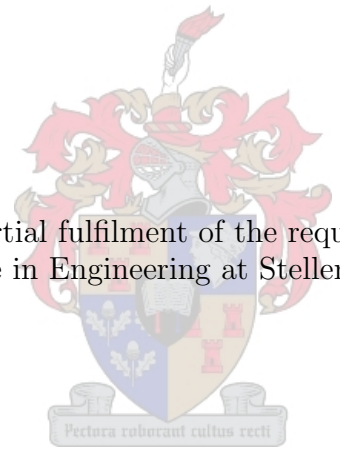
Speech Recognition of South African English Accents

by

Herman Kamper



Thesis presented in partial fulfilment of the requirements for the degree
Master of Science in Engineering at Stellenbosch University.

Supervisor: Prof. T. R. Niesler
Department of Electrical and Electronic Engineering

March 2012

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: March 2012

Summary

Several accents of English are spoken in South Africa. Automatic speech recognition (ASR) systems should therefore be able to process the different accents of South African English (SAE). In South Africa, however, system development is hampered by the limited availability of speech resources. In this thesis we consider different acoustic modelling approaches and system configurations in order to determine which strategies take best advantage of a limited corpus of the five accents of SAE for the purpose of ASR. Three acoustic modelling approaches are considered: (i) accent-specific modelling, in which accents are modelled separately; (ii) accent-independent modelling, in which acoustic training data is pooled across accents; and (iii) multi-accent modelling, which allows selective data sharing between accents. For the latter approach, selective sharing is enabled by extending the decision-tree state clustering process normally used to construct tied-state hidden Markov models (HMMs) by allowing accent-based questions.

In a first set of experiments, we investigate phone and word recognition performance achieved by the three modelling approaches in a configuration where the accent of each test utterance is assumed to be known. Each utterance is therefore presented only to the matching model set. We show that, in terms of best recognition performance, the decision of whether to separate or to pool training data depends on the particular accents in question. Multi-accent acoustic modelling, however, allows this decision to be made automatically in a data-driven manner. When modelling the five accents of SAE, multi-accent models yield a statistically significant improvement of 1.25% absolute in word recognition accuracy over accent-specific and accent-independent models.

In a second set of experiments, we consider the practical scenario where the accent of each test utterance is assumed to be unknown. Each utterance is presented simultaneously to a bank of recognisers, one for each accent, running in parallel. In this setup, accent identification is performed implicitly during the speech recognition process. A system employing multi-accent acoustic models in this parallel configuration is shown to achieve slightly improved performance relative to the configuration in which the accents are known. This demonstrates that accent identification errors made during the parallel recognition process do not affect recognition performance. Furthermore, the parallel approach is also shown to outperform an accent-independent system obtained by pooling acoustic and language model training data.

In a final set of experiments, we consider the unsupervised reclassification of training set accent labels. Accent labels are assigned by human annotators based on a speaker's mother-tongue or ethnicity. These might not be optimal for modelling purposes. By classifying the accent of each utterance in the training set by using first-pass acoustic models and then retraining the models, reclassified acoustic models are obtained. We show that the proposed relabelling procedure does not lead to any improvements and that training on the originally labelled data remains the best approach.

Opsomming

Verskeie aksente van Engels word in Suid Afrika gepraat. Outomatiese spraakherkenningstelsels moet dus in staat wees om verskillende aksente van Suid Afrikaanse Engels (SAE) te kan hanteer. In Suid Afrika word die ontwikkeling van spraakherkenningstegnologie egter deur die beperkte beskikbaarheid van geannoteerde spraakdata belemmer. In hierdie tesis ondersoek ons verskillende akoestiese modelleringstegnieke en stelselkonfigurasies ten einde te bepaal watter strategieë die beste gebruik maak van 'n databasis van die vyf aksente van SAE. Drie akoestiese modelleringstegnieke word ondersoek: (i) aksent-spesifieke modellering, waarin elke aksent apart gemodelleer word; (ii) aksent-onafhanklike modellering, waarin die akoestiese afrigdata van verskillende aksente saamgegooi word; en (iii) multi-aksent modellering, waarin data selektief tussen aksente gedeel word. Vir laasgenoemde word selektiewe deling moontlik gemaak deur die besluitnemingsboom-toestandbondeling-algoritme, wat gebruik word in die afrig van gebinde-toestand verskuilde Markov-modelle, uit te brei deur aksent-gebaseerde vrae toe te laat.

