Acoustic model selection using limited data for accent robust speech recognition

Najafian, Maryam; Safavi, Saeid; Hanani, Abualsoud; Russell, Martin

ACOUSTIC MODEL SELECTION USING LIMITED DATA FOR ACCENT ROBUST SPEECH RECOGNITION

Maryam Najafian[1], *Saeid Safavi*[1], *Abualsoud Hanani*[2], *Martin Russell*[1]

[1]School of EECE
University of Birmingham
Birmingham, UK
mxn978,sxs796,m.j.russell@bham.ac.uk

[2]Faculty of Information Technology
Birzeit University
West Bank, Palestine
ahanani@birzeit.edu

ABSTRACT

This paper investigates techniques to compensate for the effects of regional accents of British English on automatic speech recognition (ASR) performance. Given a small amount of speech from a new speaker, is it better to apply speaker adaptation, or to use accent identification (AID) to identify the speaker's accent followed by accent-dependent ASR? Three approaches to accent-dependent modelling are investigated: using the 'correct' accent model, choosing a model using supervised (ACCDIST-based) accent identification (AID), and building a model using data from neighbouring speakers in 'AID space'. All of the methods outperform the accent-independent model, with relative reductions in ASR error rate of up to 44%. Using on average 43s of speech to identify an appropriate accent-dependent model outperforms using it for supervised speaker-adaptation, by 7%.

Index Terms— speech recognition, acoustic data selection, accent identification

1. INTRODUCTION

A major limitation of hidden Markov model (HMM) based approaches to ASR is the difficulty of adapting to new speaker populations, because of the need for a significant quantity of representative speech data for model parameter adaptation. One approach is to try to exploit predictable, systematic variations in speech that characterise the population. Gender and accent have been identified as the primary sources of variation in speech [2]. Although the acoustic components of ASR systems often factor out gender, accent has proved difficult.

In [24], Wells defines 'accent of English' 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". This differentiates accent from dialect, which includes the use of words or phrases that are characteristic of that community. It includes varieties of English spoken as a first language in different countries (for example, US vs Australian English), geographical variations within a country, and patterns of pronunciation associated with particular social or ethnic groups.

Regional accents of British English are associated with five broad geographical regions: the North and South of England, Scotland, Wales and Ireland. The South of England can be further divided into London, the surrounding 'Home Counties', South-West and East Anglia, and the North into the Midlands, the 'mid-North', and the 'far-North' [24]. For example, when a speaker from Yorkshire in the North of England pronounces 'bath' with the same vowel quality as 'cat' rather than 'cart' he or she is exhibiting a Yorkshire (or at least north of England) accent. In this paper it is shown that, using an ASR system trained on the WSJCAM0 corpus of British English speech [25], error rates can be up to seven times higher for accented speech than for standard English.

The remainder of the paper is concerned with adaptation to a new user's regional accented British English speech, given minimal speaker-dependent training material. The focus is on acoustic, rather than pronunciation, modelling (although a complete solution will clearly involve both). Previous research has shown that, using 43s of speech, an individual's accent can be determined with 95% accuracy with supervised accent identification (AID) (using ACCDIST) [10] . Thus, a possible solution is to apply AID and then use an appropriate accent-dependent ASR model.

Each of these approaches treat regional accents as well-defined, disjoint phenomena with clear boundaries, whereas in reality this is not the case. Individuals who were born in the same region and have lived there for all of their lives, can still exhibit quite different patterns of pronunciation, and most users will have lived in several different locations during their lifetime. There is clearly considerable variation within an accent group and near 'accent boundaries' there may be individuals whose speech exhibits patterns of pronunciation associated with several regional accents. This is likely to be typical of individuals who have lived in many different geographical regions. This is the motivation for the final techniques that are investigated. The metrics employed in our AID systems are used to identify the set of N speakers who are 'closest in accent space' to the new speaker, again using just 43s of speech. All of the data associated with these N speakers is then used to create an ASR model.

This raises a number of questions. Given limited data from the test speaker,is it better to use that training sample for supervised speaker adaptation, using maximum likelihood linear regression (MLLR) [16] or to use that data for AID and identify a suitable accent-dependent ASR system? Is it better to use the data from neighbouring speakers in 'AID space', the 'correct' accent of the user (if it is known) or the result of AID to build a suitable acoustic model for ASR?

A relative reduction in error rate of 44% is obtained for accent-dependent models compared with the baseline system. It is also shown that using the 43s of speech to identify an appropriate accent-dependent model outperforms using it for supervised speaker-adaptation, by 7%.

