experiments are unsupervised (*_U*), since the test speaker's true accent is unknown (and not required), and are labelled *D_U_MA2.25* and *D_U_MA8.96*. As in the

previous experiments the different training sets have matching distributions of gender and vocabulary but the effects of using different 2.25 and 8.96 hour training sets were not investigated.

*7.4.2. Effect of accent diversity (*D_U_AD**)*

In contrast to the previous experiments, these experiments investigate the effect of the *accent diversity* (AD) of the training data (the number of different accents represented in the training set), rather than simply the quantity of data. WSJT is supplemented with a total of 2.25 hours of data from (a) 2 accents from the NE (North of England) accent group (low accent diversity, *D_U_AD2*), (b) 4 accents from the NE, SC (Scottish) and IR (Irish) accent groups (medium accent diversity *D_U_AD4*), (c) 8 accents from the NE, SC, IR and SE (Southern England) accent groups (medium accent diversity *D_U_AD8*), and (d) all 14 accents in the ABI corpus (high accent diversity, (*D_U_AD14).* Note that case (d) is identical to the first multiple-accent system in (Section 7.4.1). Hence *D_U_AD14* and *D_U_MA2.25* are the same.

These experiments are just a small sample from the many ways that the accent groups and specific accents could be chosen in each condition, and this should be taken into account when interpreting the results. As in the previous experiments the distribution of gender and vocabulary is consistent between training sets.

*7.4.3. Effect of different broad accents groups (*D_U_BAG(*)*)*

Figure 3 shows the relationship between WSJCAM0 and the various accents in ABI-1. The significant overlap between WSJCAM0 and the **ean** accent in the i-vector accent space suggests that adding **ean** data to WSJST would have little effect, whereas adding **gla** data might have a significant effect, because **gla** is the most distant accent from WSJCAM0 and adding **gla** data to WSJT would increase the accent diversity of the training data. Figure 2 also indicates minimal overlap between WSJCAM0 and **gla**. The purpose of this experiment is to investigate possible relationships between the proximity of a regional accent to

WSJCAM0 in figure 3 and the utility of that accent as additional training material. Four systems are created by augmenting WSJT separately with 2.25 hours of data from each of the 4 BAGs, SE (*D_U_BAG(SE)*), NE (*D_U_BAG(NE)*), SC (*D_U_BAG(SC)*) and IR (*D_U_BAG(IR)*). As in the previous experiments the distribution of gender and vocabulary is consistent between training sets to try to ensure that accent is the only variable that is changed.

8. ASR experiment results

This section presents the performances of the ASR systems described in section 7. Results are presented in terms of percentage Word Error Rate (%WER) and the Average WER Reduction (%AWR) with respect to the baseline DNN-HMM system (6.1). The results are summarised in Table 4. The final column of the table shows the amount of accented data that is added to WSJT. Where the training set contains material from ABI-1, the %AWR for accents in the training set ('Target') and not in the training set ('Off target') are also included. All recognition experiments on ABI-1 use the SPA data (Section 3.1) for test. Experiment parameters were optimized empirically using cross-validation.

*8.1. Results: Baseline GMM- and DNN-HMM ASR (*G_0 *and* D_0*)*

The baseline GMM-HMM (*G_0*) and DNN-HMM (*D_0*) systems (Section 6.1) achieve average %WERs of 12.89% and 6.85%, respectively, on the ABI-1 test set. For the GMM-HMM system, WER increases from 3.5% for **sse** to 26.7% for **gla**. The baseline DNN-HMM system also achieves its best and poorest performances, 2.9% and 13.4% WER, on **sse** and **gla**, respectively. Figure 5 shows the %WERs for each ABI-1 accent for both systems. Overall it is clear that the DNN-based system performs better than the GMM-based system. The average %WERs for the baseline GMM-HMM and DNN-HMM systems (*G_0* and *D_0*) are shown in block 2 of Table 4.

Table 3: *Summary of the motivations for testing each of the ASR systems from Section 4.*

