Multi-accent acoustic modelling of South African English

Herman Kamper, Félicien Jeje Muamba Mukanya, Thomas Niesler*

Department of Electrical and Electronic Engineering, Stellenbosch University, South Africa

Abstract

Although English is spoken throughout South Africa, it is most often used as a second or third language, resulting in several prevalent accents within the same population. When dealing with multiple accents in this under-resourced environment, automatic speech recognition (ASR) is complicated by the need to compile multiple, accent-specific speech corpora. We investigate how best to combine speech data from five South African accents of English in order to improve overall speech recognition performance. Three acoustic modelling approaches are considered: separate accent-specific models, accent-independent models obtained by pooling training data across accents, and multi-accent models. The latter approach extends the decision-tree clustering process normally used to construct tied-state hidden Markov models (HMMs) by allowing questions relating to accent. We find that multi-accent modelling outperforms accent-specific and accent-independent modelling in both phone and word recognition experiments, and that these improvements are statistically significant. Furthermore we find that the relative merits of the accent-independent and accent-specific approaches depend on the particular accents involved. Multi-accent modelling therefore offers a mechanism by which speech recognition performance can be optimised automatically, and for hard decisions regarding which data to pool and which to separate to be avoided.

Keywords: Multi-accent acoustic modelling, Multi-accent speech

*Corresponding author. Tel.: +27 21 808 4118.

Email addresses: kamperh@sun.ac.za (Herman Kamper), trn@sun.ac.za (Thomas Niesler)

1. Introduction

Despite steady improvement in the performance of automatic speech recognition (ASR) systems in controlled environments, the accuracy of these systems still deteriorates strongly when confronted with highly accented speech. In countries with non-homogeneous populations, non-mother-tongue speech is highly prevalent. When the language in question is also under-resourced, it is important to know how best to make use of the limited speech resources to provide the best possible recognition performance in the prevalent accents.

The South African constitution gives official status to eleven different languages, as summarised in Figure 1. Although English is the lingua franca, as well as the language of government, commerce and science, only 8.2% of the population use it as a first language. Hence, English is used predominantly by non-mother-tongue speakers, and this results in a large number of accents. These accents are in general not bound to geographic regions, as is often the case for other world accents. South African English (SAE) therefore provides a challenging and relevant scenario for the modelling of accents in ASR. It also can be classified as an under-resourced variety of English, since the annotated speech available for the development of ASR systems is exceedingly limited.

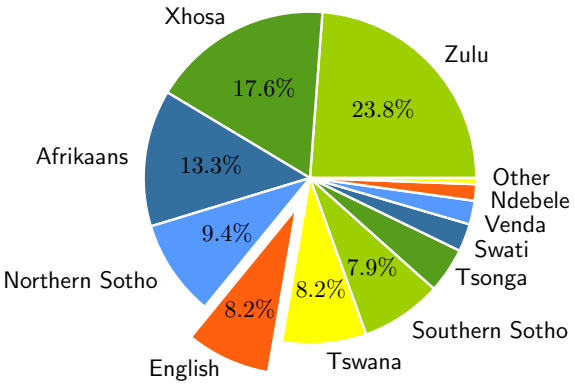

Figure 1: Mother-tongue speakers of the eleven official languages in South Africa, as a percentage of the population (Statistics South Africa, 2004).

The research presented in this paper considers the question of how best to optimise HMM-based acoustic models, when presented with a very limited

corpus of different accents[1] of SAE. Although the speech databases used in this research are small compared to those used in state-of-the-art systems, the scenario considered here is representative of an under-resourced environment in which the presence of multiple accents further aggravates the development of ASR technology.

2. Related Research

Two main approaches are encountered when considering the literature dealing with multi-accent or multidialectal speech recognition. Some authors consider modelling accents as pronunciation variants which are added to the pronunciation dictionary employed by a speech recogniser (Humphries and Woodland, 1997). Other authors focus on multi-accent acoustic modelling. We will take the latter approach and begin by presenting a brief review.

2.1. Multi-Accent Acoustic Modelling

A popular approach to multi-accent acoustic modelling is to pool data from all accents considered, resulting in a single accent-independent acoustic model set. An alternative is to train separate accent-specific systems that allow no sharing between accents. These two contrasting approaches have been considered and compared by many authors, including those summarised in Table 1. In most cases, accent-specific models lead to superior speech recognition performance when compared with accent-independent models. However, this is not always the case, as demonstrated by Chengalvarayan (2001), and the comparative merits of the two approaches appear to depend on factors such as the abundance of training data, the type of task and the degree of similarity between the accents involved.

In cases where the quantity of data is insufficient for the training of 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 (Kirchhoff and Vergyri, 2005).

[1] According to Crystal (1991), the term 'accent' refers only to pronunciation differences, while 'dialect' refers also to differences in grammar and vocabulary. It is not always obvious whether we are dealing with accents or dialects when considering varieties of SAE, and we shall therefore use the term 'accent' exclusively to avoid confusion.

Table 1: Literature comparing accent-specific and accent-independent modelling approaches, as well as various forms of adaptation.

Authors	Accents	Task	Training corpus	Best approach
Van Compernolle et al. (1991)	Dutch and Flemish	Isolated digit recognition	3993 Dutch and 4804 Flemish utterances	Accent-specific modelling
Beattie et al. (1995)	Three dialects of American English	Command and control (200 words)	Not indicated	Gender- and dialect-specific modelling
Fischer et al. (1998)	German and Austrian dialects	Large vocabulary continuous speech recognition	90h German; 15h Austrian speech	Accent-specific modelling
Chengalvarayan (2001)	American, Australian and British dialects of English	Connected digit recognition	7461 American, 5298 Australian and 2561 British digit strings	Accent-independent modelling
Caballero et al. (2009)	Five Spanish dialects (Spain, Argentina, Venezuela, Columbia, Mexico)	Isolated word recognition	50 000 Spanish utterances and 10 000 from each remaining dialect	Multi-dialect, followed by accent-independent modelling
Diakoloukas et al. (1997)	Stockholm and Scanian dialects of Swedish	Travel information task	21 000 Stockholm sentences; different amounts of Scanian adaptation data	Less data: adaptation; more data: accent-specific modelling
Wang et al. (2003)	Non-native English from German speakers	Spontaneous face-to-face dialogues	34h native English; 52min non-native adaptation data	Decision-tree-based adaptation, followed by MAP
Kirchhoff and Vergyri (2005)	Modern Standard Arabic and Egyptian Conversational Arabic	Large vocabulary continuous speech recognition	40h Modern Standard Arabic; 20h Egyptian Conversational Arabic	An approach employing both MAP and MLLR
Despres et al. (2009)	Northern and Southern dialects of Dutch	Broadcast news	100h Northern Dutch; 50h Southern Dutch	MAP

Although the results obtained by Diakoloukas et al. (1997) suggest that training acoustic models on target accented data alone is superior to adaptation when larger amounts of accented data are available, Despres et al. (2009) found that accent-independent models which have been adapted with accented data outperformed both accent-specific and accent-independent models for two varieties of Dutch. Similarly, Wang et al. (2003) showed that MAP adapted models outperformed pooled models when considering recognition of non-native English by German speakers. In that study, which considered several pooling and adaptation strategies, models obtained using decision-tree-based adaptation outperformed pooled, MAP-adapted and interpolated models.

2.2. Multilingual Acoustic Modelling

The question of how best to construct acoustic models for multiple accents is similar in some respects to the question of how to construct acoustic models for multiple languages. Multilingual speech recognition has received some attention over the last decade, most notably by Schultz and Waibel (2001). 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 described (pooling and separate models), these authors evaluated acoustic models in which selective sharing of data 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 by maximising an entropy calculated using the mixture weight vectors. 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, Zulu and Xhosa) was addressed in (Niesler, 2007). 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. Multilingual HMMs showed modest average performance improvements over language-specific and language-independent systems for the languages considered.

2.3. Recent Research

More recently, Caballero et al. (2009) considered five dialects of Spanish spoken in Spain and Latin America. Experiments were based on databases recorded in Spain, Argentina, Venezuela, Colombia and Mexico. Different approaches to multidialectal acoustic modelling were compared based on decision-tree clustering algorithms using tied-mixture systems, as also employed by Schultz and Waibel (2001). Dialect-independent modelling (pooling across dialects), dialect-specific modelling (separate modelling) and multidialectal modelling (obtained by allowing decision-tree questions relating to both phonetic context and dialect) were compared. The training material consisted of approximately 50 000 utterances by 3500 speakers from Spain, and approximately 10 000 utterances by 800 speakers from each of the Latin American countries. In isolated word recognition experiments, the multidialectal models, achieving a word error rate (WER) of 6.63%, were shown to outperform the dialect-independent model set (7.02% WER), which in turn outperformed the dialect-specific model set (7.40% WER).

3. Speech Databases

3.1. The AST Databases

Our experiments were based on the African Speech Technology (AST) databases (Roux et al., 2004). The databases consist 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

Table 2: Percentage of the South African population falling into specific speaker groups, loosely indicating the proportion of speakers of a corresponding SAE accent (Statistics South Africa, 2004). 'Other' refers to speakers not falling into one of the relevant groups, for example a White speaker using Xhosa as a first language.