2. PREVIOUS WORK

In the ASR literature 'accent adaptation' addresses a range of problems caused by 'accent' variations. Considerable research has been reported but comparisons are difficult because of this diversity. Approaches include accent-specific pronunciation adaptation (for example, [8, 13, 23]), multi accent and accent-specific acoustic modelling (e.g. [15]), accent-specific polyphone decision tree

(e.g. [19, 21]), knowledge- and data-driven acoustic model adaptation (e.g. [3, 4, 9]), feature based adaptation (e.g. [6, 7]), the use of accent discriminative acoustic features (e.g. [27]), acoustic data selection from existing corpora (e.g [1, 22]), Kullback-Leibler divergence-based HMM (e.g. [14]) and Subspace Gaussian Mixture Model (SGMM) (e.g. [18]) acoustic model adaptation. It is also possible that new approaches to ASR based on Deep Neural Networks (DNN) (e.g. [17, 20]) will go some way towards accommodating accent-related variation.

3. THE ABI SPEECH CORPUS

The Accents of the British Isles (ABI) speech corpus [5] represents 13 different regional accents of the British Isles, and standard (southern) British English (sse). The sse speakers were selected by a phonetician. ABI corpus, was recorded on location in the 13 regions listed in Table 1 and contains speech from 285 subjects. For each regional accent, 20 people (normally 10 women and 10 men) were recorded around 15 minutes of read speech. The subjects were born in the region and had lived there for all of their lives. Each subject read the same 20 prompt texts. The experiments in this paper focus on a subset of these texts, namely the 'short passages' (SPA), 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'. The corresponding 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 (three- or four-word) 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)

ABI code	Location	Broad accent
brm	Birmingham	North, Midlands
crn	Truro, Cornwall	South, South West
ean	Lowestoft, East Anglia	South, East Anglia
eyk	Hull, East Yorkshire	North, Mid-North
gla	Glasgow, Scotland	Scotland
ilo	Inner London	South, London
lan	Burnley, Lancashire	North, Mid-North
lvp	Liverpool, NW Eng.	North, Mid-North
ncl	Newcastle, Tyneside	North, Far-North
nwa	Denbigh, N Wales	Wales
roi	Dublin, Ulster	Ireland
shl	Elgin, Scottish Highlands	Scotland
sse	Standard Southern English	South
uls	Belfast, Ulster	Ireland

Table 1. *Accents represented in the ABI Corpus.*

4. REGIONAL ACCENT IDENTIFICATION

Three approaches to accent-dependent modelling are investigated: using the 'correct' accent model, choosing a model using AID, and building a model using data from neighbouring speakers in 'AID space'.

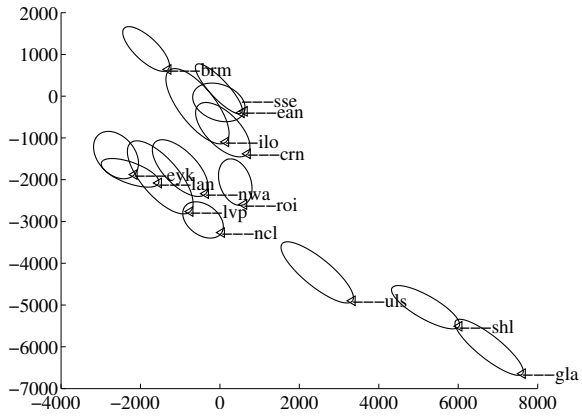

Fig. 1. *Visualization of the ACCDIST feature space*

For the purpose of this paper, ABI speakers were divided into three subsets; two with 93 and one with 94 speakers. Gender and accent were distributed equally in each subset. A 'jack-knife' procedure was used in which two subsets were used for training and the remaining subset for testing. This procedure was repeated three times with different training and test sets, so that each ABI speaker was used for testing, and no speaker appeared simultaneously in the training and test sets.

4.1. Supervised AID

The supervised AID system is based on the ACCDIST measure [12]. ACCDIST exploits the fact that British English accents are characterized by similarities and differences between the realizations of vowels in specific words. Our system differs from that described in [12], in that our classifier is based on Support Vector Machines (SVMs) rather than correlation distance, and uses tri-phone rather than word contexts. A transcription of each SPA recording was force-aligned with the speech data, and the most common vowel tri-phones were found. Each occurrence of a vowel tri-phone is split into two halves by time, and the average feature vectors (MFCCs 0 to 18 plus energy) plus duration for each half are concatenated into a 40-dimensional vector. For repeated tri-phones the average of these 40-dimensional vectors was used. Distances are calculated between vectors from different tri-phones using Euclidean distance, and stored in a distance 'super-vector'. A SVM is built for each accent by labeling the distance super-vectors of that accent as the target class and the remaining super-vectors as the background class. A test utterance super-vector is evaluated against every accent model. The SVMs used the correlation distance kernel. This accent recognition system is described fully in [10].

