Accent reclassification and speech recognition of Afrikaans, Black and White South African English

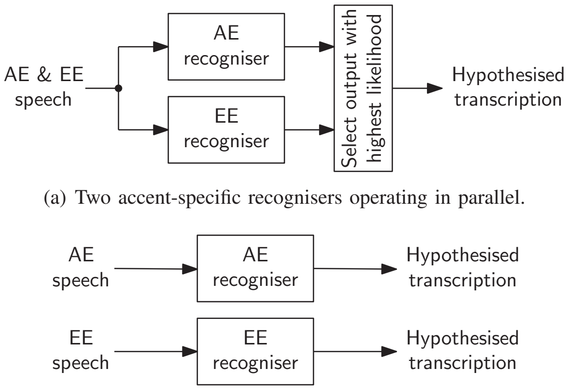
Herman Kamper and Thomas Niesler
Department of Electrical and Electronic Engineering
Stellenbosch University, South Africa
kamperh@sun.ac.za, trn@sun.ac.za

Abstract—**We consider the unsupervised re-assignment of training set accent labels and the associated effect on recognition accuracy of multi-accent automatic speech recognition (ASR) systems. Since training set accent labels are assigned by human annotators or are based on a speaker's mother-tongue or ethnicity, these may not be optimal for acoustic modelling purposes. We reclassify the accents of training set utterances for Afrikaans, Black and White accents of South African English using first-pass acoustic models trained using the original accent labels. We find that the proposed relabelling does not lead to improvements in speech recognition accuracy and that the best strategy remains the use of the originally labelled training data.**

I. Introduction

In South Africa, English is used predominantly by non-mother-tongue speakers resulting in a large number of English accents. Since these are in general not bound to geographic regions, ASR systems must be robust to multiple accents to ensure that speech-based automated services are accessible to the wider population. This motivates our aim to develop a system able to simultaneously recognise multiple accents of South African English (SAE) given limited speech resources.

In order to train multi-accent speech recognition systems, accent labels must be assigned to training set utterances. These labels are provided by human annotators or are determined based on the speaker's mother-tongue or ethnicity. However, previous research has indicated that accent labels assigned to some utterances in the accent-specific databases employed might be inappropriate [1]. In this paper we consider the iterative reclassification of the accent of training set utterances in an attempt to improve the labelling consistency of the training set. The speech recognition performance of models trained on the reclassified data is compared with that of models trained on the original data, as well as with systems in which training data are pooled across accents. Two acoustic modelling approaches are considered for the reclassification configuration: accent-specific acoustic modelling in which training data are kept totally separate for each accent, and multi-accent acoustic modelling in which selective cross-accent sharing of acoustic training data is allowed. For this investigation, we consider three of the five accents of SAE identified in the literature [2]: Afrikaans English (AE), Black South African English (BE), and White South African English (EE). These accents are considered in two pairs: AE+EE and BE+EE. The former pair represents two relatively similar accents while the latter represents two accents that are more different.



(a) Two accent-specific recognisers operating in parallel.

(b) Separate accent-specific recognisers for each accent.

Fig. 1. The two recognition configurations considered in [1] for recognition of Afrikaans English (AE) and White South African English (EE).

II. Accent Reclassification

We have previously considered speech recognition of AE and EE using a system of two accent-specific recognisers operating in parallel, as illustrated in Figure 1(a) [1]. It was shown that this configuration outperformed one in which accented speech was presented to the matching accent-specific recogniser, illustrated in Figure 1(b). When performing recognition by running multiple accent-specific recognisers in parallel and selecting the output with the highest associated likelihood, as in Figure 1(a), accent identification (AID) is performed implicitly during recognition. Thus, the finding that configuration (a) outperforms configuration (b) indicates that accent misclassifications do not always lead to deteriorated speech recognition accuracies. Instead, in some cases a different accent's recogniser produces a better accuracy than the recogniser of the correct accent. It appears then that the accent to which an utterance has been consigned in the training/test data is not always the most appropriate. In light of these results we proceed to reclassify the accent of each utterance in the training set using a set of first-pass acoustic models obtained using the original databases, and then to retrain the acoustic models using these newly assigned accent classifications.