Speaker group (ethnic group and first language)	Speakers (%)
White Afrikaans speakers (AE)	5.66
Black speakers of an official Black language (BE)	77.78
Coloured Afrikaans or English speakers (CE)	8.77
White English speakers (EE)	3.77
Indian or Asian English speakers (IE)	2.33
Other	1.70

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 orthographically as well as phonetically.

As part of the AST Project, five English speech databases were compiled, corresponding to the following accents of English described by Schneider et al. (2004): Afrikaans English (AE), Black South African English (BE), Cape Flats English (CE), White South African English (EE), and Indian South African English (IE). It is important to note that although the labels used to differentiate between these accents are not intended to reflect Apartheid classifications, there exists an undeniable correlation between the different accents of English used in South Africa and the different ethnic groups. In Table 2 an indication is given of the proportion of the South African population using each of these accents. It is evident from the table that non-mother-tongue variants of English (spoken by AE, BE and some CE speakers) are used by the overwhelming majority of the population. Notwithstanding the uneven distribution of speakers shown in Table 2, the five AST databases have approximately equal size. Approximately 7 hours of annotated speech data have been collected per accent, and this currently represents the largest such resource of SAE available for research. A brief description of each accent is presented in the following.

3.2. Varieties of South African English

English was originally brought to South Africa by British occupying forces at the end of the 18^{th} century. Today approximately 8.2% of the South African population use English as a first language (Statistics South Africa, 2004). White South African English refers to the first language English spoken by White South Africans, chiefly of British descent. When considering the phonology, morphology and syntax of White South African English as described by Bowerman (2004a,b), the influence of Afrikaans on White South African English is noted as an important feature.

Afrikaans English refers to the accent used by White South African second language English speakers of Afrikaans descent. Afrikaans is a Germanic language with its origins in 17^{th} century Dutch brought to South Africa by settlers from the Netherlands. It was influenced by various other languages including Malay, Portuguese and the Bantu and Khoisan languages, although the Afrikaans vocabulary still has a predominantly Dutch origin. As indicated in Table 2, White Afrikaans speakers comprise approximately 5.7% of

the South African population.

Black South African English refers to the English spoken by non-mother-tongue Black South Africans. Since 77.8% of the South African population are considered Black Africans who employ one of the 9 official indigenous African languages as a first language (Table 2), it is not surprising that Black South African English has become prominent in government, commerce and the media since 1994 (Van Rooy, 2004). Speech recognition of this accent is therefore particularly important in the South African context. The AST BE database contains English speech gathered from mother-tongue speakers of the Nguni languages (Zulu, Xhosa, Swati, Ndebele) as well as speakers of the Sotho languages (Northern Sotho, Southern Sotho, Tswana).

Indian languages were brought to South Africa by labourers who were recruited from India after the abolition of slavery in European colonies in the 19th century. These Indian languages have existed in South Africa since 1860, mainly in Natal (KwaZulu-Natal today). Indian South African English presents an interesting sociolinguistic case: the dialect shifted from being associated with second language speakers (originally as a lingua franca) to a first language, despite the Apartheid policy (1948-1991) preventing contact between Indian children and first language English speakers (Mesthrie, 2004b). Today, the majority of South African Indians use English as a first language. According to Statistics South Africa (2004), approximately 2.5% of the South African population are considered Indian or Asian and 94% speak English as a first language. The influence of not only English, but also Zulu and (to a lesser extent) Afrikaans on the development of Indian South African English is noted by Mesthrie (2004a,b).

Cape Flats English has its roots in 19th century working class residential areas in inner-city Cape Town, where residents from many different ethnic affiliations, religions and languages came into regular contact with one another. The accent spread as residents from these mixed neighbourhoods moved or were forced to move to the Cape Flats (a low-lying, flat expanse bordered by mountain ranges and the sea) in the 1960s and 1970s (Finn, 2004). The term 'Coloured' applies to the mixed-race ethnic group most closely associated with the Cape Flats English accent today. The diverse ancestry of these speakers includes Europe, Indonesia, Madagascar, Malaysia, Mozambique, Mauritius, Saint Helena and Southern Africa. While many Coloured speakers use a dialect of Afrikaans as a home language, English is also often considered a first language (McCormick, 2004). These people comprise approximately 8.8% of the South African population (Table 2). The connec-

tion between Cape Flats English and Afrikaans English, which are both also closely associated with White South African English, is emphasised by Finn (2004).

3.3. Training and Test Sets

The five English AST databases were each divided into training, development and evaluation sets, as indicated in Tables 3, 4 and 5 respectively. The training sets each contain between 6 and 7 hours of speech from approximately 250 speakers, while the development and evaluation sets contain approximately 14 minutes from 10 speakers and 25 minutes from 20 speakers respectively. The development set was used only for the optimisation of the recognition parameters before final testing on the evaluation set. 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. All data in the five accented English AST databases is used in our experiments, for training, for development, or for testing.

Table 3: Training sets for each database.

Database	Speech (h)	No. of utterances	No. of speakers	Phone tokens	Word tokens
AE	7.02	11 344	276	199 336	52 540
BE	5.45	7779	193	140 331	37 807
CE	6.15	10 004	231	174 068	46 185
EE	5.95	9879	245	178 954	47 279
IE	7.21	15 073	295	218 372	57 253
Total	31.78	54 079	1240	911 061	241 064

3.4. Phone Set

The AST project included the phonetic transcription of the five English-accented databases by linguistic experts using a large IPA-based phone set, similar to that described in (Niesler, 2007). Since certain phones occurred only in some of the databases and with very low frequency, these were mapped to a smaller set of 50 phones common to all five accents. Fewer than 1.2% of all the phone tokens were affected by this process.

Table 4: Development sets for each database.

Database	Speech (min)	No. of utterances	No. of speakers	Phone tokens	Word tokens
AE	14.36	429	12	6869	1855
BE	10.31	303	8	4658	1279
CE	13.49	377	10	6217	1700
EE	14.18	401	10	6344	1728
IE	14.53	620	13	7508	2044
Total	66.87	2130	53	31 596	8606

Table 5: Evaluation sets for each database.

Database	Speech (min)	No. of utterances	No. of speakers	Phone tokens	Word tokens
AE	24.16	689	21	10 708	2913
BE	25.77	745	20	11 219	3100
CE	23.83	709	20	11 180	3073
EE	23.96	702	18	11 304	3059
IE	25.41	865	20	12 684	3362
Total	123.13	3710	99	57 095	15 507

Table 6: Accent-specific phone bigram language model perplexities measured on the evaluation sets.

Database	Bigram types	Perplexity
AE	1891	14.40
BE	1761	15.44
CE	1834	14.12
EE	1542	12.64
IE	1760	14.24

Table 7: Accent-independent word bigram language model perplexities and OOV rates measured on the evaluation sets.

Database	Bigram types	Perplexity	OOV rate
AE	11 580	24.07	1.82%
BE	9639	27.87	2.84%
CE	10 641	27.45	1.40%
EE	10 451	24.90	1.08%
IE	11 677	25.55	1.73%

3.5. Language Models

Speech recognition performance was assessed in terms of both phone and word error rates. For the phone recognition experiments, separate accent-specific phone backoff bigram language models (Katz, 1987) were trained for each accent using the corresponding training set transcriptions and the SRILM toolkit (Stolcke, 2002). For the word recognition experiments, the same tools were used to train accent-independent bigram language models on the combined set of training transcriptions of all five accents in the AST databases (approximately 240k words). Initial word recognition experiments had indicated that such accent-independent language models significantly outperformed accent-specific models trained individually on the training set transcriptions of each accent. For phone recognition, the opposite was observed, an effect that we ascribe to the larger sizes of the phone training sets and the observation that, unlike the word sequences, the phone sequences are clearly accent-specific. Absolute discounting was used for the estimation of language model probabilities (Ney et al., 1994). The phone and word language model perplexities are shown in Tables 6 and 7 respectively.

3.6. Pronunciation Dictionaries

As part of the AST Project, five separate accent-specific English pronunciation dictionaries were compiled by human annotators, corresponding to the five English-accented AST databases described above. For our experiments, rare pronunciations were omitted without allowing training set words to be lost. Pronunciations for truncated, fragmented and mispronounced words were also not retained in the dictionaries.

In order to obtain an indication of the similarity of the five accent-specific dictionaries, we have considered the pair-wise phone alignment of corresponding pronunciations. For each pair of dictionaries, the pronunciations of all words common to both were aligned and the Levenshtein (or edit) distance was calculated. This distance is simply the sum of the minimum number of phone substitutions, insertions and deletions required to transform one pronunciation into the other. The average Levenshtein distance between each pair of dictionaries is presented in Table 8. The analysis shows that the pronunciation differences are particularly large between BE and the other accents. For example, on average 1.71 phone substitutions, insertions or deletions are required to transform a BE into an EE pronunciation, while the corresponding figure for EE and IE is just 0.60. The Levenshtein distance has been used in other studies to estimate accent and dialect similarity, for example in (ten Bosch, 2000).