Code	Description	Section
Baselines		
G_0	Baseline GMM-HMM trained on WSJCAM0 only.	6.1
D_0	Baseline DNN-HMM trained on WSJCAM0 only.	6.1
D_U_iV	*D_0* with acoustic vectors supplemented with accent-dependent information: each acoustic feature vector is augmented with an utterance-level 100 dimensional i-vector	7.1
GMM-HMM with unsupervised model selection		
G_U_MSel	Accent-specific GMM-HMM chosen by i-vector AID. Enables GMM-HMM model-selection method from [10] to be compared with DNN-HMM baselines	7.2
'Oracle' DNN-HMM systems - test speaker 'true' accent known		
D_S_AS	Accent-specific DNN-HMM trained on WSJT + 40min data for test speaker's true accent	7.3
D_S_BAG	BAG-specific DNN-HMM trained on WSJT + 2.25hrs ABI-1 data for speaker's true BAG. *D_S_AS* and *D_S_BAG* enable comparison of accent-specific and BAG-specific training.	7.3
Accent-independent 'multi-accent' DNN-HMM systems		
D_U_MA2.25	Multi-accent DNN-HMM trained on WSJT + 2.25hrs data from all 14 accents. Enables comparison of 'multi-accent' system with accent- and BAG-dependent systems	7.4.1
D_U_MA8.96	Multiple-accent DNN-HMM trained on WST + 8.96hrs data from all 14 accents. Comparison with *D_U_MA2.25* indicates the effect of the quantity of multi-accent data	7.4.1
Effect of 'accent diversity' of training set on DNN-HMM performance on accented speech		
D_U_AD2	DNN-HMM trained on WST + 2.25hrs ABI-1 data from 2 accents (NE BAG)	7.4.2
D_U_AD4	DNN-HMM trained on WST + 2.25hrs ABI-1 data from 4 accents (NE, SC & IR BAGs)	7.4.2
D_U_AD8	DNN-HMM trained on WST + 2.25hrs ABI-1 data from 8 accents (NE, SC, IR & SE BAGs)	7.4.2
D_U_AD14	DNN-HMM trained on WST + 2.25hrs ABI-1 data from all accents	7.4.2
Effect of selecting training data from different BAGs on DNN-HMM performance on accented speech		
D_U_BAG(SC)	DNN-HMM trained on WST + 2.25hrs ABI-1 data from SC BAG	7.4.3
D_U_BAG(IR)	DNN-HMM trained on WST + 2.25hrs ABI-1 data from IR BAG	7.4.3
D_U_BAG(SE)	DNN-HMM trained on WST + 2.25hrs ABI-1 data from SE BAG)	7.4.3
D_U_BAG(NE)	DNN-HMM trained on WST + 2.25hrs ABI-1 data from NE BAG) These systems enable the utility of different BAGs for training an accent-robust DNN-HMM system to be compared with reference to their locations in AID space (Figure 3)	7.4.3

8.2. The relationship between ASR and AID error rates

Figure 4 is a scatter plot of %WER for the baseline ASR system *D_0* as a function of AID error rate, showing the best straight-line fit to the data. In this experiment, as the AID error rate increases the ASR error rate tends to decrease, and the correlation coefficient is -0.746. Intuitively, if an accent lies at the extremes of the i-vector accent space and has minimal overlap with other accents in that space, then the AID error rate for this accent will be small. At the same time, because of the location of the WSJCAM0 training set in the i-vector space (figure 3), this accent is poorly represented in the ASR training set and so the ASR error rate for this accent is large. For example, Glasgow (**gla**) and the Scottish Highlands (**shl**) have the smallest AID error rates of 5% and 0%, respectively, and the biggest ASR WERs of 13.3% and 11.5%, respectively. Note that **gla** and **shl** are also the most distinct from the other accents and furthest from WSJCAM0 in the i-vector space (Figure 3). The accent i-vector space provides a means to visualize the relationship between the training set and regions that correspond to accents that may be encountered in testing. The objective should be to ensure that the training set covers these regions.

This relationship suggests that i-vector based analysis of an ASR training set, as described in Section 4.1, can anticipate the types of speech for which that ASR system will perform poorly, and hence indicate categories of speech that could most usefully be added to the training set to improve ASR performance. The effect of supplementing the WSJCAM0 training set with different types of accented speech is investigated in the remainder of this section.

*8.3. Results: DNN-HMM, i-vector augmentation of feature vectors (*D_U_iV*)*

Augmenting the input to the DNN with an utterance dependent i-vector results in a WER of 6.2%, an absolute improvement of 0.7% (9.4% AWR) relative to the DNN baseline *D_0*. These results are presented in block 3 of Table 4.

*8.4. Results: GMM-HMM, unsupervised AID-based model selection (*G_U_MSel*)*

The GMM-HMM system with unsupervised AID-based model selection and speaker adaptation from [10] (Section 7.2) achieves an average WER of 7.4%.

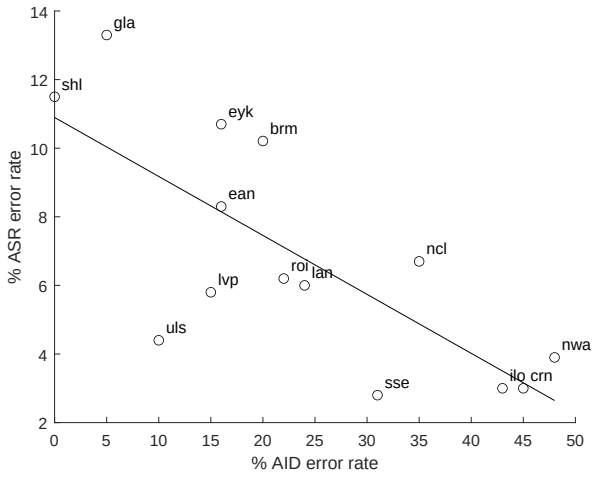

Figure 4: *Plot of ASR %WER for the baseline* D_0 *system as a function of % AID error rate*

This is substantially better than the %WER for the baseline GMM-HMM system *G_0* (5.5% absolute improvement, 42.9% AWR), but slightly poorer than for the baseline DNN-HMM system *D_0* (0.5% absolute increase in %WER, -7.5% AWR). Thus, the baseline DNN-HMM system, trained on WSJCAM0 only, achieves better performance on accented British English speech than the best GMM-HMM adaptation method from [10]. No further GMM-HMM experiments were conducted. Figure 5 shows the %WER for each ABI accent for *G_U_MSel*. These results are shown in block 4 of Table 4.