4.2. Visualisation

Our AID system maps an utterance into a 5253 dimensional super-vector space for classification. To obtain insight into how AID works, this space can be visualised by projecting it onto a suitable 2-dimensional subspace. This suggests linear discriminant analysis (LDA), but due to the small sample size ($N = 145$) and high dimensionality ($D = 5253$), it is not possible to invert the within-class covariance matrix. A solution is to use principal components analysis (PCA) to reduce the dimensionality of the data to a new value n, chosen empirically such that $C \leq n \leq D - C$, where $C = 14$ is the number of classes [11] and then apply LDA. Since $N \ll D$ we use EM-PCA [28] instead of PCA.

Fig. 1 shows mean values and 1-standard-deviation contours for each accent in this 2-dimensional projection of the ACCDIST-SVM super-vector space. The figure shows 3 clusters, corresponding to northern England, southern England and Scotland, but there is no separate cluster for the Irish accents. The proximity of Belfast (uls) to the Scotish accents (gla and shl) rather than Dublin (roi) may reflect close historic ties between Glasgow and Belfast. The North Wales (nwa) recordings were made in Denbigh, which is close to Liverpool, and this explains their location in Fig. 1. Unexpected features of Fig. 1 include the grouping of Birmingham (brm) with the southern English accents, and the positioning of the Dublin (roi) data amongst the English accents.

4.3. AID performance

For our experiments the speakers were partitioned into three approximately equal sized subsets and a three-way cross-validation procedure was applied. Gender and accent were distributed equally in each subset and two subsets were used for training and the remaining subset for testing, so that each ABI-1 speaker was used for testing, and no speaker appeared simultaneously in the training and test sets. In this way an AID result is available based on the SPA recording for each of the 285 ABI subjects (average duration 43.2s). The overall AID error rates are 4.82% [10].

5. AUTOMATIC SPEECH RECOGNITION

5.1. Baseline speech recognition system

Our baseline British English speech recognizer was built using HTK [26]. It is a phone-decision tree tied tri-phone HMM based system with 5500 tied states, each associated with an 8 component Gaussian Mixture Model (GMM). It was trained on the SI training set (92 speakers, 7861 utterances) of the WSJCAM0 corpus of read British English speech [25]. The feature vectors comprise MFCCs 0 to 12 plus their velocity and acceleration parameters. We used the British English Example Pronunciations (BEEP) dictionary [25], extended to include all of the words in the ABI corpus. The experiments reported in this paper use a weighted combination of the 5k WSJ0 bigram language model and a bigram language model based on the ABI corpus(excluding the test data), so that for a given bigram b, $P_{comb}(b) = \lambda P_{ABI}(b) + (1-\lambda)P_{WSJ0}(b)$. The choice of $\lambda \in [0,1]$ was determined empirically as 0.175, so that the bigram probabilities are strongly biased towards WSJ0. With this bigram language model we achieve similar error rates of 10.4% on the WSJCAM0 test set and 10% on the ABI sse test set. The same dictionary and grammar was used in all experiments in order to purely analyse the effect of using different acoustic models on the ASR performance.

5.2. Adaptation

5.2.1. Supervised speaker adaptation

For each speaker we conducted supervised (correct transcription) MLLR speaker adaptation with 48.1s (SPB), 101.5s (SPB+SPC), 136s (SPB+SPC+'Short phrases') and 221s (SPB+SPC+ 'Short phrases'+'Short sentences') of speaker-dependent data (Section 3).

5.2.2. Supervised Accent adaptation

For each subject in the ABI corpus, the SPA recording (section 3) was used as test data, and a gender- and accent-dependent model was created by applying supervised MLLR accent adaptation to the baseline WSJCAM0 system. Adaptation used the SPB, SPC, 'short sentences' and 'short phrases' (section 3) data from 9 other subjects (on average) with the same gender and accent as the test speaker (approximately 31.5 minutes of adaptation speech).