For the word recognition experiments, a single pronunciation dictionary was obtained by simply pooling the five accent-specific pronunciation dictionaries. This simple approach allows the same dictionary to be used by all recogniser configurations, and ensures a single fixed vocabulary for all experiments. The alternative is to use the accent-specific dictionaries, which differ in their vocabularies and therefore lead to higher out-of-vocabulary rates on

Table 8: Average Levenshtein distances for different pairs of accent-specific pronunciation dictionaries, corresponding to the five accents of SAE. A larger value indicates a larger difference between the dictionaries.

AE	BE	CE	EE	IE	
0.0	1.52	0.85	0.79	0.92	AE
	0.0	1.50	1.71	1.71	BE
		0.0	0.68	0.79	CE
			0.0	0.60	EE
				0.0	IE

the test set. We have evaluated this use of accent-specific dictionaries in a set of experiments parallel to those that we will describe, and found that word error rates were approximately 3% higher, but that the same relative performance differences were observed between the competing systems, and that precisely the same conclusions could be drawn. We have therefore restricted ourselves to the use of the single, pooled dictionary in the following experiments. This approach has the advantage that both the pronunciation dictionary and the language model are common to all systems to be described, and that any observed performance differences must therefore be a result of differences in the acoustic modelling approaches we are investigating.

4. General Experimental Methodology

4.1. General Setup

Speech recognition systems were developed using the HTK tools (Young et al., 2009). 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 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 (Young et al., 1994). 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 tied-state triphone HMMs with three states per model and eight Gaussian mixtures per state.

4.2. Acoustic Modelling Approaches

We considered three acoustic modelling approaches. The same three approaches have been previously applied to multilingual acoustic modelling in tied-state systems (Niesler, 2007), and similar approaches were followed in (Schultz and Waibel, 2001) and in (Caballero et al., 2009) for tied-mixture topologies. The three approaches are distinguished by different methods of decision-tree state clustering:

1. *Accent-Specific Acoustic Modelling:* Accent-specific acoustic models are obtained by not allowing any sharing of data between accents. By growing separate decision-trees for the different accents, triphone HMM states are clustered separately. Only questions relating to phonetic context are employed, resulting in completely distinct sets of acoustic models for each accent.
2. *Accent-Independent Acoustic Modelling:* A single accent-independent model set is obtained by blindly pooling accent-specific data across accents for phones with the same IPA symbol. This means that phones from different accents but with the same IPA classification are considered identical. A single set of decision-trees is constructed across all accents and the clustering process employs only questions relating to phonetic context, resulting in a single, accent-independent set of triphone HMMs for all accents.
3. *Multi-Accent Acoustic Modelling:* As for accent-independent modelling, a single set of decision-trees is grown across all accents. However, in this case the decision-tree questions take into account not only the phonetic context, but also the accent of the basephone. The HMM states of triphones with the same IPA symbols but from different accents can therefore be kept separate if there is a significant acoustic difference between them, or can be merged if there is not. 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.

The overview given in Section 2 has indicated that the relative merits of the accent-specific and accent-independent modelling approaches appear to depend on the recognition setup, the corpus size and the accents involved. However, the experiments described in (Caballero et al., 2009) indicate that multi-accent modelling presents a strategy by which the choice of the training data partitioning can be achieved automatically when using tied-mixture HMMs as acoustic models. We consider whether the multi-accent modelling approach can yield similar improvements for accents of SAE, and when using tied-state acoustic models, which may behave differently.

4.3. System Configuration and Optimisation

The three acoustic modelling approaches described in Section 4.2 were applied to the combination of the Afrikaans English (AE), Black South African

English (BE), Cape Flats English (CE), White South African English (EE), and Indian South African English (IE) training sets described in Section 3.3.

Our experiments consider the scenario where the accent of each test utterance is assumed to be known during testing, as was also assumed in (Caballero et al., 2009). For each of the acoustic modelling approaches considered, evaluation involved presenting each test utterance only to the recogniser matching the accent of that utterance. By configuring the recognition setup in this way, we are isolating the effect of the acoustic models on recognition performance, and are not taking into account the effects of accent misclassifications, which would occur if the accent were unknown during testing.

Initial parameter optimisation on the development set indicated that recognition performance measured separately for each accent and each acoustic modelling approach was very robust toward the word insertion penalty (WIP) and language model scaling factor (LMS). The optimal WIP and LMS values for the individual accents and acoustic modelling approaches were also very similar. Based on this initial optimisation on the development set, a single pair of WIP and LMS values was used across accents and acoustic modelling approaches in all experiments.

The initial development set optimisation did however indicate that recognition performance was sensitive to the number of independent parameters used by the acoustic model set. Several sets of HMMs were therefore produced by varying the likelihood improvement threshold used during decision-tree state clustering. For each acoustic modelling approach, this value was then optimised separately on the development set. However, for a particular acoustic modelling approach, the same threshold value was used regardless of the accent. The minimum cluster occupancy was set to 100 frames for all experiments.

5. Experimental Results and Analysis

5.1. Analysis of Phone Recognition Performance

Figure 2 shows the average phone recognition accuracy measured on the evaluation set using the eight-mixture triphone models. Due to the sensitivity of this accuracy on the number of physical states used by the acoustic models, recognition performance was determined for a range of model sizes. The particular acoustic models which deliver optimal performance on the development set are indicated by circular markers, and these systems will in the following be referred to as the optimal systems.

For each acoustic modelling approach a single curve indicating the average accuracy between the five accents is shown. The number of physical states for the accent-specific systems is taken to be the sum of the number of unique states in each component accent-specific HMM set. The number of physical states for the multi-accent systems is taken to be the total number of unique states remaining after decision-tree state clustering and hence takes cross-accent sharing into account. For all three approaches, the number of physical states in the acoustic model is therefore directly related to the total number of independent parameters. Hence, systems containing the same number of independent parameters are aligned vertically in Figure 2.

The results presented in Figure 2 show that multi-accent acoustic modelling and accent-independent modelling both yield consistently superior performance compared to accent-specific modelling. Except for one system (13 681 states), all multi-accent systems outperform their accent-independent counterparts. Table 9 summarises the performance and the number of states for the systems giving optimal average performance on the development set, as indicated by the circular markers in Figure 2. For these optimal systems, multi-accent acoustic modelling outperforms both accent-specific and accent-independent modelling. The absolute improvements in phone accuracies (in the order of 0.2% or more) are higher than those achieved for a set of similar experiments performed for multiple languages, where improvements were in the order of 0.1% (Niesler, 2007).

Since the triphone clustering process optimises the overall likelihood, improvements for each individual accent are not guaranteed. Table 10 shows the per-accent phone accuracies for the optimal systems. These results show that multi-accent acoustic models improve the phone recognition accuracy relative to the remaining two approaches for CE and IE. For BE, accent-specific (i.e. separate) modelling yields the best performance while for AE and EE, accent-independent modelling (i.e. data pooling) leads to the best performance. Nevertheless, the average accuracy over all five accents is highest for the multi-accent models.

Table 10 also shows that for AE, CE and EE it is better to pool the training data and build an accent-independent acoustic model set than to build separate, accent-specific models. For BE, on the other hand, it is better to do the opposite. For IE, the accuracies are very similar. In contrast, when considering multilingual speech recognition, it is always better to train separate, language-specific acoustic models (Schultz and Waibel, 2001; Niesler, 2007). A further important observation, which will be supported by the word

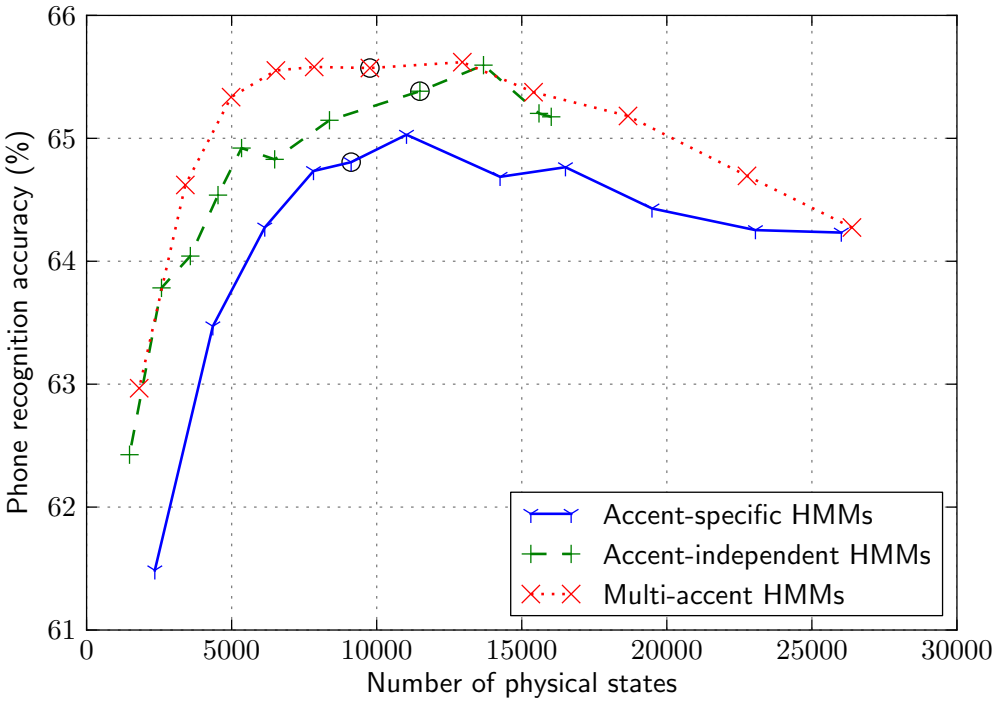

Figure 2: Average evaluation set phone accuracies of accent-specific, accent-independent and multi-accent systems as a function of the total number of distinct HMM states. Circular markers indicate the systems delivering optimal performance on the development set.