*8.5. Results: Supervised accent-dependent data augmentation (*D_S_**)*

This experiment compares the baseline DNN-HMM result (*D_0*) with DNN-HMM systems trained on WSJT plus 40 minutes of accent-specific data (*D_S_AS*), or 2.25 hours of data from a BAG (*D_S_BAG*). This is supervised training because the model that is used for recognition is chosen to match the test speaker's accent or BAG. The average WERs over the ABI test data for *D_S_AS* and *D_S_BAG* are 5.1% and 4.9%, suggesting that the larger amount of data available for BAG training outweighs any advantage from the specificity

of accent-dependent training. Augmenting the WSJT training set with accent-specific (*D_S_AS*) and broad accent group specific data (*D_S_BAG*) results in absolute WER reductions of 1.8% (25.8% AWR) and 1.9% (28.0% AWR), respectively, compared with the DNN-HMM baseline. These results are block 4 of Table 4.

A further comparison is between *DS_S_AS* and a version of GMM-HMM model selection where the accent-specific GMM-HMM is chosen according to the speaker's true accent rather than the result of AID. Although this experiment was not conducted in the present study, the results in [10] suggest that the WER would be approximately 3% better than the 7.4% achieved by *G_U_MSel*, or 7.2%. This is poorer than both *DS_S_AS* and the DNN-HMM baseline.

*8.6. Results: Accent-independent, unsupervised DNN-HMM systems (*D_U_**)*

*8.6.1. Results: Multiple-accent DNN-HMMs (*D_U_MA2.25 *and* D_U_MA8.96*)*

The experiment investigates the effect of supplementing WSJT with multi-accent data with fixed accent diversity but varying size. We compare the results of training the DNN-HMM system on WSJT augmented with 2.25 hours (*D_U_MA2.25*) and 8.96 hours (*D_U_MA8.96*) of data selected uniformly across all 14 ABI-1 accents. The results are shown in block 5 of Table 4. Adding 2.25 hours of multi-accent training data to WSJT reduces the average WER on accented data to 4.9%, an absolute reduction of 2% (28.3% AWR) relative to the DNN-HMM baseline. This accent-independent, multi-accent training result is very similar to the result for the 'oracle' accent-dependent system trained on the same amount of data from the correct BAG (*D_S_BAG*). In other words, knowing the test subject's accent and explicitly augmenting the training set with data from the correct accent or BAG appears to have no advantage over multi-accent training. Quadrupling the quantity of supplementary multi-accented data leads to a %WER to 4.4%, an absolute reduction of 2.5% (35.9% AWR). This suggests that accent-diversity is a key factor.

Table 4: *Summary of ASR results from Section 8. Columns 3, 4, 5, 6, 7, 8 and 9 indicate the %WER, the % reduction in average WER relative to the baseline, the % reduction in average WER for accents represented in the supplementary data, the % reduction in average WER for accents that are not represented in the supplementary data, the number of regional accents, the number of broad accent groups, and the amount of data in the supplementary training data.*

Code	System	WER	AWR(%)			Regional	Broad	length
		(%)	Avg	target	off target	accents	accents	(hours)
Baselines								
G_0	Baseline GMM-HMM trained on WSJCAM0 only	12.9	-	-	-	-	-	-
D_0	Baseline DNN-HMM trained on WSJCAM0 only	6.9	-	-	-	-	-	-
D_U_iV	*D_0* with acoustic vectors augmented with i-vectors	6.2	9.4	-	-	-	-	-
GMM-HMM with unsupervised model selection								
G_U_MSel	Accent-specific GMM-HMM selected using AID	7.4	-7.5	-	-	-	-	-
"Oracle" DNN-HMM systems - test speaker "true" accent known								
D_S_AS	Accent-specific DNN-GMM for "true" accent	5.1	25.8	-	-	-	-	-
D_S_BAG	BAG-specific DNN-HMM for "true" BAG	4.9	28.0	-	-	-	-	-
Accent-independent 'multi-accent' DNN-HMM systems								
D_U_MA2.25	Multi-accent DNN-HMM (2.25hrs data, 14 accents)	4.9	28.3	-	-	14	4	2.3
D_U_MA8.96	Multi-accent DNN-HMM (8.965hrs data, 14 accents)	4.4	35.9	-	-	14	4	9.0
Effect of 'accent diversity' of training set on DNN-HMM performance on accented speech								
D_U_AD2	Low-diversity DNN-HMM (2.25hrs data, 2 accents)	5.1	25.4	42.5	14.2	2	1	2.3
D_U_AD4	Med.-diversity DNN-HMM (2.25hrs data, 4 accents)	5.0	26.9	16.8	22.0	4	3	2.3
D_U_AD8	Med.-diversity DNN-HMM (2.25hrs data, 8 accents)	4.9	28.3	18.2	29.0	8	4	2.3
D_U_AD14	High-diversity DNN-HMM (2.25hrs data, all accents)	4.9	28.3	21.9	-	14	4	2.3
Effect of selecting training data from different BAGs on DNN-HMM performance on accented speech								
D_U_BAG(SC)	DNN-HMM (2.25hrs data from SC BAG)	4.7	31.7	28.7	24.4	2	SC	2.3
D_U_BAG(IR)	DNN-HMM (2.25hrs data from IR BAG)	4.9	28.3	23.2	21.6	2	IR	2.3
D_U_BAG(SE)	DNN-HMM (2.25hrs data from SE BAG)	5.6	17.7	-4.3	8.7	4	SE	2.3
D_U_BAG(NE)	DNN-HMM (2.25hrs data from NE BAG)	5.3	22.3	19.3	12.4	6	NE	2.3

