Unsupervised Model Selection for Recognition of Regional Accented Speech

Maryam Najafian[1], *Andrea DeMarco*[2], *Stephen Cox*[2], *Martin Russell*[1]

[1]School of EECE, University of Birmingham, Birmingham, UK
[2]School of Computing Sciences, University of East Anglia, Norwich, UK
mxn978,m.j.russell@bham.ac.uk, a.de-marco,s.j.cox@uea.ac.uk

10.21437/Interspeech.2014-495

Abstract

This paper is concerned with automatic speech recognition (ASR) for accented speech. Given a small amount of speech from a new speaker, is it better to apply speaker adaptation to the baseline, or to use accent identification (AID) to identify the speaker's accent and select an accent-dependent acoustic model? Three accent-based model selection methods are investigated: using the 'true' accent model, and unsupervised model selection using i-Vector and phonotactic-based AID. All three methods outperform the unadapted baseline. Most significantly, AID-based model selection using 43s of speech performs better than unsupervised speaker adaptation, even if the latter uses five times more adaptation data. Combining unsupervised AID-based model selection and speaker adaptation gives an average relative reduction in ASR error rate of up to 47%.

Index Terms: speech recognition, acoustic model selection, accent identification, i-Vector, phonotactics

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 [1]. Although the acoustic components of ASR systems often factor out gender, accent has proved difficult.

In [2], 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.

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). Although the work targets British English it is very likely that the techniques described are applicable to accented speech in other languages.

It has been shown that with 43s of speech an individual's accent can be determined with 95% and 81% accuracy with supervised (ACCDIST) [3] and unsupervised (i-Vectors) [4] AID, respectively. Thus, a possible solution is to use AID for accent-dependent ASR model selection. This raises a number of questions. Is it better to use the "true" accent of the user (if it is known) or the result of AID, and how sensitive is the resulting ASR accuracy to AID accuracy? Also, do significant differences result from using phonotactic or i-Vector based AID?

In this paper we demonstrate that by using state-of-the-art, text-independent, unsupervised AID methods, it is possible to select effective accent-specific models for ASR. Although the AID systems do not give perfect accent classification, they are reliable enough to improve ASR error rates by a wide margin. The distinct advantage, apart from improved ASR performance is that the AID step is unsupervised and does not require any transcription. Furthermore, the results obtained are in most cases on a par with adapting an ASR system based on knowing the "correct" accent for an utterance.

A relative reduction in error rate of 44% is obtained from unsupervised selection of accent-dependent models compared with the baseline system. Moreover, the resulting ASR accuracy is almost unaffected by the inherent error margin of the utilized AID systems. Finally, it is shown that unsupervised AID-based model selection outperforms unsupervised speaker adaptation, even if the latter is given five times more training data. The relative reduction in error rate goes up even further to 47% when AID-based model selection is combined with unsupervised speaker adaptation.

2. Previous work

Just as the term "accent" covers a range of phenomena, "accent adaptation" refers to a number of different of problems in the ASR literature. Considerable research has been reported but comparisons are difficult because of this diversity. Approaches include accent-specific pronunciation adaptation [5, 6, 7], accent-independent acoustic modelling incorporating multi-accent training [8], integration of accent into HMM decsion-tree tri-phone clustering [8], knowledge- and data-driven acoustic model adaptation [9, 10, 11], SGMM and tandem features [12] feature based adaptation [13], the use of accent discriminative acoustic features [14], and selection of relevant training material from existing corpora [15, 16]. It is also possible that new approaches to ASR based on Deep Neural Networks (DNN) [17, 18] will go some way towards accommodating accent-related variation.

3. The ABI speech corpus

The Accents of the British Isles (ABI) speech corpus [19] represents 13 different regional accents of the British Isles (Table 1), and standard (southern) British English (sse). For the purposes of the ABI corpus, regional accented speech was defined as speech of individuals who had lived in the region since birth. The sse speakers were selected by a phonetician. Each of 285 subjects read the same 20 prompt texts. The experiments

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

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

in this paper focus on a subset of these texts, namely the 'short passages' (SPA, SPB and SPC), the 'short sentences' and the 'short phrases'. These are described below:

- 'SPA', 'SPB' and 'SPC' are short paragraphs, of lengths 92, 92 and 107 words, respectively, which together form the accent-diagnostic "sailor passage". The recordings have average durations 43.2s, 48.1s and 53.4s.
- 'Short sentences'. These are 20 phonetically balanced sentences (e.g."Kangaroo Point overlooked the ocean"). They are a subset of the 200 Pre-Scribe B sentences (a version of the TIMIT sentences for British English), chosen to avoid some of the more 'difficult' of those sentences, whilst maintaining coverage (146 words, average duration 85.0s)
- 'Short phrases' are 18 phonetically rich short phrases (e.g."while we were away") containing English phonemes in particular contexts in as condensed form as possible (58 words, average duration 34.5s)

4. Automatic speech recognition

4.1. Baseline speech recognition system

Our baseline British English speech recognizer was built using HTK [20]. 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 [21]. The feature vectors comprise MFCCs 0 to 12, plus their velocity and acceleration parameters. We used the British English Example Pronunciations (BEEP) dictionary [21], 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, 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 were used in all experiments.

4.2. Adaptation

4.2.1. Unsupervised speaker adaptation

For each speaker we conducted unsupervised (transcription from WSJCAM0 baseline ASR) 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).

4.2.2. 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 with the same gender and accent as the test speaker (approximately 31.5 minutes of speech).

5. Regional accent identification

We perform model selection for ASR based on two AID methods - both of which are popular for language/dialect/accent identification. These AID methods do not require any transcription of the data. The i-Vector AID method is based on plain acoustic features, whilst the phonotactic AID method is based on acoustic-phonetic features. In the next subsections, we give an overview of how each system is constructed.

5.1. i-Vector AID

Our first AID system is based on i-Vectors, a technique introduced in [22] for speaker verification. This technique has also been proven to work well in language identification [23]. In previous work [4, 24], the i-Vector classification framework was also utilized for AID on the ABI corpus. An i-Vector AID classifier is based on a configuration determined by the size of the universal background model (UBM), the number of factor dimensions for the total variability subspace, as well as the various compensation methods to attenuate within-accent speaker variability. The results obtained in [4] suggest that the best results are not obtained by selecting a particular triple configuration, but rather to treat all possible configurations as 'weak classifiers', utilizing them in a fusion for a final classification.

The i-Vector AID classifier uses various UBM sizes (128, 256, 512, 1024) and different factor dimensions (100 to 400 in steps of 50). The different speaker-compensation methods are all linear projection methods: Linear Discriminant Analysis (LDA) [25], Regularized LDA (R-LDA) [26], Semi-supervised LDA (SDA) [27], and Neighbourhood Component Analysis (NCA) [28].

The i-Vectors are derived directly from the feature vectors over an entire utterance. Each feature vector is composed of 62 dimensions, which includes 13 MFCCs that have been normalized to a standard normal distribution with a three second time window, concatenated with the respective shifted-delta cepstra (SDC) features, calculated with a 7-1-3-7 parametrization. The feature vectors from training and test utterances are used to construct multiple i-Vectors per utterance, dependent on the i-Vector configuration. LDA classification is used for i-Vectors projected with LDA/R-LDA/SDA, whilst Euclidean 1-Nearest Neighbour classification is used with NCA projections. During testing, the i-Vectors of an utterance are evaluated by the different classifiers, and a Genetic Algorithm (GA)-optimised majority voting technique is used to give the final classification. The i-Vector AID system utilized in this work is described fully in [4].

5.2. Phonotactic AID

Our second AID system is based on parallel phone recognition followed by Language Modelling (PPRLM), another popular technique in language/accent/dialect identification [29, 30]. Our

implementation is similar to [3]. The idea is to exploit the fact that the rate of utilization of certain phonetic sequences can provide information about the underlying accent. There are three stages to this system:

5.2.1. Phone recognition

The speech signal is converted into a sequence of phones using either our baseline (section 4.1) or one of our 14 accent-adapted (section 4.2.2) ASR systems. This system differs from that described in [3] in that we use multiple ASR systems for different accents rather than different languages. Each of these systems is expected to perform better on accents that are similar to the one it was trained on and poorly on the rest. As with multiple-language PPRLM systems, the hope is that the different ASR systems will capture complimentary information that can subsequently be exploited using fusion. All systems use the same bi-gram phone-level language model, constructed using the WSJCAM0 SI and ABI training subsets.

5.2.2. Vectorization

For each n and each of our 15 ASR systems, we can, in principle, construct a separate n-gram-based AID system. Let D_n be the number of different n-grams that occur in the training data. For each utterance, a D_n dimensional n-gram probability vector is computed as in [31]. In these experiments we used values of $n \leq 5$. The final choice of $n = 4$ was made empirically using cross-validation.