Table 9: The number of states and evaluation set phone accuracies for the optimal systems identified by the circular markers in Figure 2.

Model set	No. of states	Accuracy (%)
Accent-specific	9119	64.81
Accent-independent	11 489	65.38
Multi-accent	9765	65.57

Table 10: Phone accuracies for each accent individually using the optimal systems.

Model set	AE	BE	CE	EE	IE	Average
Accent-specific	64.80	56.77	65.23	72.97	64.27	64.81
Accent-independent	66.51	55.61	66.07	74.44	64.40	65.38
Multi-accent	66.48	56.69	66.34	73.79	64.66	65.57

Table 11: The number of states and evaluation set word accuracies for the optimal systems identified by the circular markers in Figure 3.

Model set	No. of states	Accuracy (%)
Accent-specific	6141	81.53
Accent-independent	2582	81.52
Multi-accent	4982	82.78

recognition experiments, is that, while the decision to pool or to separate the training data depends on the particular accent in question, multi-accent modelling allows almost all of this gain to be obtained in a data-driven manner.

5.2. Analysis of Word Recognition Performance

Figure 3 shows the average word recognition accuracy measured on the evaluation set using the eight-mixture triphone models. For each acoustic modelling approach a single curve indicating the average accuracy between the five accents is shown. Once again, a range of acoustic models with differing numbers of physical states (and therefore differing numbers of independent parameters) are considered, with the systems leading to optimal performance on the development sets identified by circular markers. Figure 3 indicates that, over the range of models considered, multi-accent modelling consistently outperforms both accent-specific and accent-independent acoustic modelling. The results for the optimal systems are summarised in Table 11. The performance improvements exhibited by the optimal multi-accent system relative to both the optimal accent-specific as well as the optimal accent-independent system were found to be significant at the 99.9% level using bootstrap confidence interval estimation (Bisani and Ney, 2004).

Table 12 presents the word accuracies of the optimal systems separately

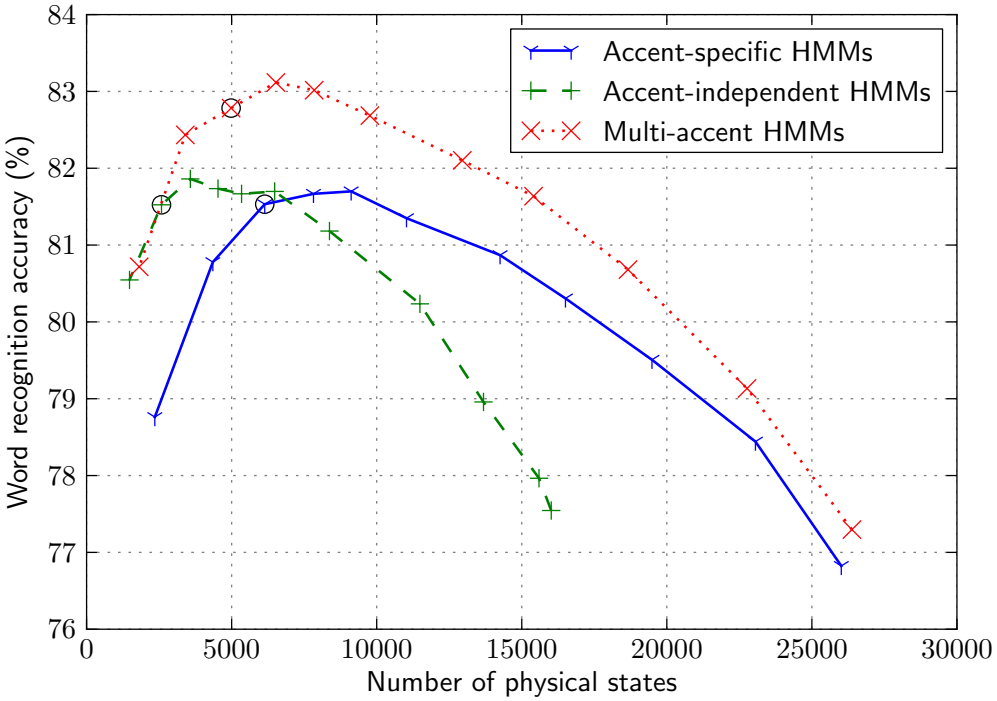

Figure 3: Average evaluation set word accuracies of accent-specific, accent-independent and multi-accent systems as a function of the total number of distinct HMM states. Circular markers indicate the systems delivering optimal performance on the development set.

for each accent. For all accents best performance is achieved using the multi-accent models. For CE, EE and IE accent-independent models obtained by pooling the training data result in better performance than is achieved by accent-specific modelling. The opposite is true for BE, while the two approaches yield very similar results for AE. Hence it is once again apparent that the decision of whether to pool or to separate the training data depends on the accents in question. The application of multi-accent acoustic modelling allows this decision to be avoided, and sharing to be configured in a data-driven manner instead.

Interestingly, Table 12 indicates that the accent-specific and multi-accent modelling approaches yield slightly better performance for AE than for EE, while this was not true for the corresponding phone accuracies in Table 10. We believe that this can be attributed to the word language model perplexities, which are lower for AE than for EE (Table 7). In contrast, the phone

Table 12: Word accuracies for each accent individually using the optimal systems.

Model set	AE	BE	CE	EE	IE	Average
Accent-specific	84.72	72.84	83.57	84.15	82.54	81.53
Accent-independent	84.72	71.10	83.86	84.90	83.16	81.52
Multi-accent	86.65	73.71	85.00	85.29	83.49	82.78

language model perplexity is substantially higher for AE than for EE (Table 6). This discrepancy may, however, at least partially, be a result of the different test sets used for the different accents. Furthermore, we have noticed that the English proficiency of the Afrikaans speakers in the AE data is very high in general, which may explain the relatively high AE recognition accuracies.

5.3. Analysis of Decision-Trees

Inspection of the type of questions most frequently used during clustering reveals that accent-based questions are most common at the root nodes of the decision-trees and become increasingly less frequent towards the leaves. Figure 4 analyses the decision-trees of the multi-accent system delivering optimal word accuracy (4982 states, Table 11). The figure shows that approximately 47% of all questions at the root nodes are accent-based and that this proportion drops to 34% and 29% for the roots' children and grandchildren respectively. For the first, second and third levels of depth, BE and IE questions are asked most often. This indicates that, close to the root nodes, separation of IE and BE states tend to occur more frequently than for the other accents. As discovered in Sections 5.1 and 5.2, BE was also the only accent for which accent-dependent (i.e. separate) modelling led to higher phone and word recognition accuracies compared to accent-independent (i.e. pooled) models.

The contribution to the log likelihood improvement made by the accent-based and phonetically-based questions respectively during the decision-tree growing process are shown in Figure 5 as a function of depth within the decision-trees. The analysis indicates that phonetically-based questions make a larger contribution to the log likelihood improvement than the accent-based questions at all levels in the decision-trees. In Figure 5, approximately 37%

