Acoustic modelling of English-accented and Afrikaans-accented South African English

H. Kamper, F. J. Muamba Mukanya and T. R. Niesler
Department of Electrical and Electronic Engineering
Stellenbosch University, South Africa
kamperh@sun.ac.za, trn@sun.ac.za

***Abstract*—In this paper we investigate whether it is possible to combine speech data from two South African accents of English in order to improve speech recognition in any one accent. Our investigation is based on Afrikaans-accented English and South African English speech data. We compare three acoustic modelling approaches: separate accent-specific models, accent-independent models obtained by straightforward pooling of data across accents, and multi-accent models. For the latter approach we extend the decision-tree clustering process normally used to construct tied-state hidden Markov models by allowing accent-specific questions. We compare systems that allow such sharing between accents with those that do not. We find that accent-independent and multi-accent acoustic modelling yield similar results, both improving on accent-specific acoustic modelling.**

I. Introduction

In South Africa, English is the lingua franca as well as the language of government, commerce and science. However, the country has 11 official languages and only 8.2% of the population use English as a first language [1]. English is therefore usually used by non-mother-tongue speakers resulting in a large variety of accents. Furthermore, the use of different accents is not regionally bound as is often the case in related research. Multi-accent speech recognition is thus especially relevant in the South African context.

For the development of any speech recognition system a large quantity of annotated speech data is required. In general, the more data are available, the better the performance of the system. It is in this light that we would like to determine whether data from different South African accents of English can be combined to improve the performance of a speech recognition system in any one accent. This involves exploring phonetic similarities between accents and exploiting these to obtain more robust and effective acoustic models. In this paper we present different acoustic modelling approaches for two South African accents of English: Afrikaans-accented English and South African English.

II. Related Research

Two main approaches are encountered when considering literature dealing with multi-accent or multidialectal[1] speech recognition. Some authors consider modelling accents as pronunciation variants, which are added to the pronunciation dictionary employed by a speech recogniser [3]. Other authors focus on multi-accent acoustic modelling. These acoustic modelling approaches are often similar to techniques employed in multilingual speech recognition.

A. Multi-Accent Acoustic Modelling

One approach to multi-accent acoustic modelling is to train a single accent-independent acoustic model set by pooling accent-specific data across all accents considered. An alternative is to train separate accent-specific systems that allow no sharing between accents. These two "traditional" approaches have been considered and compared by various authors, including Van Compernolle et al. [4] for Dutch and Flemish, Beattie et al. [5] for three regional dialects of American English, Fischer et al. [6] for German and Austrian dialects and Chengalvarayan [7] who considered American, Australian and British dialects of English. From the findings of these authors it seems that in the majority of cases accent-specific modelling leads to superior speech recognition performance compared to accent-independent modelling. However, this is not always the case (e.g. [7]) and the comparative merits of the two approaches appear to depend on factors such as the abundance of training data as well as the degree of similarity between the accents involved.

In cases where accent-specific data are insufficient to train accent-specific models, adaptation techniques such as maximum likelihood linear regression (MLLR) and maximum a posteriori (MAP) adaptation can be employed. For example, MAP and MLLR have been successfully employed in the adaptation of Modern Standard Arabic acoustic models for improved recognition of Egyptian Conversational Arabic [8]. However, results obtained by Diakoloukas et al. [9] in the development of a multidialectal system for two dialects of Swedish suggest that, when larger amounts of target accent data are available, it is advantageous to simply train models on the target accented data alone.

B. Multilingual Acoustic Modelling

The question of how best to construct acoustic models for multiple accents is similar to the question of how to construct acoustic models for multiple languages. Multilingual speech recognition has received some attention over the last decade,

[1]According to [2], the term *accent* refers only to pronunciation differences, while *dialect* refers to differences in both grammar and vocabulary. *Non-native speech* refers to speech from a speaker using a language different from his or her first language. We will adhere to these definitions.

most notably by Schultz and Waibel [10]. Their research considered large vocabulary continuous speech recognition of 10 languages spoken in different countries and forming part of the GlobalPhone corpus. In addition to the two traditional approaches already mentioned (pooling and separate models), these authors evaluated acoustic models in which selective sharing between languages was allowed by means of appropriate decision-tree training of tied-mixture HMM systems. In tied-mixture systems, the HMMs share a single large set of Gaussian distributions with state-specific mixture weights. This configuration allows similar states to be clustered using entropy decrease calculated using the mixture weights as a measure of similarity. The research found that language-specific systems exhibited the best performance among the three approaches.