*8.6.2. Effect of accent diversity (*D_U_AD**)*

Augmenting WSJT with data from just 2 accents from the NE BAG results in a WER of 5.1%, an absolute reduction of 1.8% (25.4% AWR) in WER across all accents compared with the DNN-HMM baseline. The improvement for test utterances whose accent is represented in the supplementary training data ('Target') is much greater than for accents that are not represented ('Off-target'). As the accent diversity of the additional training material increases, the average WER decreases until it matches the multiple accent system *D_U_MA2.25* when 8 accents are represented in the training data. This is achieved mainly through improvements in the off-target %AWR. The results are shown in block 6 of Table 4. The largest incremental improvement in %WER is obtained with (*D_U_AD2*). Although the additional training data from the NE BAG has low accent-diversity, Figure 3 shows that it differs from WSJCAM0 in AID space. A possible explanation is that extending the training data in this way into the NE region of AID space exposes the DNN to more accent variability, enabling it to cope better even with unseen accents. This raises the question of whether adding low-diversity data from other BAGs will lead to a similar result. This is investigated in the next experiment.

*8.6.3. Effect of different accent groups (*D_U_BAG(*)*)*

This experiment measures the effect of adding 2.25 hours of data separately from each of the four BAGs to WSJT. The results are presented in block 7 of Table 4. The lowest %WER is obtained by supplementing WST with Scottish (SC) BAG data, which gives a WER of 4.7%, an absolute improvement of 2.2% (28.7% AWR) relative to the DNN-HMM baseline. This is the best performing of all the systems with 2.25 hours of additional training data. Moreover, it gives the biggest average reduction in %WER not only for the Scottish accents ('Target') but also for the other accents ('Off-target'). For example, it is better to augment WSJT with 2.25 hours of Scottish (SC) data than with 2.25 hours of multi-accent data taken from all 14 ABI accents. There appears to be a relationship between the utility of a BAG as additional training data, the distance of that BAG from

WSJCAM0 in the i-vector space in Figure 3 and the proportion of WSJCAM0 that is allocated to that BAG by AID in Figure 2. The Scottish and Irish BAGs are least represented in WSJCAM0 in Figure 2, furthest from WSJCAM0 in Figure 3 and give the best results in this experiment. Conversely, the Northern England (NE) and Southern England (SE) BAGs are already well-represented in WSJCAM0, accounting for 79% of the speakers (Figure 2), are closer to WSJCAM0 in Figure 3, and augmenting WSJT with data from the SE or NE BAGs results in the lowest reductions in %WER.