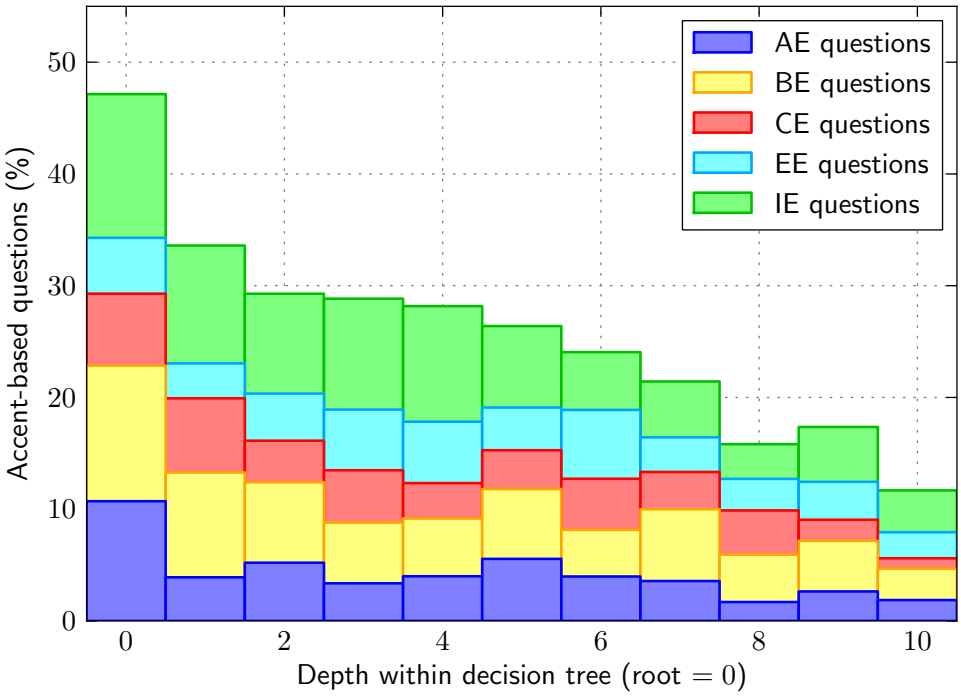

Figure 4: Analysis showing the percentage of questions that are accent-based at various depths within the multi-accent decision-trees for the multi-accent system with optimal word accuracy. The questions enquire whether the accent of a basephone is either AE, BE, CE, EE or IE and the individual proportions of these questions are also shown.

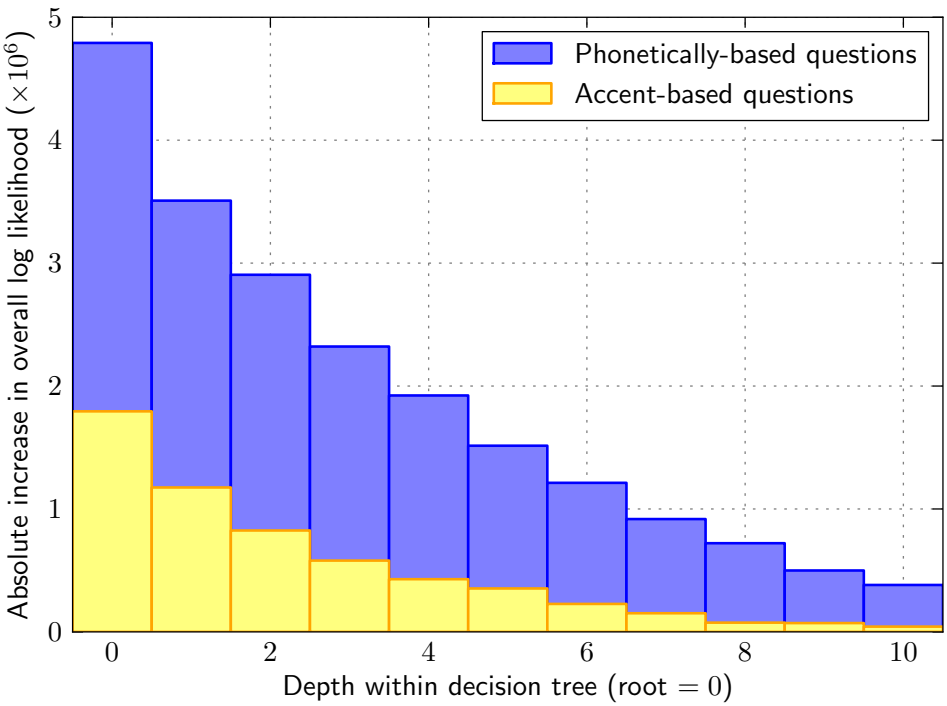

Figure 5: Analysis showing the contribution made to the increase in overall log likelihood by the accent-based questions and phonetically-based questions respectively for the multi-accent system with optimal word accuracy.

of the total increase at the root nodes is afforded by accent-based questions. This proportion is lower than the corresponding figure of 74% for multilingual acoustic modelling (Niesler, 2007). While Figures 4 and 5 both analyse the decision-trees of the multi-accent system with optimal development-set word accuracy, a repetition of the same analysis for the multi-accent system with optimal phone accuracy, as well as for the largest multi-accent system, revealed similar trends.

5.4. Analysis of Cross-Accent Data Sharing

In order to determine to what extent and for which accents data sharing ultimately takes place for a multi-accent system, we considered the proportion of decision-tree leaf nodes (which correspond to the state clusters) that are populated by states from exactly one, two, three, four or all five accents respectively. A cluster populated by states from a single accent indicates that no sharing is taking place, while a cluster populated by states from all five accents indicates that sharing is taking place across all accents. Figure 6 illustrates how these proportions change as a function of the total number of clustered states in a system.

From Figure 6 it is apparent that, as the number of clustered states increases, so does the proportion of clusters containing a single accent. This indicates that the multi-accent decision-trees tend towards separate clusters for each accent as the likelihood improvement threshold is lowered, as one might expect. The proportion of clusters containing two, three, four or all five accents show a commensurate decrease as the number of clustered states increase. For the multi-accent system yielding optimal phone recognition accuracy (9765 states, Table 9), approximately 33% of state clusters contain a mixture of accents, while 44% of state clusters contain a mixture of accents for the optimal word recognition system (4982 states, Table 11). This demonstrates that a considerable degree of sharing is taking place across accents. In contrast, for a comparable multilingual system, only 20% of state clusters contained more than one language (Niesler, 2007).

In order to determine which accents are being shared most often by the clustering process, Figures 7, 8 and 9 analyse the proportion of state clusters consisting of groups of two, three and four accents respectively. Proportions for the combinations not shown fall below 0.5%. It is evident from Figure 7 that the largest proportion of two-accent clusters are due to the combination of AE and EE and of AE and CE. All other combinations are far less common. In Section 3.2 the influence of Afrikaans on EE was noted, and this may

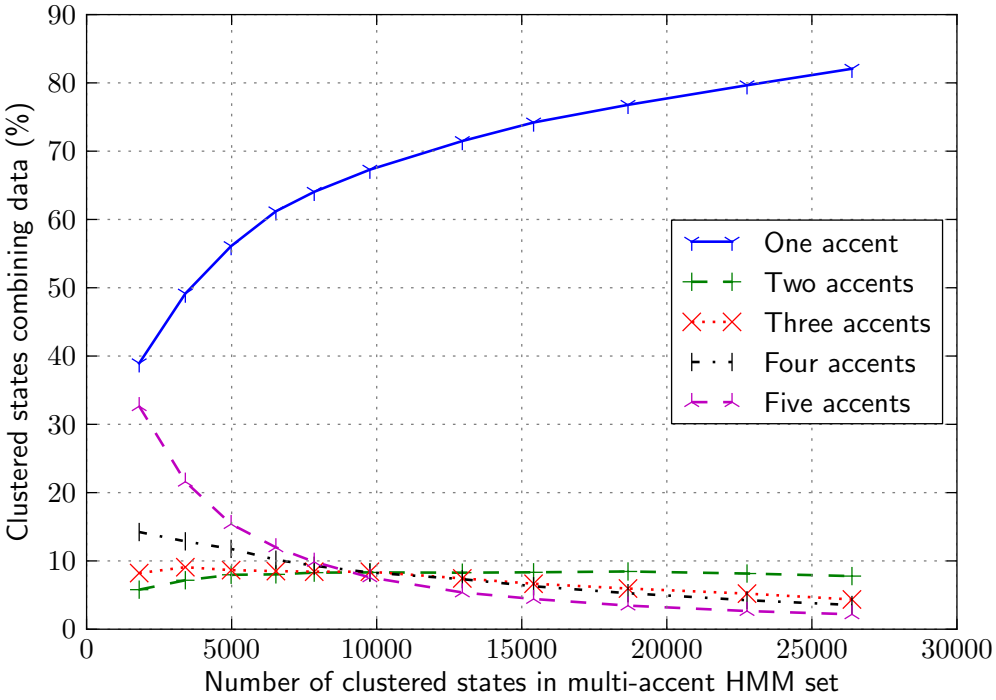

Figure 6: Proportion of state clusters combining data from one, two, three, four or five accents.

account for a higher degree of similarity between these accents. The influence of Afrikaans on CE and the use of Afrikaans as a first language by many CE speakers may in turn explain a phonetic similarity and therefore higher degree of sharing between AE and CE. Figure 8 indicates that AE, CE and EE are the most frequent three-accent combination, followed by the combination of BE, CE and EE. Furthermore, Figure 9 shows that the two most frequent four-accent combinations are AE, CE, EE, IE and AE, BE, CE, EE which both include AE, CE and EE. The similarity of these three accents is therefore emphasised in all three figures.