5.2.3. SVM backend

Finally, for each n and each ASR system, a SVM n-gram language model is trained for each accent on the vectors from utterances in the training sets [32].

During recognition, a phone sequence is extracted from the test utterance using the ASR system. The resulting D_n dimensional probability vector is then evaluated with the various class SVMs to obtain a classification. In our system, the final result of PPRLM is obtained by fusing the outputs of 15 individual phonotactic systems (15 accent-dependent phone recognizers with 4th order n-gram SVM accent classifiers), using Brummer's multi-class linear logistic regression (LLR) toolkit [33].

5.3. AID performance

In our AID experiments all 285 ABI speakers were divided into three equal sized subsets. Gender and accent were distributed equally in each subset. A "jack-knife" training procedure was used in which two subsets were used for training and the remaining subset for testing, so no speaker appeared simultaneously in the training and test sets. SPA utterances from each ABI speaker was only used for testing and not for training. An AID result is available based on the SPA recording for each of the 285 ABI subjects. The overall AID error rates are 18.95% for i-Vector AID and 19.30% for phonotactic AID.

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 4.2.2.

6.4. Unsupervised (S0) speaker adaptation

The AID-based model-selection ASR experiments 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 unsupervised (S0) MLLR adaptation (Section 4.2.1).

6.5. Accent-dependent models chosen using i-Vector (A0) and phonotactic (A1) AID

In these experiments, for each subject speech recognition is performed using the accent-adapted model (Section 4.2.2) corresponding to the result of AID for that speaker, using either i-Vector-based AID (A0) or phonotactic-based AID (A1).

6.6. Accent-dependent models followed by unsupervised speaker adaptation - (BS) and (AS)

For each speaker, model selection is applied, based on either the "true" accent (BS) or the one determined by phonotactic AID (AS), to obtain an "accent adapted " acoustic model. Unsupervised MLLR speaker adaptation is then applied to that model and recognition is performed.

7. Results and discussion

A summary of results for all tested methods is given in Table 2. Surprisingly, although the AID error rates for i-Vector-based and phonotactic-based AID are quite high, the corresponding ASR %WERs are close to that obtained with the 'correct' accent model (B2). Specifically, the WER obtained with the 'true' accent model is 14.7%, compared with 15.2% for model selection using i-vector-based AID (A0) and 15.3% using phonotactic-based AID (A1).

Regarding speaker adaptation, Table 2 compares the performance of unsupervised model selection using 43.2s of speech and unsupervised speaker adaptation using up to 221s of speech. The performance obtained with unsupervised speaker adaptation is never as good as unsupervised model selection using AID, using either i-Vector or phonotactic AID, even if speaker adaptation is given five times more adaptation data.

A more detailed breakdown of results of individual systems is shown in Figure 1. The accents are ordered on the horizontal axes according to the baseline B0 results. As one would

Table 2: *Comparison of results for all experiments*

Experiment	B0	B1	B2	S0				A0	A1	BS	AS
Data from test speaker(s)	–	–	–	48.0	101.5	136.0	221.0	43.2	43.2	43.2	43.2
WER(%)	26.0	28.7	14.7	20.37	18.75	18.99	17.83	15.2	15.3	13.7	14.1

expect (assuming that WSJCAM0 is close to 'standard' British English), the best performance of the baseline WSJCAM0 system (B0) is for standard British English (sse) (8.7 %WER). The poorest (59 %WER) is for the Glasgow accent (gla). Error rates tend to be higher for the northern English accents, and lower for the southern accents. The word error rates for the Scottish Highland (shl) and Ulster (uls) accents are grouped with the northern English accents.