In 'n eerste stel eksperimente word die foon- en woordherkenningsakkuraathede van die drie modelleringstegnieke vergelyk in 'n konfigurasie waarin daar aanvaar word dat die aksent van elke toetsspraakdeel bekend is. In hierdie konfigurasie word elke spraakdeel slegs gebied aan die modelstel wat ooreenstem met die aksent van die spraakdeel. In terme van herkenningsakkuraathede, wys ons dat die keuse tussen aksent-spesifieke en aksent-onafhanklike modellering afhanklik is van die spesifieke aksente wat ondersoek word. Multi-aksent akoestiese modellering stel ons egter in staat om hierdie besluit outomaties op 'n data-gedrewe wyse te neem. Vir die modellering van die vyf aksente van SAE lewer multi-aksent modelle 'n statisties beduidende verbetering van 1.25% absoluut in woordherkenningsakkuraatheid op in vergelyking met aksent-spesifieke en aksent-onafhanklike modelle.

In 'n tweede stel eksperimente word die praktiese scenario ondersoek waar daar aanvaar word dat die aksent van elke toetsspraakdeel onbekend is. Elke spraakdeel word gelyktydig gebied aan 'n stel herkenners, een vir elke aksent, wat in parallel hardloop. In hierdie opstelling word aksentidentifikasie implisiet uitgevoer. Ons vind dat 'n stelsel wat multi-aksent akoestiese modelle in parallel inspan, effense verbeterde werkverrigting toon in vergelyking met die opstelling waar die aksent bekend is. Dit dui daarop dat aksentidentifiseringsfoute wat gemaak word gedurende herkenning, nie werkverrigting beïnvloed nie. Verder wys ons dat die parallelle benadering ook beter werkverrigting toon as 'n aksent-onafhanklike stelsel wat verkry word deur akoestiese en taalmodelleringsafrigdata saam te gooi.

In 'n finale stel eksperimente ondersoek ons die ongekontroleerde herklassifikasie van aksenttoekennings van die spraakdele in ons afrigstel. Aksente word gemerk deur menslike transkribeerders op grond van 'n spreker se moedertaal en ras. Hierdie toekennings is nie noodwendig optimaal vir modelleringsdoeleindes nie. Deur die aksent van elke spraakdeel in die afrigstel te klassifiseer deur van aanvanklike akoestiese modelle gebruik te maak en dan weer modelle af te rig, word hergeklassifiseerde akoestiese modelle verkry. Ons wys dat die voorgestelde herklassifiseringsalgoritme nie tot enige verbeterings lei nie en dat dit die beste is om modelle op die oorspronklike data af te rig.

Acknowledgements

This work would not have been possible without the continued help and support of the following people. My sincere thanks to:

- My supervisor, Prof. Thomas Niesler, for being the best supervisor I could ask for. I do not believe that there is a supervisor more committed to aiding and encouraging his students.
- Paps and Mams, for your unfailing love and wisdom – I would be lost without you.
- Franna, for being there for me, even in difficult times.
- To Miens, for making my life interesting and always keeping me on my toes; and to Femke, for all the encouragement and loving post-its.
- Alison Wileman, for all the long hours of transcribing, for showing so much interest in my work, and for answering so many of my questions.
- The technical staff who worked behind the scenes: Heine de Jager, for being a great help with the High Performance Computer; and Tannie Thea, for all the proofreading.
- All my great friends for your continued support, help and love and for providing distraction whenever you deemed it necessary. Tuesday ten o'clock tea will always remain a highlight.
- Helena, for your love and patience.
- The National Research Foundation (NRF), for financial assistance in 2011.
- "Lord, when I am afraid, I trust in you. For you have delivered me from death and my feet from stumbling. In God, whose word I praise, in God I trust; I will not be afraid." *Psalm 56*

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 author and are not necessarily to be attributed to the NRF.

Contents

Nomenclature

Variables and Functions

$p(x)$	Probability density function with respect to variable x.
$P(A)$	Probability of event A occurring.
ε	The Bayes error.
ε_u	The Bhattacharyya bound.
B	The Bhattacharyya distance.
s	An HMM state. A subscript is used to refer to a particular state, e.g. s_i refers to the i^{th} state of an HMM.
$\mathbf{S}$	A set of HMM states.
$\mathbf{F}$	A set of frames.
$\mathbf{o}_f$	Observation (feature) vector associated with frame f.
$\gamma_s(\mathbf{o}_f)$	A posteriori probability of the observation vector $\mathbf{o}_f$ being generated by HMM state s.
μ	Statistical mean vector.
Σ	Statistical covariance matrix.
$L(\mathbf{S})$	Log likelihood of the set of HMM states $\mathbf{S}$ generating the training set observation vectors assigned to the states in that set.
$\mathcal{N}(\mathbf{x}\|\mu, \Sigma)$	Multivariate Gaussian PDF with mean μ and covariance matrix Σ.
a_{ij}	The probability of a transition from HMM state s_i to state s_j.
N	Total number of frames or number of tokens, depending on the context.
D	Number of deletion errors.
I	Number of insertion errors.
S	Number of substitution errors.

