Multi-accent speech recognition of Afrikaans, Black and White varieties of 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

10.21437/Interspeech.2011-798

Abstract

In this paper we investigate speech recognition performance of systems employing several accent-specific recognisers in parallel for the simultaneous recognition of multiple accents. We compare these systems with oracle systems, in which test utterances are presented to matching accent-specific recognisers, and with accent-independent systems, in which acoustic and language model training data are pooled. Our investigation is based on Afrikaans (AE), Black (BE) and White (EE) accents of South African English. We find that, when accent is classified on a per-utterance basis, parallel systems outperform oracle systems for the AE+EE accent pair while the opposite is observed for BE+EE. When accent identification is carried out on a per-speaker basis, oracle or better performance is obtained for both accent pairs. Furthermore, parallel systems based on multi-accent acoustic modelling, which allows selective cross-accent sharing of acoustic training data, outperform parallel systems using accent-specific acoustic models. The former also yields better performance than accent-independent recognition, which uses pooled acoustic and language models.

Index Terms: multi-accent speech recognition, parallel recognition, accent identification, South African English accents

1. Introduction

Despite steady improvement in the performance of automatic speech recognition (ASR) systems, accuracy still deteriorates strongly when confronted with accented speech. In South Africa, where official status is given to 11 different languages, non-mother-tongue speech is highly prevalent. Although English is the language of government, commerce and science, only 8.2% of the population use it as a first language [1]. English is therefore used predominantly by non-mother-tongue speakers, resulting in a large number of accents. These are in general not bound to geographic regions and hence ASR systems must be robust to multiple accents to ensure that speech-based automated services are accessible to the wider population.

For the development of any speech recognition system a large quantity of annotated speech data is required. However, speech corpora are scarce and expensive to develop, especially for under-resourced languages and accents such as South African English (SAE). It is in this light that we would like to determine the best strategy to follow when developing a system able to simultaneously recognise multiple accents of SAE given a limited corpus. We analyse the performance of multi-accent recognition systems employing multiple accent-specific recognisers in parallel. These systems are compared with oracle systems in which accented speech is presented to the matching accent-specific recogniser, and with accent-independent systems in which acoustic and language model training data are pooled across accents. For this investigation, we consider three accents of SAE: 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.

2. Accents of English in South Africa

A total of five varieties of SAE are recognised in the literature [2]. Apart from the three aforementioned accents, these include Cape Flats English (CE) and Indian South African English (IE). In the following we briefly discuss AE, BE and EE.

English was originally brought to South Africa by British occupying forces at the end of the 18th century. White South African English (EE) refers to the first language English spoken by White South Africans, chiefly of British descent. In the literature, the influence of Afrikaans on White South African English is noted as an important feature [2]. Afrikaans English (AE) refers to the accent used by second language White South African English speakers of Afrikaans descent. Afrikaans is a Germanic language with its origins in 17th century Dutch brought to South Africa by settlers from the Netherlands. Although its vocabulary still has a predominantly Dutch origin, Afrikaans has been influenced by several languages including Malay, Portuguese and the Bantu and Khoisan languages. Black South African English (BE) refers to the English variety spoken by non-mother-tongue Black South Africans. Of the South African population, 77.8% are considered Black Africans who employ one of the nine official indigenous African languages as a first language [1]. Speech recognition in this accent is therefore particularly important in the South African context.

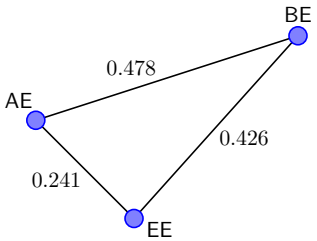

Figure 1: *Average Bhattacharyya distances between Afrikaans English (AE), Black South African English (BE) and White South African English (EE).*

To obtain some initial intuition regarding the relative similarity of the three accents, we used the Bhattacharyya distance which gives a measure of similarity between two PDFs [3]. Three-state single-mixture monophone HMMs were trained for each accent and the average of the three distances between

HMM states was calculated to obtain a measure of between-accent similarity for a particular monophone. The average distance between corresponding monophones was subsequently determined to obtain a figure indicating the similarity of two accents. For the AE, BE and EE accents of SAE, these distances are shown to scale in Figure 1. It is clear that AE+EE represents a pair of relatively similar accents while BE+EE represents a pair of relatively different accents.

3. Related research

When simultaneously recognising multiple accents, one approach is to explicitly precede accent-specific speech recognition with accent identification (AID) [4]. Two alternatives to this approach are described by Chengalvarayan [5]. The first is a system where a bank of accent-specific recognisers is run in parallel and the output with the highest associated likelihood is selected. AID is thus performed implicitly during recognition. The second alternative is to train a single model set by pooling data across accents. For recognition of American, Australian and British English, the latter showed the best performance in [5]. In [6], recognition of non-native English from six European countries was considered. Identification followed by recognition gave compareable performance to an oracle system, but both were outperformed by a pooled system.