This approach is illustrated in Figure 2. Using the unmodified training data, initial accent-specific hidden Markov models (HMMs) are obtained. These models are then used

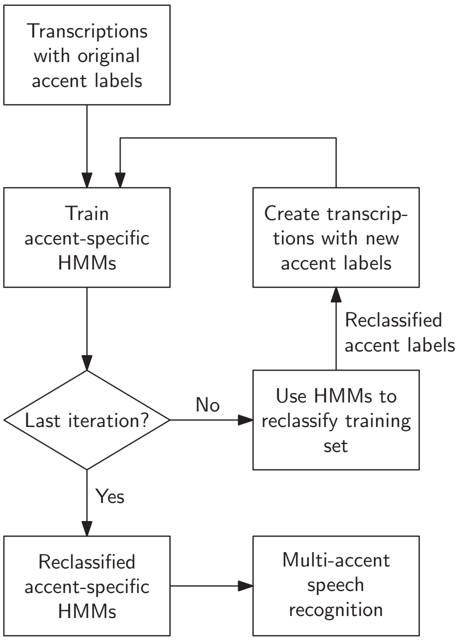

Transcriptions with original accent labels
Train accent-specific HMMs
Create transcriptions with new accent labels
Reclassified accent labels
Last iteration?
No
Use HMMs to reclassify training set
Yes
Reclassified accent-specific HMMs
Multi-accent speech recognition

Fig. 2. Reclassification of training data and subsequent retraining of acoustic models using the relabelled data.

TABLE I
TRAINING SETS FOR EACH ACCENT.

Accent	Speech (h)	No. of utterances	No. of speakers	Word tokens
AE	7.02	11 344	276	52 540
BE	5.45	7779	193	37 807
EE	5.95	9878	245	47 279

TABLE II
EVALUATION SETS FOR EACH ACCENT.

Accent	Speech (min)	No. of utterances	No. of speakers	Word tokens
AE	24.16	689	21	2913
BE	25.77	745	20	3100
EE	23.96	702	18	3059

TABLE III
WORD BIGRAM LANGUAGE MODEL PERPLEXITIES MEASURED ON THE EVALUATION SETS.

Accent	AE+EE LM perplexity	BE+EE LM perplexity
AE	23.48	-
BE	-	26.74
EE	23.85	24.04

to reclassify the training data. Transcriptions reflecting the new accent classifications are subsequently used to train new accent-specific HMMs. Multiple iterations of reclassification can be performed in this manner, although we only performed a single iteration. The proposed reclassification approach aims to compensate, during training, for the inexact assignment of accent labels to some utterances in the original data.

III. Speech Databases

A. Training and test sets

Our experiments were based on the African Speech Technology (AST) databases [3]. The databases consist of annotated telephone speech recorded over both mobile and fixed telephone networks and contain a mix of read and spontaneous speech. As part of the AST Project, five English accented speech databases were compiled, corresponding to the five South African accents of English identified in the literature [2]. In this research we made use of only the AE, BE and EE databases. These databases were transcribed both phonetically, using a common IPA-based phone set consisting of 50 phones, as well as orthographically. The assignment of a speaker's accent was guided by the speaker's first language and race.

The three databases were each divided into training, development and evaluation sets. As indicated in Tables I and II, the training sets each contain between 5.5 and 7 hours of speech from approximately 250 speakers, while the evaluation sets contain approximately 25 minutes from 20 speakers for each accent. The development sets were used only for the optimisation of the recognition parameters before final testing on the evaluation data. For the development and evaluation sets, the ratio of male to female speakers is approximately equal and all sets contain utterances from both land-line and mobile phones. There is no speaker-overlap between any of the sets. The average length of a test utterances is approximately 2 seconds.