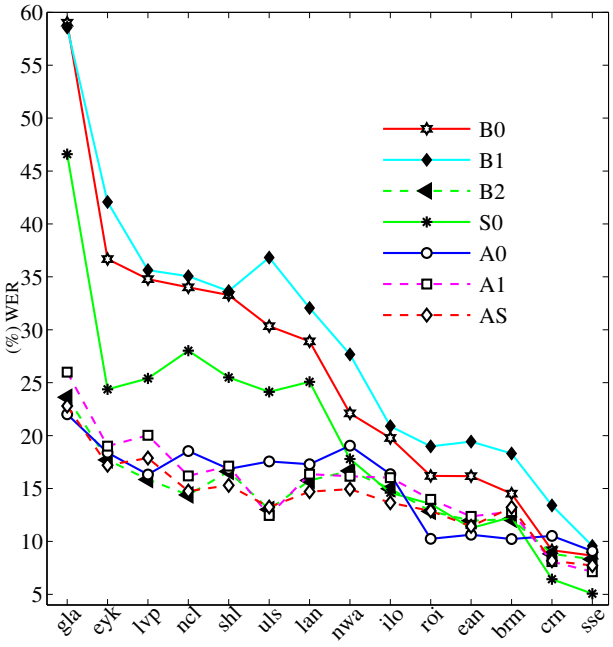

Figure 1: *Comparison of results by accent*

The graph labelled B1 in Figure 1 shows results for 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 selecting the accent-dependent ASR model corresponding to the speaker's "true" accent are shown in graph labelled B2 in the figures. The relative reduction in error rate varies between 60% (gla) and 4% (sse & crn), with an average reduction of 47%.

Graph S0 in Figure 1 shows results for unsupervised speaker adaptation of the baseline (B0) with 43s of speech. The relative reduction in error rate relative to the baseline (B0) for unsupervised speaker adaptation is 22%. It can be noted that there is an improvement in ASR performance across all accents when compared to the baseline (B0). However the relative improvement seems to be wider for the "easier" accents (up to lan), and gets narrower after that. Presumably, this is due to the poor baseline ASR performance on the more "difficult" accents. The results of choosing the accent model returned by AID, rather than the correct accent, are shown in the graphs labeled A0 (i-Vector AID) and A1 (phonotactic AID) and summarised in Table 2.

Looking at the results for speaker adaptation (S0), it has already been noted that for accents that are 'close' to sse speaker adaptation performs better that accent adaptation, but as the accent moves further from sse the opposite is true. This suggests that for speaker adaptation to work well, it is important that the acoustic model has already been adapted to the speaker's accent. We test this in experiments (BS and AS), which reduce error rates further. Interestingly, AS, based on phonotactic-AID followed by speaker adaptation performs better than B2, which uses the "correct" accent, despite the AID error margin for A1.

Finally, we assert that the comparisons made in these experiments are "fair". Although the underlying accent-dependent models have certainly been trained with far more data that is used for speaker adaptation, the key *practical* question is whether a good accent-dependent model can be reliably selected using the same small amount of data that is available for speaker adaptation. We are evaluating the consequences of using this small amount of data in different ways.

8. Conclusions

This paper investigates whether 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 accent-dependent ASR model selection are investigated, namely using the acoustic model for the speaker's "correct" accent and using the acoustic model for the accent chosen by two different unsupervised AID systems. In fact, 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 compared with the baseline WSJCAM0 system is 47% for models chosen according to the "correct" accents followed by speaker adaptation, and 42% and 41% for ASR models selected using i-Vector or phonotactic based AID, respectively. Interestingly, the performance obtained with AID model selection does not appear to be very sensitive to the accuracy of the AID system, with AID error rates ranging from 18.95% to 19.30% leading to similar ASR error rates.

Compared with unsupervised MLLR-based speaker adaptation, both AID-based model selection methods show substantial benefits. Both methods outperform unsupervised speaker adaptation, even if the latter uses five times more adaptation data.

These experiments demonstrate the utility of AID based model selection for practical ASR, where the unsupervised nature of AID is particularly advantageous. Given the apparent insensitivity of ASR performance to AID error rate, it will be interesting to apply the same methods to shorter enrollment utterances, since it would be desirable for an ASR system not to require around 40 seconds of data for reliable model selection. An alternative to model selection is to use AID measures to identify a set of known speakers that are acoustically and phonologically close to a new test speaker, and to estimate the parameters of a new model from these speakers' data or models.

9. References

[1] M. Benzeghiba, R. D. Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens, "Automatic speech recognition and speech variability: A review," *Speech Communication*, vol. 49, no. 1011, pp. 763 – 786, 2007.

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

[3] A. Hanani, M. J. Russell, and M. J. Carey, "Human and computer recognition of regional accents and ethnic groups from british english speech." *Computer Speech and Language*, vol. 27, no. 1, pp. 59–74, 2013.

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

[5] S. Goronzy, *Robust Adaptation to Non-Native Accents in Automatic Speech Recognition*, ser. Lecture Notes in Computer Science. Springer, 2002, vol. 2560.

[6] J. J. Humphries and P. C. Woodland, "Using accent-specific pronunciation modelling for improved large vocabulary continuous speech recognition." in *EUROSPEECH*, G. Kokkinakis, N. Fakotakis, and E. Dermatas, Eds. ISCA, 1997.