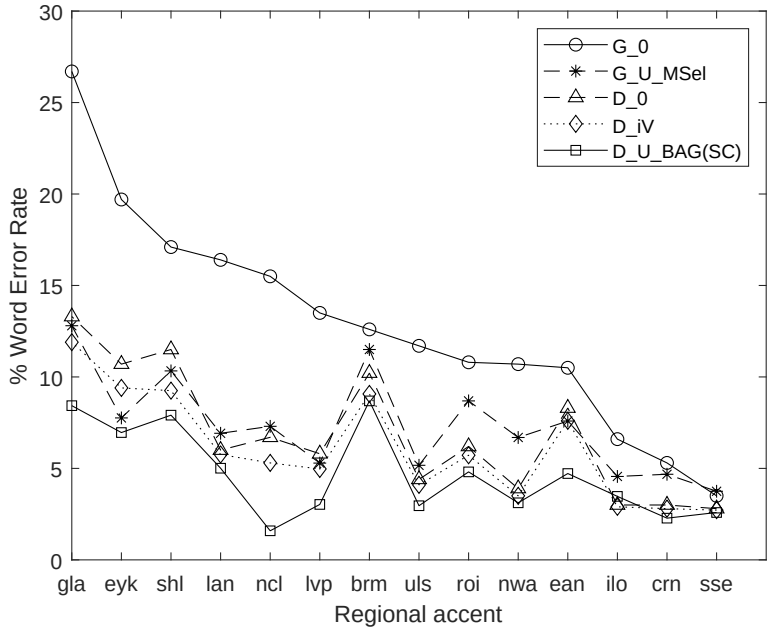

Figure 5: *%WER per regional accent for the GMM-HMM (*G_0*) and DNN-HMM (*D_0*) baseline systems, GMM-HMM AID-based model selection plus speaker adaptation (*G_U_MSel*), i-vector based DNN-HMM adaptation (*D_U_iV*) and DNN-HMM training using WSJT augmented with Scottish accented data (*D_U_BAG(SC)*)*

8.7. Detailed results for different regional accents

Figure 5 shows %WER for each accent in the ABI corpus for the GMM-HMM (*G_0*) and DNN-HMM (*D_0*) baseline systems, the DNN-HMM system with

i-vector augmentation of acoustic feature vectors (*D_U_iV*), accent-dependent GMM-HMM systems with AID-based model selection followed by speaker adaptation (*G_U_MSel*), and training using WSJT augmented with 2.25 hours of Scottish accented speech (*D_U_BAG(SC)*, the best system with 2.25 hours of additional training material).

8.8. Discussion of ASR results

The DNN-HMM baseline system (*D_0*) achieves a WER of 6.85%, a reduction of approximately 47% relative to the GMM-HMM baseline (*G_0*), indicating that the DNN-HMM is inherently more able to accommodate variability due to regional accent. Augmenting the acoustic feature vectors with an utterance-dependent i-vector (*D_U_iV*) results in a further reduction in %WER of 6.2% relative to the DNN-HMM baseline. The approach from [10] which uses AID to select an accent-dependent GMM-HMM system (*G_U_MSel*) is outperformed by both the DNN-HMM baseline (*D_0*) and i-vector augmentation. A negative correlation, with respect to accent, is noted between AID accuracy and the performance of the baseline DNN-HMM system. Accents at the edge of the AID space achieve good AID accuracy, because they are relatively separate from other accents, but poor ASR accuracy because they are not well-represented in the training data. Thus, AID may be useful for anticipating ASR problems.

The lowest WER (4.4%) is obtained by supplementing WST with 8.96 hours of multi-accent data (*D_U_MA8.96*), reinforcing the importance of the quantity of data for DNN training.

In the remaining experiments, in which WSJT is supplemented with 2.25 hours of accented data, the WER ranges from 5.6% (*D_U_BAG(SE)*) to 4.7% (*D_U_BAG(SC)*). For multi-accent training, performance improves with accent diversity. The best performance, 4.7%, is obtained by supplementing WSJT with 2.25 hours of Scottish data, and is a reduction in %WER of 5% relative to the best result for multi-accent training with 2.25 hours of data (4.9%, *D_U_AD8*). By analysing the accent content of WSJCAM0 (Figure 2) and visualising its relationship with the accents in the ABI-1 corpus in i-vector AID

space (Figure 3), the AID experiments quantify the relationships between the different accented data sets and enable the subsequent ASR experiments to be interpreted in terms of accent diversity and the properties of the AID space. These ASR experiments suggest that the utility of additional training material depends on the extent to which it introduces new variability that is not included in the original training set, and that this can be inferred from the AID space and the distribution of accents in the training set. The most striking result is that ASR performance for Irish (IR), Northern English (NE) and Southern English (SE) accented speech is improved by supplementing the WSJCAM0 training set with Scottish (SC) accented data. It is hypothesised that this is related to the extreme position of the SC data in AID space and its absence from the WSJT training set.

9. Conclusions

This paper has explored the relationships between AID and ASR accuracy for regional accented British English speech, using i-vector based AID and GMM-HMM and DNN-HMM based ASR. In particular it has shown that the i-vector based AID feature space contains information that is useful to understand ASR performance on particular types of accented data, can shed light on the relationships between different accents, and can anticipate the utility of different types of training data for ASR.

The first contribution (Section 5.2) is an analysis of the accent diversity of the WSJCAM0 corpus. This is achieved by applying an i-vector based AID system to WSJCAM0 (Section 4.1). This indicates that 47%, 32%, 13% and 8% of its subjects are categorised as speakers of Northern English (NE), Southern English (SE), Scottish (SC) and Irish (IR) English, respectively (Figure 2). This is quantitative evidence that the creators of WSJCAM0 were able to take advantage of Cambridge University's diverse population to find a wide range of regional accents [12]. The next contribution of the paper is to provide insights into the relationships between the different accents in the ABI corpus through

visualisation of the i-vector AID space (Section 5.3). A two-dimensional projection of the space is created, using PCA and LDA, in which each accent is represented by its mean and a 0.7 standard deviation contour (Figure 3). Sectors of the i-vector space can be seen to correspond, approximately, to the broad accent groups of Northern and Southern England, Scotland and Ireland. The 0.7 standard-deviation contour for WSJCAM0 is located in the centre of the figure, between the Northern and Southern English BAGs, which is consistent with Figure 2.