B. Language models and pronunciation dictionaries

Using the SRILM toolkit [4], accent-independent bigram language models (LMs) were trained on the combined set of training transcriptions of all five accents in the AST databases (approximately 240k words). This was done based on initial experiments which indicated that, given the limited amount of LM training data, accent-independent LMs trained on the combination of all the English data in the AST databases outperformed accent-specific LMs trained individually on the training set transcriptions of each accent. Absolute discounting was used for the estimation of LM probabilities [5]. LM perplexities are shown in Table III. The AE+EE and BE+EE LMs differ only in their vocabularies, which was taken from the respective training sets of each accent pair. There are 21 605 and 20 644 bigram types for the AE+EE and BE+EE LMs respectively. Pronunciation dictionaries were obtained from the alignment between corresponding word and phone level training set transcriptions. Out-of-vocabulary rates are below 4% for all three accents when measured on the evaluation sets.

IV. Experimental Methodology

A. General setup

Speech recognition systems were developed using the HTK tools [6]. Speech audio data were parameterised as 13 Mel-frequency cepstral coefficients (MFCCs) with their first and second order derivatives to obtain 39 dimensional feature vectors. Cepstral mean normalisation (CMN) was applied on a per-utterance basis. The parameterised training sets were used to obtain three-state left-to-right single-mixture monophone HMMs with diagonal covariance matrices using embedded Baum-Welch re-estimation. These monophone models were then cloned and re-estimated to obtain initial cross-word triphone models which were subsequently clustered using decision-tree state clustering [7]. Clustering was followed by a further five iterations of re-estimation. Finally, the number of Gaussian mixtures per state was gradually increased, each increase being followed by a further five iterations of re-estimation, yielding diagonal-covariance cross-word triphone HMMs with three states per model and eight Gaussian mixtures per state.

B. Acoustic modelling

When performing multi-accent speech recognition by running several accent-specific recognisers in parallel as in Figure 1(a), or when performing accent reclassification as described in Section II, different approaches can be followed to acquire the required accent-specific acoustic models. In this paper we consider two alternatives. The same approaches have been previously applied to modelling of AE and EE [8] and to multilingual acoustic modelling [9] in tied-state systems, while similar approaches were adopted in [10] and [11] for tied-mixture topologies. The two approaches are distinguished by different methods of decision-tree state clustering:

1) *Accent-specific acoustic modelling:*
Separate accent-specific acoustic models are obtained by not allowing any sharing of data between accents. Separate decision-trees are grown for each accent and the clustering process employs only questions relating to phonetic context.
2) *Multi-accent acoustic modelling:*
A single set of decision-trees is grown for all accents. In this case the decision-tree questions take into account not only the phonetic context, but also the accent of the basephone. Tying across accents can thus occur when triphone states are similar, while separate modelling of the same triphone state from different accents can be performed when there are differences.

In addition, we also considered *accent-independent acoustic modelling* in which a single accent-independent model set is obtained by pooling accent-specific data across accents for phones with the same IPA classification. A single set of decision-trees is constructed for all accents and the clustering process employs only questions relating to phonetic context. Such pooled models are often employed in multi-accent ASR (e.g. [12], [13]) and therefore represent an important baseline.

C. System configuration, evaluation and objectives

We performed word recognition experiments for the AE+EE and BE+EE accent pairs. Using the accent-specific and multi-accent acoustic modelling approaches described in Section IV-B, we trained reclassified models using the approach described in Section II and employed these models in parallel during recognition. As a baseline we used the same two acoustic modelling approaches to train models on the unmodified training data. Systems employing these models were used to perform both parallel recognition, in which accented speech is presented to both recognisers, as well as oracle recognition, in which each test utterance is presented only to the correct accent-specific recogniser. These two configurations are illustrated in Figure 1 for the AE+EE pair. As a further benchmark we developed accent-independent recognition systems for each of the two accent pairs. The parallel recognition systems perform AID implicitly and these accuracies can also be measured. These accuracies are calculated relative to the originally assigned accent labels and are therefore not relevant to the evaluation of the reclassified systems.