6. EXPERIMENTS

All of the following speech recognition experiments are conducted on the SPA data from each of the speakers in the ABI corpus. Hence the content of each test file corresponds to the same text. The optimal values of experiment parameters (e.g. MLLR regression class threshold) were obtained empirically using cross-validation.

6.1. Baseline experiment on the ABI corpus (B0)

We used the baseline WSJCAM0 speech recognition system with the extended WSJ0 5k bigram grammar to recognise the SPA recording for each subject in the ABI corpus. The purpose of this experiment was to measure the effect of regional accent on the performance of a 'standard' British English ASR system.

6.2. SSE adaptation (B1)

We were concerned that performance improvements resulting from accent adaptation might actually be due to adaptation to the ABI task. Since the recordings in WSJCAM0 are already close to sse, by adapting the baseline system using the ABI sse adaptation data and then testing on all of the ABI accents we can measure the amount of task adaptation. This is the purpose of B1.

6.3. Accent-dependent models — 'correct' accent (B2)

In these experiments we use the 'correct' accent of each ABI subject to apply the correct accent-dependent models. Accent adaptation of the baseline WSJCAM0 system is described in Section 5.2.2.

6.4. Supervised speaker adaptation (S0)

The accent-dependent ASR experiments based on AID that follow use AID results from 43.2s of speech. This raises two questions: (1) Is it better to use this speech for AID, so that an accent-dependent model can be selected, or directly for speaker adaptation? (2) How much speech from an individual is needed to achieve results from speaker adaptation that are comparable with the use of an accent-dependent model? To answer these questions we conducted speaker adaptation experiments for each ABI subject, using supervised (S0) MLLR adaptation (Section 5.2.1).

6.5. Accent-dependent models chosen using supervised AID (S1)

In these experiments, for each subject speech recognition is performed using the accent-adapted model (Section 5.2.2) corresponding to the result of AID for that speaker, using supervised ACCDIST-based AID (S1).

6.6. ASR Model based on N closest speakers in supervised AID space (S2)

In (S2), each ABI speaker s, is represented as an ACCDIST super-vector V_s (Section 4.1). Given a test speaker s the correlation $C(V_s, V_t)$ is calculated between V_s and V_t for each ABI subject t, and the N speakers $t_1, ..., t_N$ for which the correlation $C(V_s, V_{t_n})$ is largest are identified. A new model is then constructed by adapting the baseline WSJCAM0 model using the adaptation data from these N speakers. The values $N = 9$ (S2, 33.1 minutes of adaptation speech) was chosen using cross-validation.

7. RESULTS

Detailed results are shown in Fig. 2 and a summary is given in Table 2. The percentage word error rates (%WER) for experiments B0, B1 and B2 are included in the figure. The accents are ordered on the horizontal axes according to the baseline B0 results.

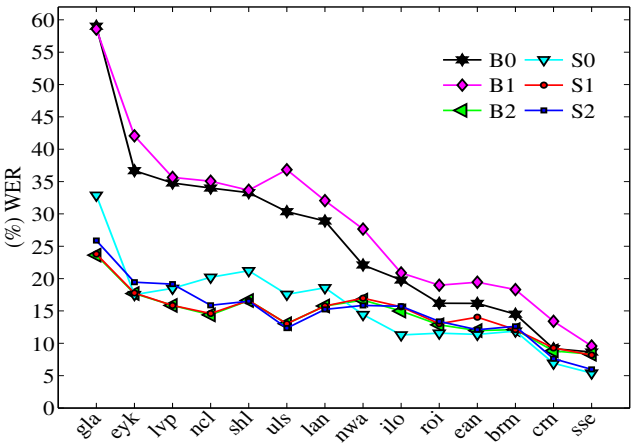

Fig. 2. *Comparison of results (supervised adaptation)*

8. DISCUSSION

The WSJCAM0 database consists mostly of standard British English (sse) speakers, so we expect, and find, that the best performance of the baseline WSJCAM0 system (B0) is for (sse) (8.7 %WER). The poorest (59 %WER) is for the Glasgow accent (gla), which is also the furthest from sse in Fig. 1. Error rates tend to be higher for the northern English accents, and lower for the southern accents, which is also consistent with Fig. 1. The word error rates for the Scottish Highland (shl) and Ulster (uls) accents are grouped with the northern English accents, and are not as poor as one might predict from Fig. 1.

The graph labelled B1 in Fig. 2 shows the result of MLLR adaptation using the sse data. Recall that the purpose of this experiment is to show that subsequent performance gains obtained by adapting to accented data in the ABI corpus result from accent, and not task adaptation. Overall, performance is 10% poorer than the baseline. As one would expect, sse performance is almost unchanged. This gives confidence that the improvements reported below are indeed due to accent adaptation.