The second part of the paper is concerned with ASR for accented speech. The baseline DNN-HMM system trained on WSJCAM0 outperforms both the corresponding GMM-HMM baseline and the model selection method proposed in [10], in which AID is used to select an accent-dependent GMM-HMM system. Thus it appears that DNN-HMMs have some inherent ability to accommodate accented speech. The performance of the DNN-HMM baseline can be further improved by augmenting the acoustic feature vectors with an utterance dependent i-vector, indicating that the DNN can exploit the accent information contained in the i-vector. Inspection reveals a negative correlation between AID accuracy for a particular accent and the performance of the baseline DNN-HMM ASR system for that accent. The remainder of the paper describes DNN-HMM experiments in which the WSJCAM0 training set WSJT is supplemented with various sets of accented data. The novelty of these experiments stems from the use of Figures 3 and 2 to understand the relationship, in terms of accent, between this supplementary data and WSJCAM0 and its diversity in accent space. The results show that the quantity and the location of the supplementary data in the AID feature space both influence ASR performance. In the case of the "oracle" systems, which use knowledge of the speaker's accent to choose an appropriate accent- or BAG-dependent DNN-HMM, the BAG-dependent system performs best. This has practical implications because AID systems can identify a speaker's BAG more accurately than his or her precise accent (89.8% compared with 76.8% (Section 5.1)). However, both of the "oracle" systems are outperformed by unsupervised multi-accent DNN-HMM systems that do

not need to know the test speaker's accent. The most interesting results are presented in Section 8.6.3 where the supplementary data is drawn from a single broad accent group. The results show that the lowest %WER is obtained by adding Scottish (SC) data, which results in the biggest reductions in %WER not only for the 'matched' Scottish accents but also for the other 'unmatched' accents. The proposed explanation is that the Scottish data, located at the periphery of the AID accent space, exposes the DNN to 'extreme' accented speech and thereby renders it also able to accommodate other accents. In fact, it is better to augment WSJT with 2.25 hours of Scottish accented data than with 2.25 hours of multi-accent data taken from all 14 ABI accents. Compared with the DNN-HMM baseline, adding this Scottish data to the training set results not only in a 28.7% improvement in WER for Scottish accented speech but also a 24.4% relative improvement for non-Scottish accents. Comparison of Table 4 and Figure 3, suggests that there is a relationship between the distance of an accent group from WSJCAM0 in the i-vector space and the recognition accuracy obtained by using that accent group as supplementary training data.

Further work is required to determine if these results generalise to other deep learning structures, larger systems, and training corpora. If this is the case then they have the potential to provide a principled way to choose effective training material and predict ASR problems in a range of applications. Although the experiments described here are restricted to the relationship between AID and ASR, it may also be possible to use other detection technologies, such as Speaker Identification (SID), as an analysis tool in ASR. It may also be effective to apply these methods to abstract classes, rather than accent groups, within a training corpus. For example, these classes could be derived using data-drive clustering. In this case "extreme accents" would be replaced by "extreme clusters" and one would hope to establish a relationship between identification accuracy with respect to a particular cluster and ASR accuracy for speech from that cluster.

References

[1] G. Hinton, L. Deng, D. Yu, G. E. Dahl, M. Abdel-Rahman, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, B. Kingsbury, Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups, IEEE Signal Process. Mag. 29 (6) (2012) 82–97.

[2] Y. Huang, D. Yu, C. Liu, Y. Gong, A comparative analytic study on the Gaussian mixture and context dependent deep neural network hidden Markov models., in: Proc. Interspeech, *Singapore*, 2014, pp. 1895–1899.

[3] Y. Huang, M. Slaney, M. L. Seltzer, Y. Gong, Towards better performance with heterogeneous training data in acoustic modeling using deep neural networks, in: Proc. Interspeech, *Singapore*, 2014, pp. 845–849.

[4] The Washington Post, The Accent Gap (July 19, 2018).
URL `https://www.washingtonpost.com/graphics/2018/business/alexa-does-not-understand-your-accent/?noredirect=on&utm_term=.e7719874c8f1`

[5] The Daily Mail, Brummie accents baffle automated phone system ... at Birmingham City Council (November 5, 2012).
URL `https://www.dailymail.co.uk/news/article-2228029/Brummie-accents-baffle-automated-phone-Birmingham-City-Council.html`

[6] C. Huang, T. Chen, E. Chang, Accent issues in large vocabulary continuous speech recognition, International Journal of Speech Technology 7 (2-3) (2004) 141–153.

[7] A. Hanani, M. Russell, M. Carey, Human and computer recognition of regional accents and ethnic groups from British English speech, Computer Speech & Language 27 (1) (2013) 59–74.

[8] J. C. Wells, Accents of English: The British Isles, Vol. 2, Cambridge University Press, 1982.