Multilingual acoustic modelling of four South African languages: Afrikaans, English, Xhosa and Zulu, was addressed in [11]. Similar techniques to those proposed by Schultz and Waibel were employed, but in this case applied to tied-state HMMs. In a tied-state system, each HMM state has an associated Gaussian mixture distribution and these distributions may be shared between corresponding states of different HMMs. The clustering procedure for tied-state systems will be described in Section IV-B. Modest average performance improvements were shown over language-specific and language-independent systems using multilingual HMMs.

C. Recent Research

More recently, Caballero et al. presented research which dealt with five dialects of Spanish spoken in Spain and Latin America [12]. Different approaches to multidialectal acoustic modelling were compared based on decision-tree clustering algorithms using tied-mixture systems. A dialect-independent model set (obtained by pooling) was compared to a multidialectal model set (obtained by allowing decision tree questions relating to both context and dialect). These approaches are similar to those applied in both [10] and [11]. In isolated word recognition experiments, the multidialectal model set was shown to outperform the dialect-independent model set.

III. Speech Databases

Our experiments were based on the African Speech Technology (AST) databases [13], which were also used in [11].

A. The AST Databases

The eleven AST databases were collected in five languages spoken in South Africa as well as a number of non-mother-tongue variants. The databases consists of annotated telephone speech recorded over both mobile and fixed telephone networks and contain a mix of read and spontaneous speech. The types of read utterances include isolated digits, digit strings, money amounts, dates, times, spellings and phonetically rich words and sentences. Spontaneous responses include references to gender, age, home language, place of residence and level of education. Utterances were transcribed both phonetically and orthographically.

TABLE I
Training and test sets for each accent of English

Accent	Set	Speech (min)	No. of utterances	No. of speakers	Phone tokens
English	train	356.95	9879	245	178 954
Afrikaans	train	421.14	11 344	276	199 336
English	dev	14.18	401	10	6344
Afrikaans	dev	14.36	429	12	6869
English	eval	23.96	702	18	11 304
Afrikaans	eval	24.16	689	21	10 708

Five English databases were compiled as part of the AST project: South African English from mother-tongue English speakers, as well as English from Black, Coloured, Asian and Afrikaans non-mother-tongue English speakers. In this research we made use of the South African English (EE) and Afrikaans English (AE) databases. The phonetic transcriptions of both these databases were obtained using a common IPA-based phone set consisting of 50 phones.

B. Training and Test Sets

Each database was divided into a training (train), development (dev) and evaluation (eval) set, as indicated in Table I. The EE and AE training sets contain 5.95 and 7.02 hours of speech audio data respectively. The evaluation set contains approximately 24 minutes of speech from 20 speakers in each accent. There is no speaker-overlap between the evaluation and training sets.

The development set consists of approximately 14 minutes of speech from 10 speakers in each accent. This data was used only for the optimisation of the recognition parameters before final evaluation on the evaluation set. There is no speaker-overlap between the development set and either the training or evaluation sets. For the development and evaluation sets the ratio of male to female speakers are approximately equal and all sets contain utterances from both land-line and mobile phones.

IV. General Experimental Methodology

Speech recognition systems were developed using the HTK tools [14] following three different acoustic modelling approaches that will be described in Section V. An overview of the common setup of these systems are given in the following.

A. General Setup

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 set from each accent was used to obtain three-state left-to-right single-mixture monophone HMMs with diagonal-covariance using embedded Baum-Welch re-estimation. These monophone models were then cloned and re-estimated to obtain initial accent-specific cross-word triphone models which were subsequently clustered using decision-tree state clustering [15]. 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.

The distinction between the different acoustic modelling approaches considered is based solely on different methods of decision-tree clustering. Since decision-tree state clustering is central to the research presented here, it is summarised below.