4. Speech databases

4.1. Training and test sets

Our experiments were based on the African Speech Technology (AST) databases [7]. 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 described in Section 2. In this research we made use of only the AE, BE and EE databases. These three databases were each divided into training, development and evaluation sets. As indicated in Tables 1 and 2 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.

4.2. Language models and pronunciation dictionaries

Using the SRILM toolkit [8], separate accent-specific word backoff bigram [9] language models (LMs) were trained for each accent individually from the corresponding training set transcriptions. Additionally, AE+EE and BE+EE accent-independent bigram LMs were trained by combining the training set transcriptions of the two accents involved prior to training. This was done in order to investigate the effect of the LMs on recognition accuracy. Absolute discounting was used for the estimation of LM probabilities [10] and LM perplexities are shown in Table 3. The matched LM refers to the accent-specific LM corresponding to the accent under evaluation.

For the experiments in which accent-specific LMs were used, separate pronunciation dictionaries were obtained for each accent individually from the corresponding word and phone level training set transcriptions. For recognition experiments where the accent-independent LMs were employed, the accent-specific dictionaries were pooled. Out-of-vocabulary rates are below 7% for the BE and below 5% for the AE and EE evaluation sets.

Table 1: *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	9879	245	47 279

Table 2: *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 3: *Word bigram language model perplexities (perp.) measured on the evaluation sets.*

Accent	Bigram types	Matched LM perp.	AE+EE LM perp.	BE+EE LM perp.
AE	11 580	25.81	25.46	-
BE	9639	30.30	-	29.63
EE	10 451	28.97	27.16	26.67

5. Experimental methodology

5.1. General setup

Speech recognition systems were developed using the HTK tools [11]. 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 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 [12]. 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.

5.2. Acoustic modelling

When performing multi-accent speech recognition by running several accent-specific recognisers in parallel, different approaches can be followed to acquire the accent-specific acoustic models. In this paper we consider two alternatives. The

Table 4: *Performance of AE+EE and BE+EE oracle and parallel recognition systems when applying per-utterance AID. Word recognition accuracies (%) and per-utterance AID accuracies (%) are given for systems using accent-specific and accent-independent LMs.*

Model set	AE+EE accent pair						BE+EE accent pair					
	Accent-specific LMs			Accent-independent LM			Accent-specific LMs			Accent-independent LM		
	Oracle	Parallel	AID	Oracle	Parallel	AID	Oracle	Parallel	AID	Oracle	Parallel	AID
Accent-specific	78.70	81.38	78.72	82.62	82.97	80.16	71.64	71.51	92.26	74.84	74.18	93.30
Accent-independent	79.57	82.38	67.72	82.97	82.97	-	71.63	72.43	86.87	74.74	74.74	-
Multi-accent	79.50	81.85	78.07	83.15	83.32	78.29	72.40	72.22	92.81	75.50	74.85	93.02

same approaches have been previously applied to acoustic modelling of AE and EE [13] and to multilingual acoustic modelling [14], while similar approaches were adopted in [15] and [16]. 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. [5] and [6]) and therefore represent an important baseline.

5.3. System configuration, evaluation and objectives

By running two accent-specific recognisers in parallel and selecting the output with the highest associated likelihood, we performed speech recognition experiments for the AE+EE and BE+EE pairs. The selection of the highest scoring result can be done independently for each utterance, leading to per-utterance AID, or for each speaker, leading to per-speaker AID. The three acoustic modelling approaches described in Section 5.2 were used in combination with both the accent-specific and accent-independent LMs described in Section 4.2, resulting in six different system configurations for each of the two accent pairs. As a further benchmark, we compare the performance of these systems to those of oracle systems in which each test utterance is presented only to the correct accent-specific recogniser. In each case the oracle system used the same acoustic and language models as the system it was compared to.

The parallel recognition systems perform AID implicitly and these accuracies can also be measured. One of the chief aims of our investigation was to determine the degree of performance degradation caused by accent misclassifications when running accent-specific recognition systems in parallel. 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). This serves as motivation for considering these two accent pairs.

6. Experimental results

Several speech recognition systems were developed using the combination of the AE and EE as well as of the BE and EE training sets described in Section 4.1. 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. Table 4 shows the average word recognition and per-utterance AID accuracies measured on the evaluation sets when performing speech recognition using the systems described in Section 5.3. Because a single recogniser is used for the systems employing both accent-independent acoustic and language models, identical results are obtained for the oracle and parallel tests. AID is not performed by these fully accent-independent systems.

The results in Table 4 show consistently superior performance for the systems employing accent-independent LMs compared to those employing accent-specific LMs. This is not only the case for recognition, but also for AID. Table 3 shows that the perplexities measured on the evaluation sets are also in all cases higher for the accent-specific LMs than for the accent-independent LMs. This is attributed to the very small amount of data available for LM training (Table 1). We therefore focus on systems using accent-independent LMs in the following comparison of the oracle and parallel recognition tests, although the accent-specific LM systems show similar trends. Although even the accent-independent LMs may be considered poorly trained, they are common to all recognition systems. Hence, the accent-specific recognition systems used in the parallel recognition approaches are distinguished solely by their acoustic models.