The chief aim of our research was to determine whether the reclassified systems could improve on parallel recognition performance compared to systems trained directly on the original databases. By performing these experiments in pairs, we are considering one scenario where accents are quite similar (AE+EE) and a second scenario where accents are relatively different (BE+EE).

V. Experimental Results

Using the combination of the AE and EE as well as the BE and EE training sets described in Section III-A, we performed speech recognition experiments using the systems described in Section IV-C. Table IV shows the average word recognition and AID accuracies measured on the evaluation sets. Oracle performance is indicated for the systems trained on the original data, but is not relevant to the reclassified systems. Because a single recogniser is used for the systems employing accent-independent models, identical results are obtained for the oracle and parallel tests. AID and accent reclassification is not possible with these fully accent-independent systems. For each configuration the development set was used to optimise the likelihood thresholds used for decision-tree clustering as well as the word insertion penalties and LM scaling factors used during recognition.

The results in Table IV indicate that, as noted in [1], the AE+EE parallel systems employing accent-specific and multi-accent acoustic models show small improvements over the corresponding oracle systems. Although the improvements are small, it is noteworthy that accent misclassifications do not lead to deteriorated system performance and instead improve overall recognition performance. In contrast we observe deteriorated performance for the BE+EE pair when using the original parallel systems compared to oracle recognition. The results also indicate that the recognition performance of the original systems employing multi-accent acoustic models is better than that achieved by the original systems employing accent-specific and accent-independent acoustic models for both accent pairs.

When comparing the performance of the original and reclassified parallel recognition systems, degradation in system

TABLE IV

Performance of AE+EE and BE+EE systems employing HMMs trained on the original databases, as well as systems employing reclassified HMMs. Word recognition accuracies (%) and AID accuracies (%) are given.

	AE+EE accent pair					BE+EE accent pair				
Model set	Original HMMs			Reclassified HMMs		Original HMMs			Reclassified HMMs	
	Oracle	Parallel	AID	Parallel	AID	Oracle	Parallel	AID	Parallel	AID
Accent-specific	84.01	84.63	80.23	84.58	78.07	76.69	76.07	93.23	75.86	93.37
Accent-independent	84.78	84.78	-	-	-	75.38	75.38	-	-	-
Multi-accent	84.78	84.88	78.22	84.61	76.99	77.35	76.75	93.16	76.60	92.40

performance is observed for both accent-specific and multi-accent acoustic modelling approaches and for both accent pairs. Except for the BE+EE accent-specific systems, the AID accuracy of the reclassified systems is lower than that of the corresponding original systems in all other cases, as one might expect. Using bootstrap confidence interval estimation [14], the statistical significance levels of the improved performance of the original parallel systems over the reclassified systems have been calculated and are shown in Table V. It is evident that the improvements are significant only at low levels varying between 56% and 80%. Nevertheless, the degradation in performance after a single iteration of reclassification is consistent across all the considered accent-pairs and acoustic modelling approaches. The AE+EE reclassified systems also show deteriorated performance in comparison to the accent-independent system, while the BE+EE reclassified systems still show superior performance.

VI. Analysis and Discussion

The comparison of oracle and parallel recognition results for the AE+EE systems trained on the originally labelled data indicates that, for some utterances, the test data is better matched to models trained on data from the other accent. However, the recognition performance of the reclassified AE+EE as well as the BE+EE systems seem to indicate that the overall mismatch between test data and models is aggravated by the reclassification process. Since the reclassification procedure is unsupervised, improvements are not guaranteed. We conclude that using the data with the originally assigned accent labels to train acoustic models is still the best strategy to follow and that no gains are achieved by using the unsupervised reclassification procedure proposed in this paper.