B. Decision-Tree State Clustering

The clustering process is normally initiated by pooling the data of corresponding states from all context-dependent phones with the same base phone in a single cluster. This is done for all context-dependent phones observed in the training set. A set of linguistically-motivated questions is then used to split these initial clusters. Such questions may, for example, ask whether the left context of a particular context-dependent phone is a vowel or whether the right context is a silence. Each potential question results in a split which yields an increase in likelihood of the training set and for each cluster the optimal question is determined. Based on this splitting criteria, clusters are subdivided repeatedly until either the increase in likelihood or the number of frames associated with a resulting cluster falls below a certain threshold (the minimum cluster occupancy).

The result is a phonetic binary decision-tree where the leaf nodes indicate clusters of context-dependent phones for which data should be pooled. The advantage of this approach is that each state of a context-dependent phone not seen in the training set can be associated with a cluster using the decision-trees. This allows the synthesis of models for unseen context-dependent phones.

C. Language Models

Comparison of recognition performance was based on phone recognition experiments. Since the presented work considers only the effect of the acoustic models, recognition of a specific test set was performed using a language model trained on the training set of the same accent. Using the SRILM toolkit [16], backoff bigram language models were trained for each accent individually from the corresponding training set phone transcriptions [17]. Absolute discounting was used for the estimation of language model probabilities [18]. Language model perplexities are shown in Table II for the two English accents. The development set was used to optimise the word insertion penalties and language model scaling factors used during recognition.

V. Acoustic Modelling Approaches

We considered three acoustic modelling approaches. Similar approaches were followed in [10] and [11] for multilingual acoustic modelling, and in [12] for multi-dialectal acoustic modelling. The fundamental aim of our research was to determine which acoustic modelling approach takes best advantage of the data available to us (Section III-B).

TABLE II
Bigram language model perplexities measured on the evaluation test-sets

Accent	Bigram types	Perplexity
English	1542	12.63
Afrikaans	1894	14.37

A. Accent-Specific Acoustic Models

As a first approach, a baseline system was developed by constructing accent-specific model sets where no sharing is allowed between accents. Corresponding states from all triphones with the same basephone are clustered separately for each accent, resulting in separate decision-trees for the two accents. The decision-tree clustering process employs only questions relating to phonetic context. The structure of the resulting acoustic models is illustrated in Figure 1 for both an Afrikaans-accented and a South African English triphone of basephone [i] in the left context of [j] and the right context of [k]. This approach results in a completely separate set of acoustic models for each accent since no data sharing is allowed between triphones from different accents. Information regarding accent is thus considered more important than information regarding phonetic context.

B. Accent-Independent Acoustic Models

For the second approach, a single accent-independent model set was obtained by pooling accent-specific data across the two accents for phones with the same IPA classification. A single set of decision-trees is constructed for both accents and employs only questions relating to phonetic context. Information regarding phonetic context is thus regarded as more important than information regarding accent. Figure 2 illustrates the acoustic models, again for both an Afrikaans-accented and a

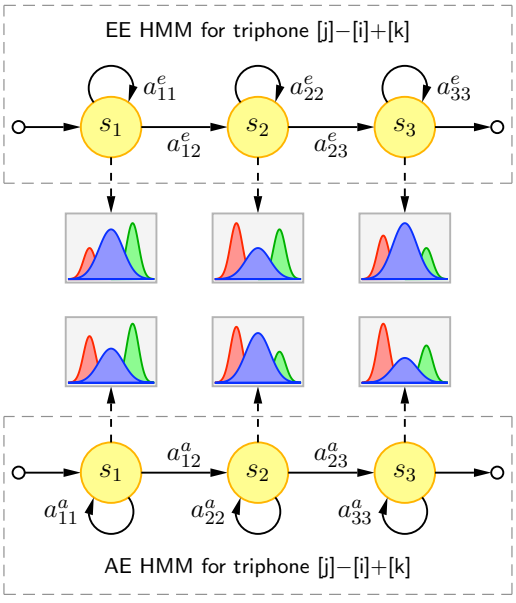

Fig. 1. Accent-specific acoustic models.