In order to determine which accents are being separated most often from the others during the clustering process, Figure 10 presents the proportion of state clusters consisting of just one accent. The most striking feature is the high degree of separation of IE. BE is found in single-accent clusters second most often, with the remaining three accents following. Figure 10 lends further support to our conclusion that AE, CE and EE are most similar,

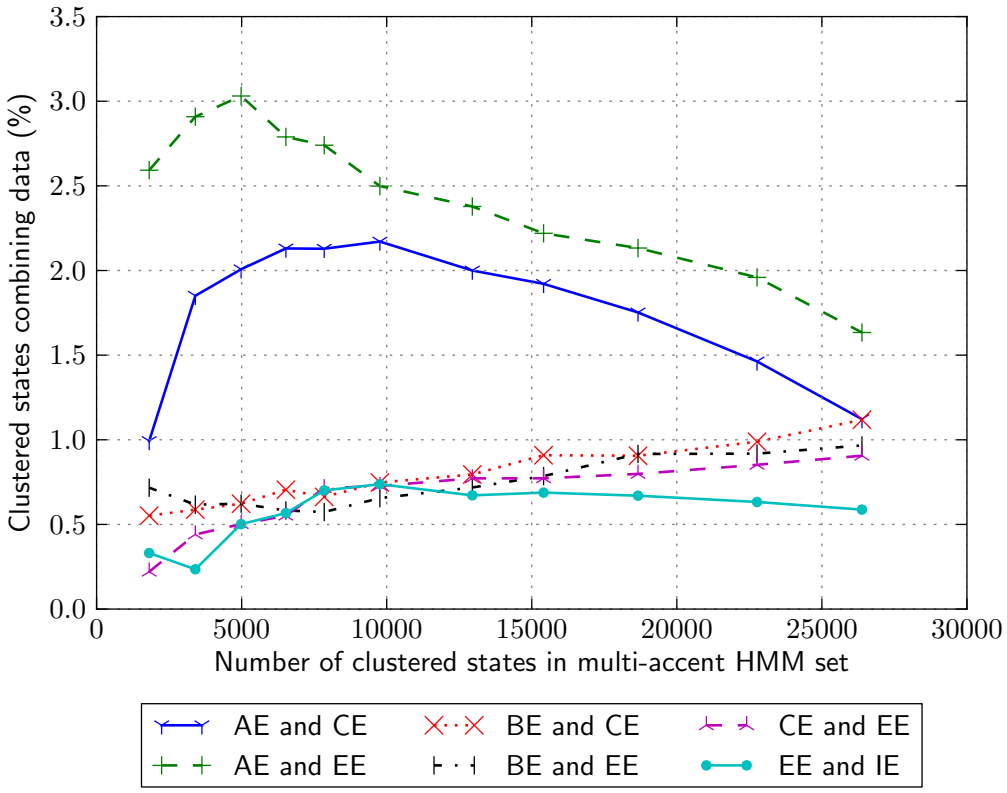

Figure 7: Proportion of state clusters combining data from various combinations of two accents.

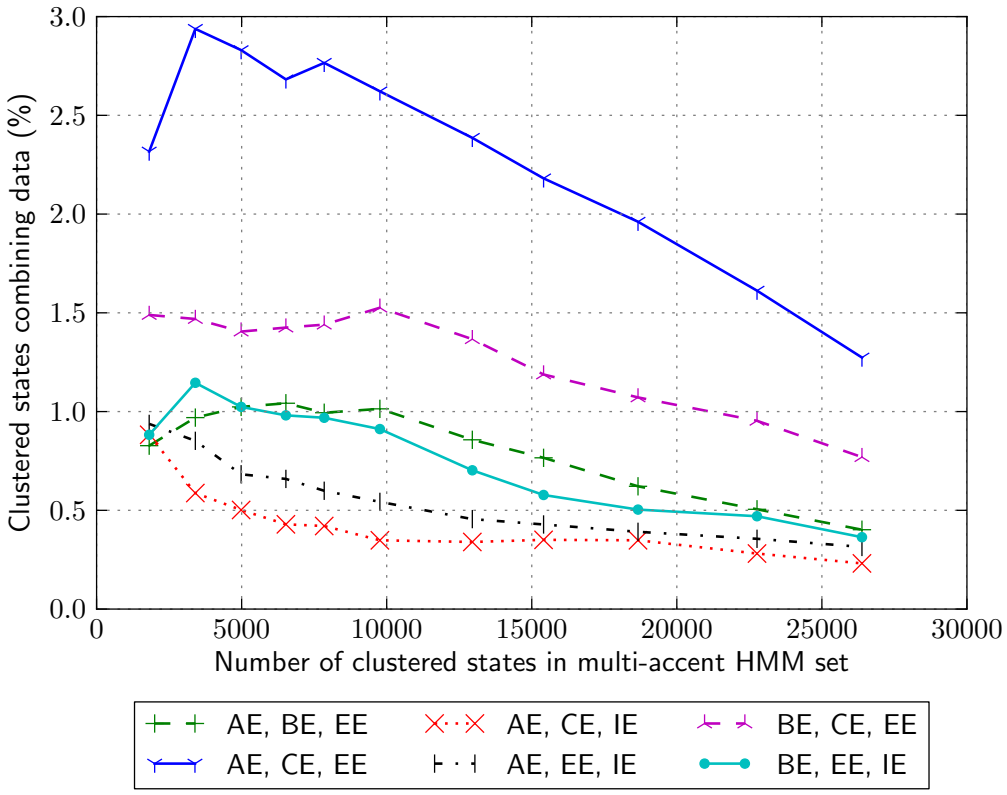

Figure 8: Proportion of state clusters combining data from various combinations of three accents.

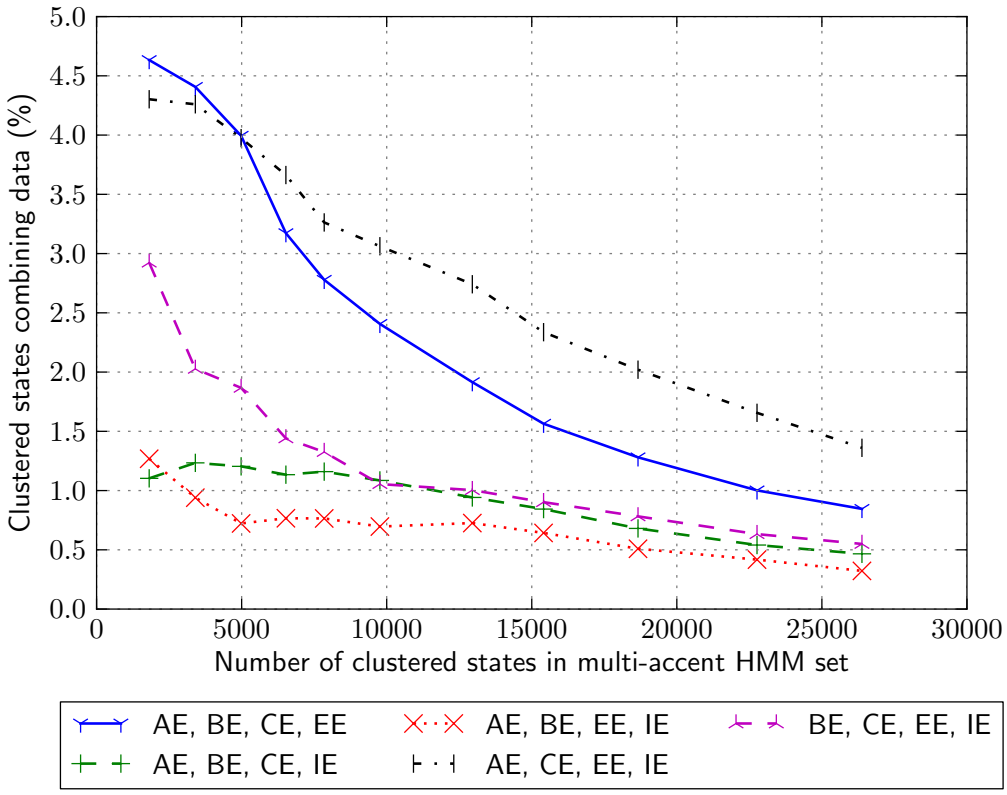

Figure 9: Proportion of state clusters combining data from various combinations of four accents.