In order to obtain some insight into the somewhat surprising results, we have analysed utterances in the training set for which the original and the reclassified accent labels differ. We performed this analysis for the AE+EE multi-accent system and the results are presented in Table VI. The analysis indicates that the utterances for which the original accent labels have been changed are generally shorter (1.07 seconds) compared to both the overall average (2.20 seconds) as well as the average length of utterances for which accent labels were unchanged (2.28 seconds). Furthermore, the number of original AE utterances reclassified as EE utterances is approximately double the number of EE utterances reclassified as AE. In general, the proficiency of Afrikaans English speakers is high [15], which might suggest that some of the AE speakers are simply better matched to the models trained on the EE data and this might explain why AE to EE relabelling is performed more often than the opposite.

As noted in Section II, AID is performed implicitly during parallel recognition. Table VII shows an analysis of test set utterances for which the accent classification according to the original AE+EE parallel multi-accent system (84.88% accuracy, Table IV) and the corresponding reclassified system (84.61%, Table IV) differ, i.e. utterances for which the accent

TABLE V

Accuracy difference (%) and corresponding significance (sig.) levels (%) of the superior performance of the original systems over the reclassified parallel systems.

Model set	AE+EE accent pair		BE+EE accent pair	
	Difference	Sig. level	Difference	Sig. level
Accent-specific	0.05	56	0.21	70
Multi-accent	0.27	70	0.15	65

TABLE VI

Analysis of training set utterances for which the original and the reclassified accent labels differ according to the AE+EE multi-accent system. The ‘labels changed’ row is the combination of the two rows that follow.

Reclassification effect	No. of utterances	No. of tokens	Average length (s)
Labels unchanged	19 775	96 488	2.28
Labels changed:	1447	3331	1.07
AE → EE	942	2251	1.11
EE → AE	505	1080	1.00
Total/average†	21 222	99 819	2.20†

TABLE VII

Analysis of the change in recogniser selection between the original and reclassified AE+EE multi-accent systems. The ‘changed’ row is the combination of the two rows that follow.

Recogniser selection	No. of utterances	Average length (s)	Original accuracy(%)	Reclassified accuracy(%)
Unchanged	1241	2.14	85.54	85.08
Changed:	150	1.53	77.19	79.10
AE → EE	63	1.39	74.21	80.00
EE → AE	87	1.63	79.21	78.50
Overall	1391	2.08	84.88	84.61

of the accent-specific recogniser selected during recognition is different for the original and reclassified systems. Table VII indicates that, again, the utterances for which classification has changed tend to be shorter with an average length of 1.53 seconds compared to both the overall average of 2.08 seconds as well as the average length of 2.14 seconds of test utterances for which accent classification was unchanged. The breakup of system accuracy indicates that the overall drop of 0.27% absolute in accuracy is mainly due to worse performance on the utterances for which accent classification was unchanged (85.54% compared to 85.08% word recognition accuracy). Performance on the utterances for which classification has changed indicates a 1.91% improvement in performance. Rows three and four in Table VII show that this is the result of superior performance on utterances which were previously classified as AE but identified as EE by the reclassified system.

While Tables VI and VII analyse only the accent classifications and performance of the AE+EE multi-accent systems, the same analysis on the AE+EE accent-specific systems indicates similar trends. Analysis of the BE+EE systems also indicates that the training set utterances for which the manual and automatically derived accent labels differ, as well as of the test utterances for which the original and reclassified systems' accent classification are inconsistent, tend to be shorter. However, for the BE+EE case many fewer training utterances are relabelled (only approximately 450 out of the total 17657) and the number of training utterance label changes from BE to EE and vice versa are approximately equal. The original and reclassified systems are also more consistent in test utterance accent assignment and many fewer classification changes occur compared to the AE+EE case.