The results of adapting to the 'correct' accent of the speaker is shown in graph labelled B2 in the figure. The relative reduction in error rate varies between 60% (gla) and 4% (sse & crn), with an average reduction of 44%.

Graph S0 in Fig. 2 shows results for supervised speaker adaptation of the baseline (B0) with 48s of speech. The reduction in error rate relative to the baseline (B0) for supervised speaker adaptation is 39%. For supervised speaker adaptation (S0) a small improvement is observed for the *easier* accents (up to lan) but poorer performance is obtained for the more *difficult* accents.

The result of choosing the accent model returned by AID, rather than the 'correct' accent, is shown in the graph labeled S1 (supervised AID). Since the supervised AID error rate is less than 5% one would expect the performance in S1 to be similar to B2, and this is the case.

The final graph (S2) is for adaptation using all data from the N closest ABI speakers to the test speaker, according to the correlations between their ACCDIST super-vectors (S2). The results are similar to B1 (adaptation to the 'correct' accent) and S1(adaptation to the supervised AID accent). This is disappointing. By definition, an ABI speaker's 'correct' accent is determined by the fact that he or she has lived all of their life in the the location where they were born. However, for some of the ABI accents there are, subjectively, large differences between speakers, for example due to economic and social factors. Hence one might expect that using AID to choose an accent-dependent model (S0), would result in better performance. Further, if a speaker is close to the boundary of an accent region in AID space, one might expect that building a model from the speech of the closest other speakers in AID space would lead to an advantage. However, there is no evidence for this in the current study.

Exp	%WER	Exp	%WER
B0	26.0	S0	15.9
B1	28.7	S1	14.8
B2	14.7	S2	15.6

Table 2. *Summary of results (Word Error Rate (%WER))*

Exp	S0				S1	S2
Utterance	48s	102s	136s	221s	43s	43s
%WER	15.88	14.17	13.81	12.30	14.80	15.60

Table 3. *Comparison of of results (%WER) for speaker adaptation, AID based accent-dependent model selection and N closest speakers data selection*

The result for speaker adaptation (S0) suggest that for southern English accents that are *closer* to sse, speaker adaptation performs better than accent adaptation. However, as the accent moves further from sse the opposite is true.

Finally, is it better to use a test speaker's data for speaker adaptation, or for AID-based accent adaptation? Table 3 compares the performance of supervised AID adaptation using 43.2s of speech (S0) and speaker adaptation using up to 221s of speech. In this case, the data required to achieve a similar result to AID adaptation with speaker adaptation is greater by a factor of 1.1.

9. CONCLUSIONS

We showed that the notion of 'regional accent' can be used explicitly to improve ASR performance. Given an average of 43s of data from a new speaker, three alternative approaches to supervised accent-dependent modelling were investigated, namely using the acoustic model for the 'correct' accent, using the acoustic model for the accent chosen by a supervised AID system, and building a model using data from the N closest speakers in the supervised 'AID feature spaces'.

All three methods give similar performance, which is significantly better than the performance obtained with the baseline, accent-independent model. The relative reduction in ASR error rate is 44% for accent-dependent models, compared with the baseline WSJCAM0 system. We also demonstrated that using the 43s of speech to identify an appropriate accent-dependent model using AID gives better performance than speaker adaptation.

In most practical applications, unsupervised adaptation approaches are preferred over supervised ones. Given small amount of speech from a new speaker, in our ongoing work we will investigate the changes in ASR performance caused by applying acoustic data selection using unsupervised AID instead of supervised AID, and we will show, how sensitive this result is to AID accuracy. Also, the choice of unsupervised AID based acoustic model selection, unsupervised speaker adaptation or the combination of both will be investigated in our future work.

REFERENCES

[1] Bacchiani, M., "Rapid adaptation for mobile speech applications", Proc. IEEE ICASSP 2013.

[2] Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L. et al. "Automatic speech recognition and speech variability: A review." Speech Communication, Vol. 49, no. 10: 763-786, 2007.

[3] Chao Huang, Tao Chen and Eric Chang, "Accent Issues in Large Vocabulary Continuous Speech Recognition", Int. J. of Speech Technology, Vol. 7, 141-153, 2004.

[4] Cincarek, T., Gruhn, R., and Nakamura, S.. "Speech recognition for multiple non-native accent groups with speaker-group-dependent acoustic models." Proc. ICSLP 2004.