For the AE+EE systems using accent-independent LMs, the parallel systems employing accent-specific and multi-accent acoustic models show small improvements over the corresponding oracle systems. These improvements have been calculated to be statistically significant at the 93% and 79% confidence levels for the two approaches respectively. Although the improvements are small, it is noteworthy that accent misclassifications do not lead to deteriorated system performance. Instead, the misclassifications improve overall recognition performance indicating that some test utterances are better matched to the acoustic models of the other accent. In contrast we observe deteriorated performance for the BE+EE pair when using a parallel recognition approach with the accent-independent LM. The superior performance of the BE+EE oracle systems is statistically significant at the 99% level for both the accent-specific and multi-accent acoustic modelling approaches. The results also indicate that the recognition performance of the multi-accent acoustic models is better than that achieved using accent-specific and accent-independent acoustic models for both accent pairs. The extent of these improvements depends on the accents involved and the recognition scenario (oracle or parallel), and confidence levels vary between 60% and 94%.

The performance of AE+EE systems when applying per-

Table 5: *Performance of AE+EE oracle and parallel recognition systems when applying per-speaker AID. Word recognition accuracies (%) and per-utterance AID (%) are given for systems employing the accent-independent LM.*

Model set	Oracle	Parallel	AID
Accent-specific	82.62	82.74	91.95
Multi-accent	83.15	83.19	94.75

speaker AID is shown in Table 5, where the oracle results are unchanged from Table 4. A comparison between these two tables shows that, although per-speaker AID improves identification accuracy, it leads to slightly deteriorated recognition performance. These, however, are still marginally higher than those achieved by the oracle systems. For BE+EE systems, on the other hand, per-speaker AID leads to perfect accent identification for both acoustic modelling approaches. Hence the BE+EE systems employing per-speaker AID achieve the performance of the oracle systems as indicated in Table 4, which represents an improvement over the per-utterance AID results.

7. Discussion

Our results indicate that the observed improvement or degradation in recognition performance of multiple parallel accent-specific systems relative to an oracle system depends on the similarity of the accents involved. A surprising conclusion from our experimental evaluation is that superior AID prior to accent-specific speech recognition does not necessarily lead to superior speech recognition accuracy. Furthermore, from our comparison of acoustic modelling approaches it is apparent that in both cases it is better to employ parallel speech recognisers with multi-accent acoustic models than to pool the acoustic training data of the accent pair and use the resulting accent-independent acoustic models. Finally, our experiments considering per-speaker AID indicate that, for both accent pairs, per-speaker AID yields oracle or better performance. Whether per-speaker AID can be employed will be determined by the practical speech recognition setup.

8. Summary and conclusions

We have evaluated the speech recognition performance of systems employing parallel accent-specific recognisers for three varieties of South African English (SAE). In order to determine the effect of misclassifications in the accent identification (AID) process that occurs implicitly during parallel recognition, the performance of these systems was compared with the performance of oracle systems in which test utterances are presented to matching accent-specific recognisers. The performance of parallel recognition systems was also compared with accent-independent speech recognition achieved by pooling acoustic and language model training data. Modelling of Afrikaans (AE), Black (BE) and White (EE) accented SAE was considered in two pairs: AE+EE and BE+EE. The former represents a relatively similar accent pair while the latter pair is relatively dissimilar. Speech recognition experiments demonstrated that, despite AID errors, parallel systems performing implicit per-utterance AID outperformed oracle systems for the AE+EE configuration. This was not the case for the BE+EE accent pair. However, parallel systems based on per-speaker AID showed oracle or better speech recognition performance for both accent pairs. We conclude that AID errors made during parallel recognition do not necessarily lead to deteriorated speech recognition accuracy and may in fact lead to improvements. Furthermore, we speculate that such improvements are possible for similar accents but less likely for accents that differ strongly from each other. Of the three acoustic modelling approaches considered, multi-accent modelling, which supports selective cross-accent sharing of acoustic training data, yields superior or comparable performance to the other two approaches. Finally, parallel systems employing multi-accent acoustic models outperformed systems employing accent-independent acoustic models obtained by cross-accent pooling of acoustic and language model training data. Future work includes considering recognition of all five accents of SAE.

9. Acknowledgements

This work was executed using the High Performance Computer (HPC) facility at Stellenbosch University. The work was financially supported by the National Research Foundation (NRF).

10. References

[1] Statistics South Africa, "Census 2001: Census in brief," 2003.

[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] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. San Diego, CA: Academice Press, 1990.

[4] A. Faria, "Accent classification for speech recognition," in *Proc. MLMI*, Eddinburgh, UK, 2005, pp. 285–293.

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

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

[7] 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.

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

[9] S. M. Katz, "Estimation of probabilities from sparse data for the language model component of a speech recognizer," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 35, no. 3, pp. 400–401, 1987.

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

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

[12] 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.

[13] 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.

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

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

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