VII. Summary and Conclusions

In this paper we have evaluated the speech recognition performance of systems employing reclassified accent-specific recognisers in parallel for three varieties of South African English (SAE). Modelling of Afrikaans (AE), Black (BE) and White (EE) accented SAE was considered in two pairs: AE+EE and BE+EE. By classifying the accent of each utterance in the training set using first-pass acoustic models trained on the original databases and then retraining the models, reclassified acoustic models were obtained. Two acoustic modelling approaches were considered for this procedure: accent-specific acoustic modelling and multi-accent acoustic modelling. Selective cross-accent sharing of data is allowed by the latter. Systems employing reclassified models were compared with systems employing the original models and with accent-independent systems in which training data were pooled. In parallel speech recognition experiments the reclassified models showed consistently deteriorated performance compared to the original models for both accent pairs and all acoustic modelling approaches considered. Analysis indicated that the training utterances for which manual and automatically derived labels differ tend to be shorter. Fewer utterances were relabelled for the BE+EE case than for the AE+EE pair. For the latter, accent label changes from AE to EE occurred much more often than the opposite. We conclude that the proposed relabelling procedure does not lead to performance improvements and that the best strategy remains the use of the originally labelled training data.

VIII. Acknowledgements

Parts of this work were executed using the High Performance Computer (HPC) facility at Stellenbosch University. The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at are those of the authors and are not necessarily to be attributed to the NRF.

References

[1] H. Kamper and T. R. Niesler, "Multi-accent speech recognition of Afrikaans, Black and White varieties of South African English," in *Proc. Interspeech*, Florence, Italy, 2011, pp. 3189–3192.

[2] E. W. Schneider, K. Burridge, B. Kortmann, R. Mesthrie, and C. Upton, Eds., *A Handbook of Varieties of English*. Berlin, Germany: Mouton de Gruyter, 2004.

[3] J. C. Roux, P. H. Louw, and T. R. Niesler, "The African Speech Technology project: An assessment," in *Proc. LREC*, Lisbon, Portugal, 2004, pp. 93–96.

[4] A. Stolcke, "SRILM – An extensible language modeling toolkit," in *Proc. ICSLP*, Denver, CO, 2002, pp. 901–904.

[5] H. Ney, U. Essen, and R. Kneser, "On structuring probabilistic dependencies in stochastic language modelling," *Comput. Speech Lang.*, vol. 8, pp. 1–38, 1994.

[6] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book, Version 3.4*. Cambridge University Engineering Department, 2009.

[7] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proc. Workshop Human Lang. Technol.*, Plainsboro, NJ, 1994, pp. 307–312.

[8] H. Kamper, F. J. Muamba Mukanya, and T. R. Niesler, "Acoustic modelling of English-accented and Afrikaans-accented South African English," in *Proc. PRASA*, Stellenbosch, South Africa, 2010, pp. 117–122.

[9] T. R. Niesler, "Language-dependent state clustering for multilingual acoustic modelling," *Speech Commun.*, vol. 49, no. 6, pp. 453–463, 2007.

[10] T. Schultz and A. Waibel, "Language-independent and language-adaptive acoustic modeling for speech recognition," *Speech Commun.*, vol. 35, pp. 31–51, 2001.

[11] M. Caballero, A. Moreno, and A. Nogueiras, "Multidialectal Spanish acoustic modeling for speech recognition," *Speech Commun.*, vol. 51, pp. 217–229, 2009.

[12] R. Chengalvarayan, "Accent-independent universal HMM-based speech recognizer for American, Australian and British English," in *Proc. Eurospeech*, Aalborg, Denmark, 2001, pp. 2733–2736.

[13] C. Teixeira, I. Trancoso, and A. Serralheiro, "Accent identification," in *Proc. ICSLP*, Philadelphia, PA, 1996, pp. 1784–1787.

[14] M. Bisani and H. Ney, "Bootstrap estimates for confidence intervals in ASR performance evaluation," in *Proc. ICASSP*, vol. 1, Montreal, Quebec, Canada, 2004, pp. 409–412.

[15] P. F. De V. Müller, F. De Wet, C. Van Der Walt, and T. R. Niesler, "Automatically assessing the oral proficiency of proficient L2 speakers," in *Proc. SLaTE*, Warwickshire, UK, 2009, pp. 29–32.