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

[10] M. Najafian, A. DeMarco, S. Cox, M. Russell, Unsupervised model selection for recognition of regional accented speech, in: Proc. Interspeech, *Singapore*, 2014.

[11] M. Najafian, S. Safavi, A. Hanani, M. Russell, Acoustic model selection for recognition of regional accented speech, in: Proc. 22nd European Signal Processing Conference (EUSIPCO) *Lisbon, Portugal*, 2014.

[12] A. Robinson, J. Fransen, D. Pye, J. Foote, S. Renals, Wsjcam0: A british english speech corpus for large vocabulary continuous speech recognition, in: Proc. IEEE-ICASSP, *Detroit, MI*, 1995.

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

[14] M. Najafian, S. Safavi, P. Weber, M. J. Russell, Identification of British English regional accents using fusion of i-vector and multi-accent phonotactic systems, in: Proc. Odyssey'16, The Speaker and Language Recognition Workshop *Bilbao, Spain*, 2016.

[15] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, Front-end factor analysis for speaker verification, IEEE Transactions on Audio, Speech, and Language Processing 19 (4) (2011) 788–798.

[16] Y. Zheng, R. Sproat, L. Gu, I. Shafran, H. Zhou, Y. Su, D. Jurafsky, R. Starr, S.-Y. Yoon, Accent detection and speech recognition for shanghai-accented mandarin, in: Proc. Interspeech, *Lisbon, Portugal*, 2005, pp. 217–220.

[17] A. W. Senior, I. Lopez-Moreno, Improving DNN speaker independence with i-vector inputs, in: Proc. IEEE-ICASSP, *Florence, Italy*, 2014, pp. 225–229.

[18] G. Saon, H. Soltau, D. Nahamoo, M. Picheny, Speaker adaptation of neural network acoustic models using i-vectors, in: Proc. ASRU 2013, IEEE Automatic Speech Recognition and Understanding Workshop *Olomouc, Czech Republic*, IEEE, 2013, pp. 55–59.

[19] V. Gupta, P. Kenny, P. Ouellet, T. Stafylakis, I-vector-based speaker adaptation of deep neural networks for French broadcast audio transcription, in: Proc. IEEE-ICASSP, *Florence, Italy*, IEEE, 2014, pp. 6334–6338.

[20] Y. Miao, H. Zhang, F. Metze, Towards speaker adaptive training of deep neural network acoustic models, in: Proc. Interspeech, *Singapore*, 2014.

[21] L. Kat, P. Fung, Fast accent identification and accented speech recognition, in: Proc. IEEE-ICASSP, *Phoenix, AZ*, Vol. 1, 1999, pp. 221–224.

[22] D. Vergyri, L. Lamel, J.-L. Gauvain, Automatic speech recognition of multiple accented English data, in: Proc. Interspeech, *Makuhari, Japan*, 2010, pp. 1652–1655.

[23] S. Goronzy, Robust adaptation to non-native accents in automatic speech recognition, Vol. 2560 of Lecture Notes in Computer Science, Springer, 2002.

[24] S. Goronzy, S. Rapp, R. Kompe, Generating non-native pronunciation variants for lexicon adaptation, Speech Communication 42 (1) (2004) 109–123.

[25] J. J. Humphries, P. C. Woodland, Using accent-specific pronunciation modelling for improved large vocabulary continuous speech recognition., in: Proc. Eurospeech '97, *Rhodes, Greece*, 1997.

[26] M. Tjalve, M. Huckvale, Pronunciation variation modelling using accent features, in: Proc. Interspeech, *Lisbon, Portugal*, 2005, pp. 1341–1344.

[27] Z. Wang, T. Schultz, A. Waibel, Comparison of acoustic model adaptation techniques on non-native speech, in: Proc. IEEE-ICASSP, *Hong Kong*, Vol. 1, IEEE, 2003, pp. I–540.

[28] U. Nallasamy, F. Metze, T. Schultz, Enhanced polyphone decision tree adaptation for accented speech recognition, in: Proc. Interspeech, *Portland, Oregon, USA*, 2012, pp. 1902–1905.

[29] J.-L. Gauvain, C. Lee, Maximum a-posteriori estimation for multivariate gaussian mixture observations of markov chains, IEEE Trans. on Spch. & Aud. Proc. 2 (1994) 291–298.

[30] U. Nallasamy, F. Metze, T. Schultz, Active learning for accent adaptation in automatic speech recognition, in: IEEE Workshop on Spoken Language Technology *Miami, Florida, USA*, 2012, pp. 360–365.

[31] H. Kamper, F. J. Muamba Mukanya, T. Niesler, Multi-accent acoustic modelling of South African English, Speech Communication 54 (6) (2012) 801–813.

[32] C. Leggetter, P. C. Woodland, Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models, Computer Speech and Language 9 (2) (1995) 171–185.

[33] M. Benzeghiba, R. de Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, C. Wellekens, Automatic speech recognition and speech variability: A review, Speech Communication 49 (10-11) (2007) 763–786.