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

[8] H. Kamper, F. J. M. Mukanya, and T. Niesler, "Multi-accent acoustic modelling of south african english," *Speech Communication*, vol. 54, no. 6, pp. 801 – 813, 2012.

[9] T. Cincarek, R. Gruhn, and S. Nakamura, "Speech recognition for multiple non-native accent groups with speaker-group-dependent acoustic models." in *Interspeech*. ISCA, 2004.

[10] M. J. F. Gales, "Cluster adaptive training for speech recognition." in *ICSLP*. ISCA, 1998.

[11] C. Huang, T. Chen, and E. Chang, "Accent issues in large vocabulary continuous speech recognition," *International Journal of Speech Technology*, vol. 7, no. 2-3, pp. 141–153, 2004.

[12] P. Motlcek, D. Imseng, and P. N. Garner, "Crosslingual tandem-sgmm: exploiting out-of-language data for acoustic model and feature level adaptation." in *Interspeech*. ISCA, 2013, pp. 510–514.

[13] Y. Deng, X. Li, C. Kwan, B. Raj, and R. Stern, "Continuous feature adaptation for non-native speech recognition," *International Journal of Computer, Information Science and Engineering*, vol. 1, no. 6, pp. 164 – 171, 2007.

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

[15] M. Bacchiani, "Rapid adaptation for mobile speech applications." in *ICASSP*. IEEE, 2013, pp. 7903–7907.

[16] O. Siohan and M. Bacchiani, "ivector-based acoustic data selection." in *Interspeech*, F. Bimbot, C. Cerisara, C. Fougeron, G. Gravier, L. Lamel, F. Pellegrino, and P. Perrier, Eds. ISCA, 2013, pp. 657–661.

[17] A. Mohamed, G. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 14–22, Jan 2012.

[18] N. T. Vu and T. Schultz, "Multilingual multilayer perceptron for rapid language adaptation between and across language families." in *Interspeech*, F. Bimbot, C. Cerisara, C. Fougeron, G. Gravier, L. Lamel, F. Pellegrino, and P. Perrier, Eds. ISCA, 2013, pp. 515–519.

[19] S. D'Arcy, J. Russell, S. Browning, and M. Tomlinson, "The Accents of the British Isles (ABI) Corpus," in *Modelisations pour l'Identification des Langues. MIDL Paris*, 2005, pp. 115–119.

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

[21] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJ-CAM0: A British English speech corpus for large vocabulary continuous speech recognition," vol. 1, Detroit, 1995, pp. 81–84.

[22] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, May 2011.

[23] D. M. Gonzlez, O. Plchot, L. Burget, O. Glembek, and P. Matejka, "Language recognition in ivectors space." in *Interspeech*. ISCA, 2011, pp. 861–864.

[24] A. DeMarco and S. J. Cox, "Iterative classification of regional british accents in i-vector space," in *Symposium on Machine Learning in Speech and Language Processing (SIGML)*, 2012.

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

[26] J. H. Friedman, "Regularized Discriminant Analysis," *Journal of the American Statistical Association*, vol. 84, no. 405, pp. 165–175, 1989.

[27] D. Cai, X. He, and J. Han, "Semi-supervised discriminant analysis," in *Proc. Int. Conf. Computer Vision (ICCV'07)*, 2007.

[28] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov, "Neighbourhood components analysis," in *Advances in Neural Information Processing Systems 17*. MIT Press, 2004, pp. 513–520.

[29] P. Matejka, P. Schwarz, J. Cernock, and P. Chytil, "Phonotactic language identification using high quality phoneme recognition." in *Interspeech*. ISCA, 2005, pp. 2237–2240.

[30] M. A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 1, p. 31, 1996.

[31] W. M. Campbell, J. P. Campbell, D. A. Reynolds, D. A. Jones, and T. R. Leek, "Phonetic speaker recognition with support vector machines," in *in Advances in Neural Information Processing Systems*, 2004.

[32] L.-F. Zhai, M.-H. Siu, X. Yang, and H. Gish, "Discriminatively trained language models using support vector machines for language identification," in *Speaker and Language Recognition Workshop, 2006. IEEE Odyssey 2006: The*, June 2006, pp. 1–6.

[33] D. A. Van Leeuwen and N. Brummer, "Channel-dependent gmm and multi-class logistic regression models for language recognition," in *Speaker and Language Recognition Workshop, 2006. IEEE Odyssey 2006: The*. IEEE, 2006, pp. 1–8.