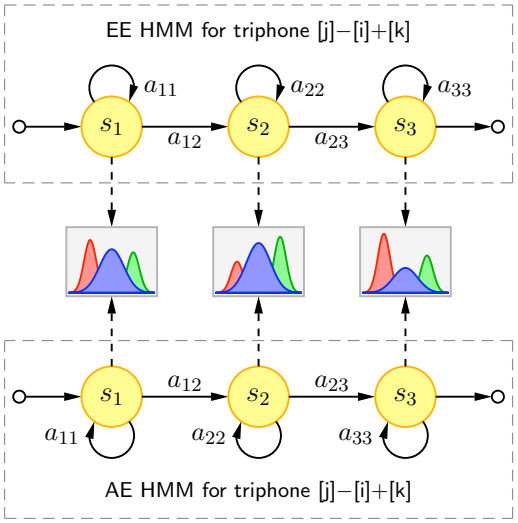

Fig. 2. Accent-independent acoustic models.

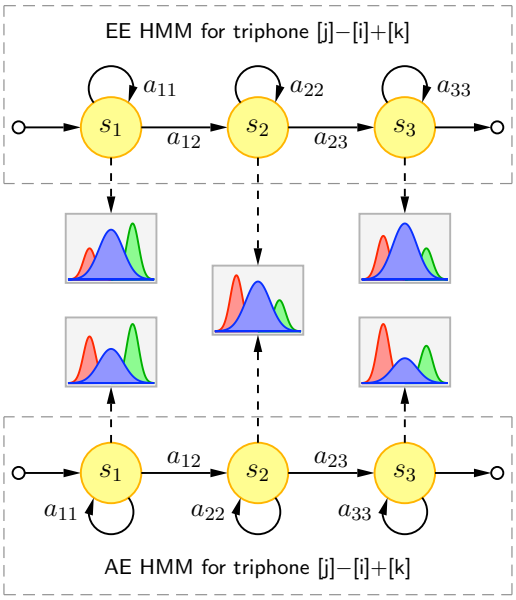

Fig. 3. Multi-accent acoustic models.

South African English triphone. Both triphone HMMs share the same Gaussian mixture probability distributions as well as transition probabilities.

C. Multi-Accent Acoustic Models

The third and final approach involved obtaining multi-accent acoustic models. This approach is similar to that followed for accent-independent acoustic modelling. Again, the state clustering process begins by pooling corresponding states from all triphones with the same basephone. However, in this case the set of decision-tree questions take into account not only the phonetic character of the left and right context, but also the accent of the basephone. The HMM states of two triphones with the same IPA symbols but from different accents can therefore be kept separate if there is a significant acoustic difference, or can be merged if there is not. Tying across accents is thus performed when triphone states are similar, and separate modelling of the same triphone state from different accents is performed when there are differences. A data-driven decision is made regarding whether accent information is more or less important than information relating to phonetic context.

The structure of such multi-accent acoustic models is illustrated in Figure 3. Here the centre state of the triphone [j]-[i]+[k] is tied across accents while the first and last states are modelled separately. As for the the accent-independent acoustic models, the transition probabilities of all triphones with the same basephone are tied across both accents.

VI. Experimental Results

The acoustic modelling approaches described in Section V were applied to the combination of the Afrikaans-accented and South African English training sets described in Section III. Since the optimal size of an acoustic model set is not known beforehand, several sets of HMMs were produced by varying the likelihood improvement threshold during the decision-tree clustering process (described in Section IV-B). The minimum cluster occupancy was set to 100 frames for all experiments.

A. Analysis of Recognition Performance

Figure 4 shows the average phone recognition accuracy measured on the evaluation set using the final eight-mixture triphone models. For each approach a single curve indicating the average accuracy between the accents is shown. The number of states for the accent-specific systems is taken to be the sum of the number of states in each component accent-specific HMM set. The number of states for the multi-accent systems is taken to be the total number of unique states remaining after decision-tree clustering and hence takes cross-accent sharing into account.