[34] Y. R. Oh, H. K. Kim, MLLR/MAP adaptation using pronunciation variation for non-native speech recognition, in: Proc. ASRU 2009, IEEE Automatic Speech Recognition and Understanding Workshop *Merano, Italy*, 2009, pp. 216–221.

[35] K. Kirchhoff, D. Vergyri, Cross-dialectal acoustic data sharing for Arabic speech recognition., in: Proc. IEEE-ICASSP, *Montreal, Quebec, Canada*, 2004, pp. 765–768.

[36] M. Najafian, Acoustic model selection for recognition of regional accented speech, Ph.D. thesis, University of Birmingham (2016).

[37] D. Povey, L. Burget, M. Agarwal, P. Akyazi, F. Kai, A. Ghoshal, O. Glembek, N. Goel, M. Karafiát, A. Rastrow, et al., The subspace Gaussian mixture model : A structured model for speech recognition, Computer Speech & Language 25 (2) (2011) 404–439.

[38] P. Motlícek, P. N. Garner, N. Kim, J. Cho, Accent adaptation using Subspace Gaussian Mixture Models, in: Proc. IEEE-ICASSP, *Vancouver, BC, Canada*, 2013, pp. 7170–7174.

[39] Y. Huang, D. Yu, C. Liu, Y. Gong, Multi-accent deep neural network acoustic model with accent-specific top layer using the KLD-regularized model adaptation, in: Proc. Interspeech, *Singapore*, 2014, pp. 2977–2981.

[40] X. Chen, J. Cheng, Deep neural network acoustic modeling for native and non-native Mandarin speech recognition, in: International Symposium on Chinese Spoken Language Processing, 2014, pp. 6–9.

[41] M. Chen, Z. Yang, J. Liang, Y. Li, W. Liu, Improving deep neural networks based multi-accent mandarin speech recognition using i-vectors and accent-specific top layer, in: Proc. Interspeech, *Dresden, Germany*, 2015.

[42] P. Kenny, G. Boulianne, P. Dumouchel, Eigenvoice modeling with sparse training data, IEEE Trans. on Spch. & Aud. Proc. 13 (3) (2005) 345–354.

[43] Y. Lei, N. Scheffer, L. Ferrer, M. McLaren, A novel scheme for speaker recognition using a phonetically-aware deep neural network, in: Proc. IEEE-ICASSP, *Florence, Italy*, 2014, pp. 1714–1718.

[44] S. O. Sadjadi, M. Slaney, L. Heck, MSR identity toolbox v1.0: A MATLAB toolbox for speaker recognition research, IEEE Speech and Language Processing Technical Committee Newsletter, November 2013.

[45] X. Zeng, J. Yang, D. Xu, Approaches to language identification using Gaussian mixture model and linear discriminant analysis, in: Intelligent Information Technology Application Workshops, 2008. IITAW'08. International Symposium on, IEEE, 2008, pp. 1109–1112.

[46] L.-F. Zhai, M.-H. Siu, X. Yang, H. Gish, Discriminatively trained language models using support vector machines for language identification, in: Proc. Odyssey'06, The Speaker and Language Recognition Workshop *San Juan, Puerto Rico*, 2006, pp. 1–6.

[47] C. Huang, T. Chen, E. Chang, Accent issues in large vocabulary continuous speech recognition, International Journal of Speech Technology 7 (2-3) (2004) 141–153.

[48] R. A. Fisher, The use of multiple measurements in taxonomic problems, Annals of Eugenics 7 (2) (1936) 179–188.

[49] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, K. Veselý, The Kaldi speech recognition toolkit, in: Proc. ASRU 2011, IEEE Automatic Speech Recognition and Understanding Workshop *Waikoloa, Hawaii*, 2011.

[50] A. Robinson, The British English Example Pronunciation (BEEP) dictionary, ftp://svrftp.eng.cam.ac.uk/pub/comp.speech/dictionaries/beep.tar.gz.

[51] G. E. Dahl, D. Yu, L. Deng, A. Acero, Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition, IEEE Transactions on Audio, Speech & Language Processing 20 (1) (2012) 30–42.

[52] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland, The HTK Book (ver. 3.2), Cambridge University Engineering Department, 2002.

[53] P. Swietojanski, S. Renals, Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models, in: IEEE Workshop on Spoken Language Technology *South Lake Tahoe, Nevada, USA*, 2014.

[54] J.-T. Huang, J. Li, D. Yu, L. Deng, Y. Gong, Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers, in: Proc. IEEE-ICASSP, *Vancouver, BC, Canada*, 2013.

[55] N. T. Vu, D. Imseng, D. Povey, P. Motlicek, T. Schultz, H. Bourlard, Multilingual deep neural network based acoustic modeling for rapid language adaptation, in: Proc. IEEE-ICASSP, *Florence, Italy*, 2014.

[56] S. Mirsamadi, J. H. L. Hansen, A study on deep neural network acoustic model adaptation for robust far-field speech recognition, in: Proc. Interspeech, *Dresden, Germany*, 2015.