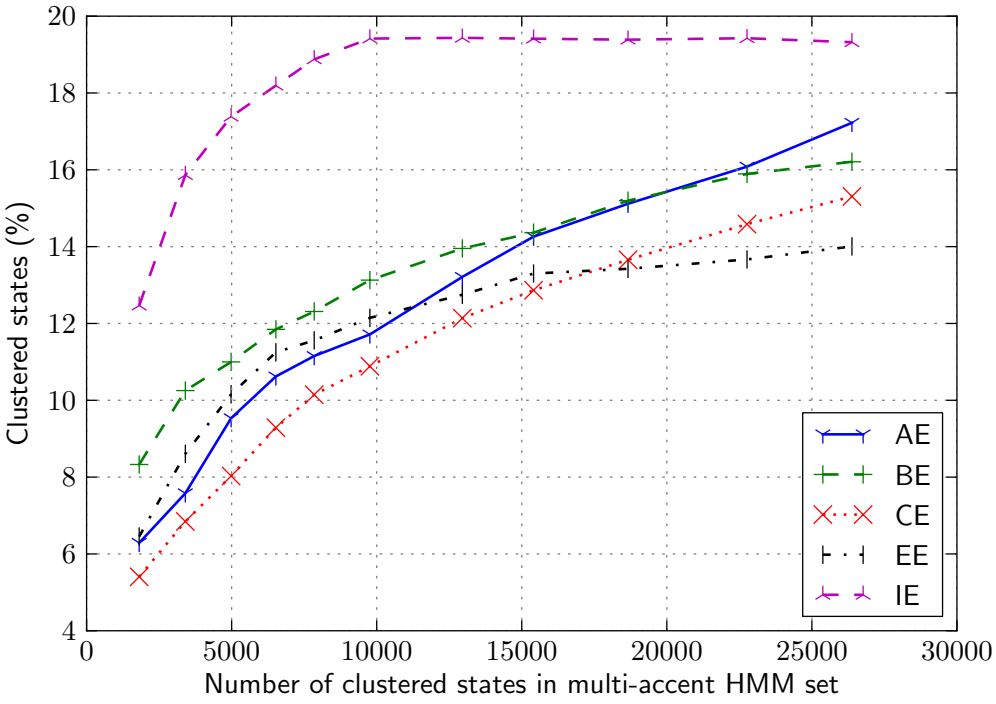

Figure 10: Proportion of state clusters containing data from just one accent.

since these three accents occur least frequently in single-accent clusters.

Finally, in order to obtain a further perspective on the results presented in Figures 7, 8 and 9, Figure 11 analyses the proportion of decision-tree leaf nodes that are populated by at least two, at least three, at least four, or all five accents. This figure also indicates the proportion of state clusters containing states from two or more of the members of the (AE, CE, EE) group of accents. For the multi-accent system yielding optimal phone recognition accuracy (9765 states, Table 9), approximately 21% of state clusters contain a mixture of these three accents, while 24% of state clusters contain a mixture of these accents for the optimal word recognition system (4982 states, Table 11). This again highlights the similarity of the AE, CE and EE accents.

6. Estimation of Acoustic Model Similarity

To obtain some further intuition regarding the relative similarity between the five SAE accents, we considered a computational method allowing the similarity between the respective acoustic models to be estimated. This can be achieved by determining the similarity between the probability density functions (PDFs) associated with two sets of HMMs. Several such measures have been proposed in the literature, including the Kullback-Leibler divergence and the Bhattacharyya bound (Vihola et al., 2002; Olsen and Hershey, 2007). We have employed the Bhattacharyya bound, which is a widely-used upper bound for the Bayes error of a classifier, and has been used in several other studies (Mak and Barnard, 1996; Badenhorst and Davel, 2008; Hershey and Olsen, 2008). The Bhattacharyya bound provides a simple closed-form solution for Gaussian distributions, and is easy to interpret because it is based on the Bayes error.

Although the analysis of pronunciation dictionary similarity presented in Section 3.6 and Table 8 might also provide indications of accent similarity, the decision-tree clustering employed in the multi-accent modelling approach is based on acoustic similarity. The analysis which follows should therefore give a better indication of the sharing potential between accents.

6.1. The Bhattacharyya Bound

Given the PDFs $p_1(\mathbf{x})$ and $p_2(\mathbf{x})$ and associated prior probabilities P_1 and P_2, the Bayes error is given by (Fukunaga, 1990):

$$\varepsilon = \int \min \Big[P_1\, p_1(\mathbf{x}), P_2\, p_2(\mathbf{x}) \Big] \, \mathrm{d}\mathbf{x} \tag{1}$$

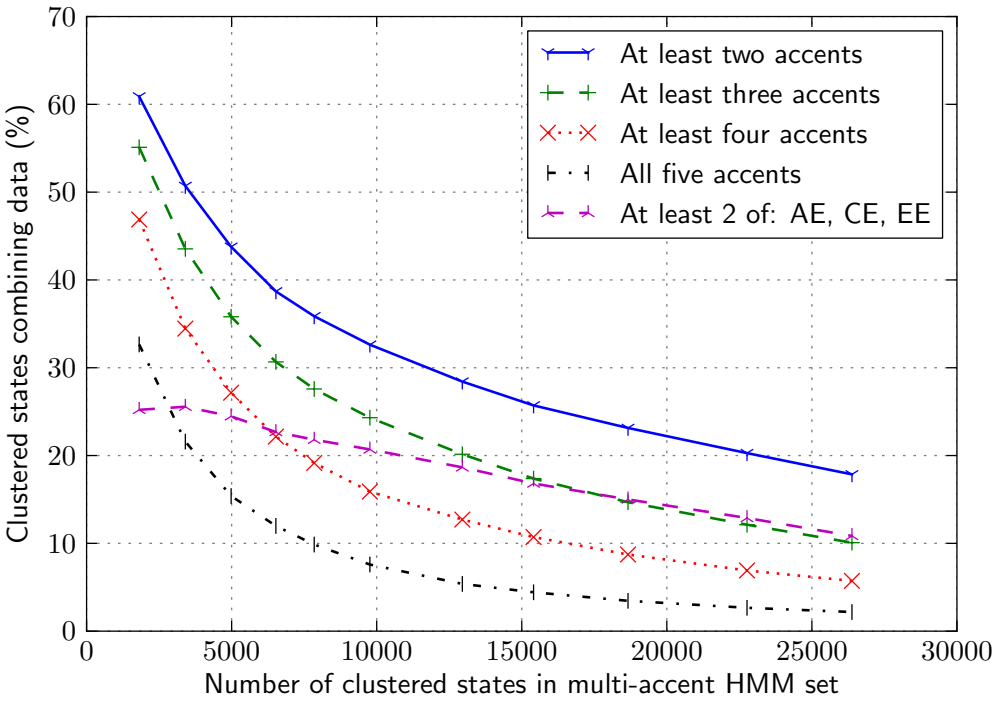

Figure 11: Proportion of state clusters combining data from at least two, at least three, at least four, or all five accents. In addition, the proportion of state clusters combining data from at least two of the members of the (AE, CE, EE) group of accents is shown.

By using the identity

$$\min[a, b] \leq a^s b^{1-s} \text{ for } 0 \leq s \leq 1 \tag{2}$$

with $a, b \geq 0$, an upper bound of equation (1) can be determined. If we do not insist on the optimisation of s and choose $s = 1/2$, we obtain the Bhattacharyya bound:

$$\varepsilon_u = \sqrt{P_1\, P_2} \int \sqrt{p_1(\mathbf{x})\, p_2(\mathbf{x})}\, \mathrm{d}\mathbf{x} \;\; \geq \varepsilon \tag{3}$$

When both $p_1(\mathbf{x})$ and $p_2(\mathbf{x})$ are Gaussian with means μ_i and covariance matrices Σ_i, the closed-form expression for ε_u can be found to be (Fukunaga, 1990):

$$\varepsilon_u = \sqrt{P_1\, P_2}\; e^{-D} \tag{4}$$

where

$$D = \frac{1}{8}(\mu_2 - \mu_1)^{\mathrm{T}} \left[\frac{\Sigma_1 + \Sigma_2}{2}\right]^{-1} (\mu_2 - \mu_1) + \frac{1}{2} \ln \frac{\left|\frac{\Sigma_1 + \Sigma_2}{2}\right|}{\sqrt{|\Sigma_1|\,|\Sigma_2|}} \tag{5}$$

The term D is known as the Bhattacharyya distance. When we assume the prior probabilities to be equal, as suggested by Mak and Barnard (1996) and by Badenhorst and Davel (2008), ε_u is bounded $0 \leq \varepsilon_u \leq 1/2$ with $\varepsilon_u = 1/2$ when the PDFs are identical. Increased similarity between PDFs is thus indicated when ε_u approaches $1/2$.

6.2. Similarity between accent pairs

In order to verify our interpretation of the experimental results presented in Section 5, and of our intuition regarding the five accents considered, we used the Bhattacharyya bound to compute the degree of similarity between each pair of accents. The single-mixture monophone HMMs which were trained for each accent as part of the model development procedure described in Section 4.1 were used for this purpose, since the Bhattacharyya bound cannot easily be determined for Gaussian mixture densities. For each monophone, the average of the three bounds calculated between corresponding HMM states was obtained. This gives a measure of between-accent similarity for a particular monophone. Finally, a weighted average of these similarities is computed, where each individual similarity is weighted by the frequency

Table 13: Average Bhattacharyya bounds for different pairs of SAE accents. A value of $\varepsilon_u = 1/2$ indicates identical models, and increased similarity between accents is indicated by ε_u approaching 1/2.

AE	BE	CE	EE	IE	
0.5	0.3101	0.4030	0.3929	0.3302	AE
	0.5	0.3516	0.3266	0.3266	BE
		0.5	0.3679	0.3629	CE
			0.5	0.3670	EE
				0.5	IE

of occurrence of the phone in the training set. This final figure gives an indication of the similarity between two acoustic models.