The results presented in Figure 4 indicate that, over the range of models considered, accent-specific modelling performs worst while accent-independent and multi-accent modelling yield similar performance improvements. The best

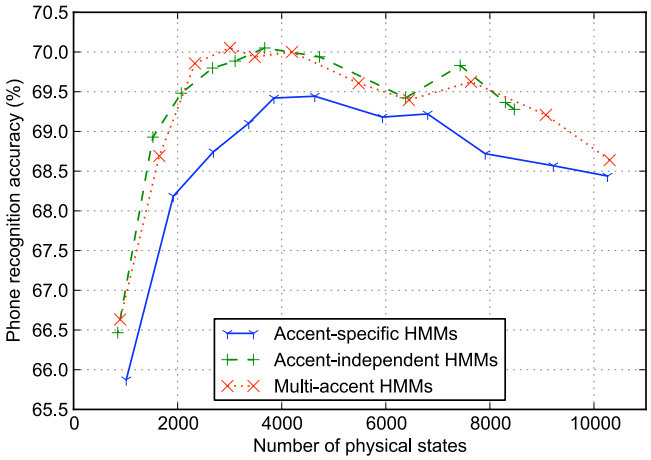

Fig. 4. Average evaluation test-set phone accuracies of accent-specific, accent-independent and multi-accent systems as a function of total number of distinct HMM states.

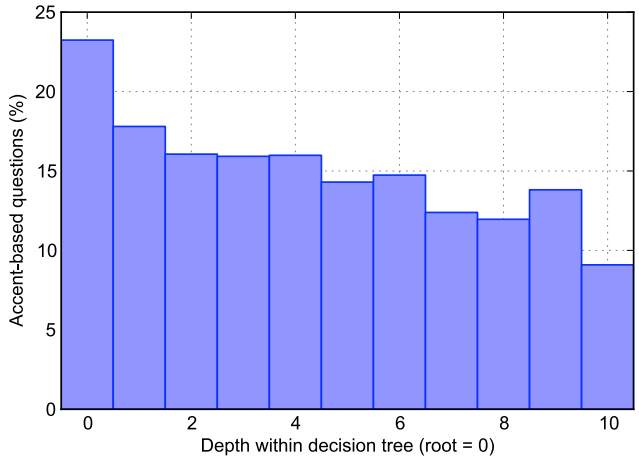

Fig. 5. Analysis showing the percentage of questions that are accent-based at various depths within the multi-accent decision-trees for the largest multi-accent system.

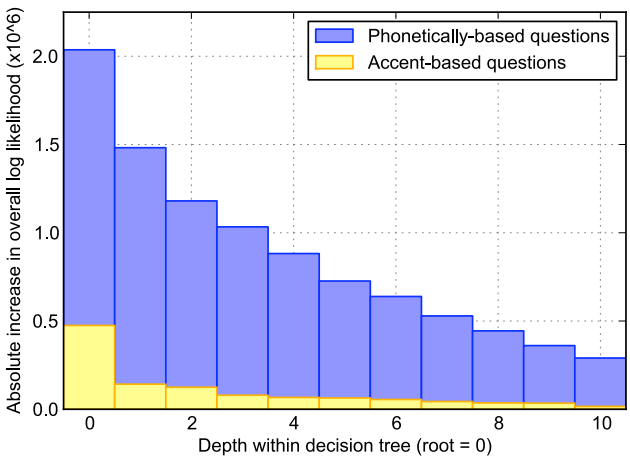

Fig. 6. Analysis showing the contribution made to the increase in overall log likelihood by the accent-based questions and phonetically-based questions respectively for the largest multi-accent system.

accent-specific system yields an average phone recognition accuracy of 69.44% (4635 states) while the best accent-independent system (3673 states) and the best multi-accent system (3006 states) both yield an average accuracy of 70.05%. The improvements of the best accent-independent and the multi-accent systems compared to the best accent-specific system were found to be statistically significant at the 95% level using bootstrap confidence interval estimation [19]. Similar trends were observed in the phone recognition accuracy measured separately on the evaluation set of each accent.

The results clearly indicate that there is little to no advantage in multi-accent acoustic modelling relative to accent-independent modelling for the two accents considered. When comparing the two approaches where the difference in performance is relatively high and the number of physical states is approximately equal (3006 states for the multi-accent system and 3104 states for the accent-independent system) the absolute improvement of 0.17% is found to be statistically significant only at the 70% level. The current practice of simply pooling data across accents when considering acoustic modelling of English is thus supported by our findings.