[5] D'Arcy, S., Russell, M., Browning, S. and Tomlinson, M., "The Accents of the British Isles (ABI) Corpus". Proc. Modelisations pour l'Identification des Langues, MIDL Paris, 115-119, 2005.

[6] Deng, Y., Li, X., Kwan, C., Raj, B., and Stern, R. "Continuous feature adaptation for non-native speech recognition." Int. J. of Signal Processing, Vol. 3, no.1, 2006.

[7] Dupont, S., Deroo, O. and Poitoux, S., "Feature Extraction And Acoustic Modeling: An Approach for Improved Generalization Across Languages and Accents" Proc. ASRU, Mexico, 29-34, 2005.

[8] Goronzy, S. "Robust adaptation to non-native accents in automatic speech recognition.", LNCS 2560, Springer, 2002.

[9] Gales, M. J. "Cluster adaptive training for speech recognition." Proc. ICSLP'98, 1998.

[10] Hanani, A., Russell, M.J., Carey, M.J., "Human and computer recognition of regional accents and ethnic groups from British English speech", Computer Speech and Language, Vol. 27, 2013.

[11] Huang, R, Liu Q, and Ma, S, "Solving the small sample size problem of LDA", Proc. ICPR'02, 2002.

[12] Huckvale, M. "ACCDIST: An accent similarity metric for accent recognition and diagnosis", in Speech Classification II, Lect. Notes in Comp. Sci., Vol. 441, 2007.

[13] Humphries, J.J.,Woodland, P.C., "Using accent-specific pronunciation modelling for improved large vocabulary continuous speech recognition", Proc. Eurospeech'97, 1997.

[14] Imseng, D., Rasipuram, R. and Magimai-Doss, M., "Fast and flexible Kullback-Leibler divergence based acoustic modeling for non-native speech recognition", Proc. ASRU, 348-353, 2011.

[15] Kamper, H., Jeje Muamba Mukanya, F. and Niesler, T. "Multi-accent acoustic modelling of South African English", Speech Comm., Vol. 54, 6, pp. 801-813, 2012.

[16] Leggetter, C.J. and Woodland, P.C., "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models", Computer Speech and Language, Vol. 9, Issue 2, 171-185, 1995.

[17] Mohamed, A., Dahl, G. E. and Hinton, G. E. "Acoustic Modeling using Deep Belief Networks", IEEE Trans. on Audio, Speech, and Language Processing, Vol. 20, 1., pp 14-22 2012.

[18] Motlicek, P., Garner, P.N., Namhoon, K., and Jeongmi, C. "Accent adaptation using Subspace Gaussian Mixture Models", Proc. IEEE ICASSP 2013, 7170-7174, May 2013.

[19] Nallasamy, U., Metze, F., Schultz, T. "Enhanced polyphone decision tree adaptation for accented speech recognition", Proc. INTERSPEECH 2012, 1902-1905.

[20] Ngoc Thang Vu and Tanja Schultz, "Multilingual Multilayer Perceptron For Rapid Language Adaptation Between and Across Language Families", Proc. Interspeech 2013, 515-519, Lyon 2013.

[21] Schultz, T. and Waibel, A. "Polyphone Decision Tree Specialization for Language Adaptation", Proc. ICASSP, Istanbul 2000.

[22] Siohan, O. and Bacchiani, M., "iVector-based Acoustic Data Selection", Proc. Interspeech 2013, Lyon 2013.

[23] Tjalve, M., Huckvale, M., "Pronunciation variation modelling using accent features", Proc. Interspeech 2005, Lisbon, 2005.

[24] Wells, J.C., "Accents of English, Volume 2: The British Isles". Cambridge University Press, 1982.

[25] Robinson, T., Fransen J., Pye D., Foote J., Renals S., "WSJ-CAM0: A British English speech corpus for large vocabulary continuous speech recognitio", Proc. IEEE ICASSP 1995, 81-84, Detroit, 1995.

[26] Young, S.J., Evermann, G., Kershaw, D.,Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V. and Woodland, P., "The HTK book (version 3.2)", Cambridge University Engineering Dept., 2002.

[27] Zheng, Y, Sproat, R., Gu, L., Shafran, I., Zhou, H., Su, Y., Jurafsky, D., Starr, R., Yoon, S-Y., "Accent Detection and Speech Recognition for Shanghai-Accented Mandarin", Proc. Interspeech 2005, Lisbon, 2005.

[28] Bishop, C.M., "Pattern recognition and machine learning", Springer New York, 2006.