A comparison for the five SAE accents using the approach described above is presented in Table 13. From this table it is evident that the highest degree of similarity exists between the AE, CE and EE accents. BE appears to be the most different from the other accents, showing the lowest similarity of 0.3101 with AE. In general, the similarity values in Table 13 appear to indicate that AE, CE and EE are rather similar, with IE lying further away and BE being the most different. This is in agreement with our interpretation of the experimental results presented in Section 5.

7. Summary and Conclusions

We have presented the evaluation of three approaches to multi-accent acoustic modelling for five accents of South African English (SAE): Afrikaans English (AE), Black South African English (BE), Cape Flats English (CE), White South African English (EE), and Indian South African English (IE). These English accents can be regarded as under-resourced, since very little annotated speech data and very few associated resources (such as pronunciation dictionaries) are currently available. Tied-state multi-accent acoustic models, obtained by introducing accent-based questions in the decision-tree clustering process and thus allowing for selective sharing between accents, were found to yield improved performance compared to both accent-independent models, obtained by simply pooling data across accents, and accent-specific models obtained by separating training data for each accent. These improvements were obtained for both phone and word recognition accuracies, and were found to be statistically significant at the 99.9% level for

the latter. It was also found that the relative merits of pooling training data (for accent-independent modelling) and separating training data (for accent-specific modelling) depend on the particular accents in question. In particular, AE, CE, EE and IE are best pooled, while separate modelling is more advantageous for BE. Nevertheless, the application of multi-accent modelling results in average improvements over both alternatives. This allows the choice of the training data partitioning strategy to become data-driven, and eliminates the need to develop and compare multiple systems when new datasets are used.

Analysis of the decision-trees constructed during the multi-accent modelling process showed that questions relating to phonetic context resulted in a much larger contribution to the likelihood increase than the accent-based questions, indicating that the multi-accent models gain more from combined modelling than from separation. The decision-tree analysis also indicated that the AE, CE and EE accents were combined most often, while the BE and IE accents were frequently modelled separately. This was supported by an analysis of the relative similarities between the different SAE accents using the Bhattacharyya bound. When contrasted with a comparable set of experiments for multilingual acoustic modelling, we find that there is substantially more data sharing for multi-accent systems, and that the improvements in recognition accuracy are larger and statistically more significant.

8. Future work

In this paper we have considered multi-accent acoustic modelling for five accents of SAE based on experiments where the accent of the test data is assumed to be known. Future work will focus on the integration of the different acoustic modelling approaches into a single multi-accent ASR system for which the accent of the input speech is unknown.

Acknowledgements

The authors would like to thank Febe de Wet for her helpful comments and suggestions. 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. Parts of this work were executed using the High Performance Computer (HPC) facility at Stellenbosch University.

References

Badenhorst, J.A.C., Davel, M.H., 2008. Data requirements for speaker independent acoustic models, in: Proc. PRASA, Cape Town, South Africa.

Beattie, V., Edmondson, S., Miller, D., Patel, Y., Talvola, G., 1995. An integrated multi-dialect speech recognition system with optional speaker adaptation, in: Proc. Eurospeech, Madrid, Spain. pp. 1123–1126.

Bisani, M., Ney, H., 2004. Bootstrap estimates for confidence intervals in ASR performance evaluation, in: Proc. ICASSP, Montreal, Quebec, Canada. pp. 409–412.

Bowerman, S., 2004a. White South African English: morphology and syntax, in: Kortmann, B., Burridge, K., Mesthrie, R., Schneider, E.W., Upton, C. (Eds.), A Handbook of Varieties of English, Mouton de Gruyter, Berlin, Germany. pp. 949–961.

Bowerman, S., 2004b. White South African English: phonology, in: Schneider, E.W., Burridge, K., Kortmann, B., Mesthrie, R., Upton, C. (Eds.), A Handbook of Varieties of English, Mouton de Gruyter, Berlin, Germany. pp. 931–942.

Caballero, M., Moreno, A., Nogueiras, A., 2009. Multidialectal Spanish acoustic modeling for speech recognition. Speech Commun. 51, 217–229.

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

Crystal, D., 1991. A Dictionary of Linguistics and Phonetics. Blackwell Publishers, Oxford, UK. third edition.

Despres, J., Fousek, P., Gauvain, J.L., Gay, S., Josse, Y., Lamel, L., Messaoudi, A., 2009. Modeling Northern and Southern varieties of Dutch for STT, in: Proc. Interspeech, Brighton. pp. 96–99.

Diakoloukas, V., Digalakis, V., Neumeyer, L., Kaja, J., 1997. Development of dialect-specific speech recognizers using adaptation methods, in: Proc. ICASSP, Munich, Germany. pp. 1455–1458.

Finn, P., 2004. Cape Flats English: phonology, in: Schneider, E.W., Burridge, K., Kortmann, B., Mesthrie, R., Upton, C. (Eds.), A Handbook of Varieties of English, Mouton de Gruyter, Berlin, Germany. pp. 964–984.

Fischer, V., Gao, Y., Janke, E., 1998. Speaker-independent upfront dialect adaptation in a large vocabulary continuous speech recognizer, in: Proc. ICSLP, Sydney, Australia. pp. 787–790.

Fukunaga, K., 1990. Introduction to Statistical Pattern Recognition. Academice Press, San Diego, CA. second edition.

Hershey, J.R., Olsen, P.A., 2008. Variational Bhattacharyya divergence for hidden Markov models, in: Proc. ICASSP, Las Vegas, NV. pp. 4557–4560.

Humphries, J.J., Woodland, P.C., 1997. Using accent-specific pronunciation modelling for improved large vocabulary continuous speech recognition, in: Proc. Eurospeech, Rhodes, Greece. pp. 2367–2370.

Katz, S.M., 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. IEEE Trans. Acoust., Speech, Signal Process. 35, 400–401.

Kirchhoff, K., Vergyri, D., 2005. Cross-dialectal data sharing for acoustic modeling in Arabic speech recognition. Speech Commun. 46, 37–51.

Mak, B., Barnard, E., 1996. Phone clustering using the Bhattacharyya distance, in: Proc. ICSLP, Philadelphia, PA. pp. 2005–2008.

McCormick, K., 2004. Cape Flats English: morphology and syntax, in: Kortmann, B., Burridge, K., Mesthrie, R., Schneider, E.W., Upton, C. (Eds.), A Handbook of Varieties of English, Mouton de Gruyter, Berlin, Germany. pp. 993–1005.

Mesthrie, R., 2004a. Indian South African English: morphology and syntax, in: Kortmann, B., Burridge, K., Mesthrie, R., Schneider, E.W., Upton, C. (Eds.), A Handbook of Varieties of English, Mouton de Gruyter, Berlin, Germany. pp. 974–992.

Mesthrie, R., 2004b. Indian South African English: phonology, in: Schneider, E.W., Burridge, K., Kortmann, B., Mesthrie, R., Upton, C. (Eds.), A Handbook of Varieties of English, Mouton de Gruyter, Berlin, Germany. pp. 953–963.

Ney, H., Essen, U., Kneser, R., 1994. On structuring probabilistic dependencies in stochastic language modelling. Comput. Speech Lang. 8, 1–38.

Niesler, T.R., 2007. Language-dependent state clustering for multilingual acoustic modelling. Speech Commun. 49, 453–463.

Olsen, P.A., Hershey, J.R., 2007. Bhattacharyya error and divergence using variational importance sampling, in: Proc. Interspeech, Antwerp, Belgium. pp. 46–49.

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

Schneider, E.W., Burridge, K., Kortmann, B., Mesthrie, R., Upton, C. (Eds.), 2004. A Handbook of Varieties of English. Mouton de Gruyter, Berlin, Germany.

Schultz, T., Waibel, A., 2001. Language-independent and language-adaptive acoustic modeling for speech recognition. Speech Commun. 35, 31–51.

Statistics South Africa, 2004. Census 2001: Primary tables South Africa: Census 1996 and 2001 compared.

Stolcke, A., 2002. SRILM – An extensible language modeling toolkit, in: Proc. ICSLP, Denver, CO. pp. 901–904.

ten Bosch, L., 2000. ASR, dialects, and acoustic/phonological distances, in: Proc. ICSLP, Beijing, China. pp. 1009–1012.

Van Compernolle, D., Smolders, J., Jaspers, P., Hellemans, T., 1991. Speaker clustering for dialectic robustness in speaker independent recognition, in: Proc. Eurospeech, Genove, Italy. pp. 723–726.

Van Rooy, B., 2004. Black South African English: phonology, in: Schneider, E.W., Burridge, K., Kortmann, B., Mesthrie, R., Upton, C. (Eds.), A Handbook of Varieties of English, Mouton de Gruyter, Berlin, Germany. pp. 943–952.

Vihola, M., Harju, M., Salmela, P., Suontausta, J., Savela, J., 2002. Two dissimilarity measures for HMMs and their application in phoneme model clustering, in: Proc. ICASSP, Orlando, FL. pp. 933–936.

Wang, Z., Schultz, T., Waibel, A., 2003. Comparison of acoustic model adaptation techniques on non-native speech, in: Proc. ICASSP, Hong Kong. pp. 540–543.

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

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