Our results are however in contrast to the findings of many authors where accent-specific modelling seemed to improve recognition performance [4]–[6], although they do agree with the findings of some studies [7]. In general, the proficiency of Afrikaans English speakers is high, which might suggest that the two accents are quite similar and thus explain why accent-independent modelling is advantageous [20]. The results are also in contrast to those presented in [11] where multilingual acoustic modelling of four South African languages was considered, and which were also based on the AST databases. In that research, modest improvements were seen using multilingual HMMs relative to language-specific and language-independent systems, while the language-independent models performed worst. While there is a strong difference between the multilingual and multi-accent cases, similar databases were used and hence the results are comparable to some degree.

B. Analysis of the Decision-Trees

Figure 5 analyses the decision-trees of the largest multi-accent system (10 302 states). The figure shows that, although accent-based questions are most common at the root node of the decision-trees and become increasingly less frequent towards the leaves, at most depths between approximately 12% and 16% of questions are accent-based. This suggests that accent-based questions are more or less evenly distributed through the different depths of the decision-trees and that early partitioning of models into accent-based groups is not necessarily performed or advantageous. This is in contrast to the multilingual case where the percentage of language-based questions drops from more than 45% at the root node to less than 5% at the 10^{th} level of depth [11].

The minimal influence of accent is emphasised further when considering the contribution to the log likelihood improvement made by the accent-based and phonetically-based questions respectively during the decision-tree growing process. Figure 6 illustrates this improvement as a function of depth within the decision-tree and clearly shows that phonetically-based questions make a much larger contribution to the log likelihood improvement than the accent-based questions. It is evident that, at the root node, the greatest log likelihood improvement is afforded by the phonetically-based questions (approximately 77% of the total improvement). At no depth do the accent-based questions yield log likelihood improvements comparable to those of the phonetically-based questions. This is again in contrast to the multilingual case, where approximately 74% of the total log likelihood improvement is due to language-based questions at the root node and the decision-trees tend to quickly partion models into language-based groups [11].

C. Analysis of Cross-Accent Data Sharing

In order to determine to what extent data sharing takes place for the various multi-accent systems, we considered the proportion of decision-tree leaf nodes (which correspond to the state clusters) that are populated by states from both accents. A

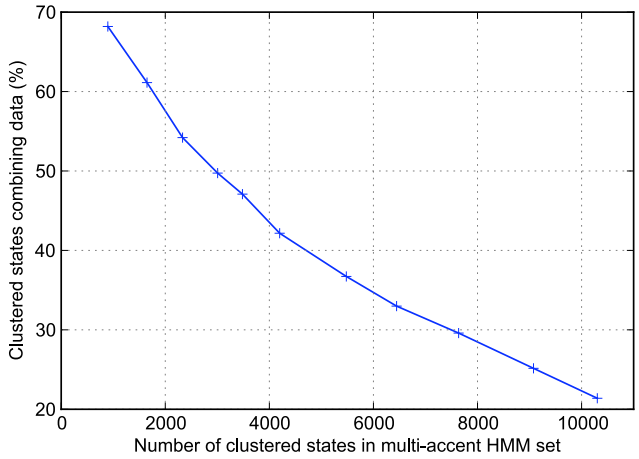

Fig. 7. Proportion of state clusters combining data from both accents.

cluster populated by states from a single accent indicates that no sharing is taking place, while a cluster populated by states from both accents indicates that sharing is taking place across accents. Figure 7 illustrates how these proportions change as a function of total number of clustered states in a system.

From Figure 7 it is apparent that as the number of clustered states is increased, the proportion of clusters consisting of both accents decreases. This indicates that the multi-accent decision-trees tend towards separate clusters for each accent as the likelihood improvement threshold is lowered, as we might expect. It is interesting to note that, although our findings suggest that multi-accent and accent-independent systems give similar performance, the optimal multi-accent system (3006 states) models approximately 50% of state clusters separately for each accent. Thus, although accent-independent modelling is advantageous when compared to accent-specific modelling, multi-accent modelling does not impair recognition performance even though a large degree of separation takes place. For the optimal multilingual system in [11], only 20% of state clusters contained more than one language, emphasising that the multi-accent case is much more prone to sharing.

VII. Conclusions and Future Work

The evaluation of three approaches to multi-accent acoustic modelling of Afrikaans-accented English and South African English has been presented. The aim was to find the best acoustic modelling approach given the available accented AST data. Tied-state multi-accent models, obtained by introducing accent-based questions into the decision-tree clustering process and thus allowing for selective sharing between accents, were found to yield similar results to accent-independent models, obtained by simply pooling data across accents. Both these approaches were found to be superior to accent-specific modelling. Further analysis of the decision-trees constructed during the multi-accent modelling process indicated that questions relating to phonetic context resulted in a much larger contribution to the likelihood increase than the accent-based questions, although a significant proporation of state clusters did contain only one accent. We conclude that, for the two accented speech databases considered, the inclusion of accent-based questions does not impair recognition performance, but also does not yield any significant gain. Future work includes considering less-similar English accents (e.g. Black English and South African English) and multi-accent acoustic modelling of all five English accents found in the AST databases.

Acknowledgements

Parts of this work were executed using the High Performance Computer (HPC) facility at Stellenbosch University.

References

[1] Statistics South Africa, "Census 2001: Census in brief," 2003.
[2] D. Crystal, *A Dictionary of Linguistics and Phonetics*, 3rd ed. Oxford, UK: Blackwell Publishers, 1991.
[3] J. J. Humphries and P. C. Woodland, "Using accent-specific pronunciation modelling for improved large vocabulary continuous speech recognition," in *Proc. Eurospeech*, vol. 5, Rhodes, Greece, 1997, pp. 2367–2370.
[4] D. Van Compernolle, J. Smolders, P. Jaspers, and T. Hellemans, "Speaker clustering for dialectic robustness in speaker independent recognition," in *Proc. Eurospeech*, Genove, Italy, 1991, pp. 723–726.
[5] V. Beattie, S. Edmondson, D. Miller, Y. Patel, and G. Talvola, "An integrated multi-dialect speech recognition system with optional speaker adaptation," in *Proc. Eurospeech*, Madrid, Spain, 1995, pp. 1123–1126.
[6] V. Fischer, Y. Gao, and E. Janke, "Speaker-independent upfront dialect adaptation in a large vocabulary continuous speech recognizer," in *Proc. ICSLP*, Sydney, Australia, 1998, pp. 787–790.
[7] R. Chengalvarayan, "Accent-independent universal HMM-based speech recognizer for American, Australian and British English," in *Proc. Eurospeech*, Aalborg, Denmark, 2001, pp. 2733–2736.
[8] K. Kirchhoff and D. Vergyri, "Cross-dialectal data sharing for acoustic modeling in Arabic speech recognition," *Speech Commun.*, vol. 46, no. 1, pp. 37–51, 2005.
[9] V. Diakoloukas, V. Digalakis, L. Neumeyer, and J. Kaja, "Development of dialect-specific speech recognizers using adaptation methods," in *Proc. ICASSP*, Munich, Germany, 1997, pp. 1455–1458.
[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] T. R. Niesler, "Language-dependent state clustering for multilingual acoustic modelling," *Speech Commun.*, vol. 49, no. 6, pp. 453–463, 2007.
[12] M. Caballero, A. Moreno, and A. Nogueiras, "Multidialectal Spanish acoustic modeling for speech recognition," *Speech Commun.*, vol. 51, pp. 217–229, 2009.
[13] 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.
[14] 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.
[15] 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.
[16] A. Stolcke, "SRILM – An extensible language modeling toolkit," in *Proc. ICSLP*, vol. 2, Denver, Co, 2002, pp. 901–904.
[17] 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.
[18] H. Ney, U. Essen, and R. Kneser, "On structuring probabilistic dependencies in stochastic language modelling," *Comput. Speech Lang.*, vol. 8, pp. 1–38, 1994.
[19] 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.
[20] 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.