Acronyms and Abbreviations

AE	Afrikaans English
AID	Accent Identification
ASR	Automatic Speech Recognition
AST	African Speech Technology
BE	Black South African English
CE	Cape Flats English
DCD	Dialect-Context-Dependent
EE	White South African English
G2P	Grapheme to Phoneme
GMM	Gaussian Mixture Model
GPS	Global Phone Set
HMM	Hidden Markov Model
HTK	Hidden Markov Model Toolkit
IE	Indian South African English
IPA	International Phonetic Alphabet
LM	Language Model
LMS	Language Model Scaling Factor
LVCSR	Large Vocabulary Continuous Speech Recognition
MAP	Maximum a Posteriori
MFCC	Mel-Frequency Cepstral Coefficient
MLLR	Maximum Likelihood Linear Regression
MR	Multiroot
OOV	Out-of-Vocabulary
OR	One-Root
PD	Pronunciation Dictionary
PDF	Probability Density Function
SAE	South African English
SAMPA	Speech Assessment Methods Phonetic Alphabet
SRILM	Stanford Research Institute Language Modelling
WER	Word Error Rate
WIP	Word Insertion Penalty

Chapter 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 accented speech. In countries with non-homogeneous populations, ASR systems are therefore greatly challenged since accented 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 illustrated in Figure 1.1. 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 within the same population. The development of speech technology in South Africa is key due to high levels of illiteracy and the accompanying lack of basic computer skills. In these circumstances it is especially important to develop ASR systems that are able to process multiple accents of South African English (SAE) in order to ensure that, for example, telephone-based information retrieval services and other speech-based automated services become accessible to the wider population. SAE therefore provides a challenging and relevant scenario for the modelling of accents in ASR. It is also classified as an under-resourced variety of English since the annotated speech available for the development of ASR systems is exceedingly limited.

The research presented in this thesis considers the development of ASR systems which explicitly cater for the different accents of SAE. Given a limited corpus of these accents, we evaluate several modelling approaches and investigate different system configurations in order to determine the best design strategies for the development of a multi-accent South African English ASR system.

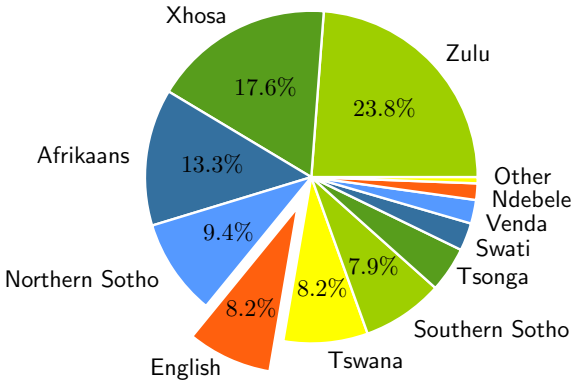


Figure 1.1: Mother-tongue speakers of the eleven official languages in South Africa as a percentage of the population [1]. Although English is widely used, only 8.2% of the population use it as a first language.

1.1 Motivation for Research

Several authors [2, 3, 4] have found that significant deterioration occurs when speech from one accent is presented to a recogniser trained on speech from another. In our own preliminary experiments, for instance, we found that a system trained on White South African English data achieved a word recognition accuracy of 50.42% when presented with a Black South African English test set. In contrast, a similar system trained on Black South African English achieved a word accuracy of 72.84% on the same test set. In the South African environment, systems are likely to be confronted with several accents since accents of SAE are not bound to specific geographic areas.

The work presented in this thesis extends previous research conducted at Stellenbosch University in 2006 and 2007 [5]. In that study, multilingual acoustic modelling of four South African languages (Afrikaans, English, Xhosa and Zulu) was considered. In phone recognition experiments, the author showed that by selectively sharing data across languages, modest improvements can be achieved over language-specific acoustic modelling (in which separate model sets are trained for each language) and language-independent acoustic modelling (performed by pooling training data across languages). A major motivation for the work presented in this thesis is to determine whether similar acoustic modelling strategies can be applied to the different accents of SAE. In particular, we investigate whether selective sharing leads to improvements over accent-specific and accent-independent acoustic modelling. Additionally, we extend the scenario dealt with in [5] and consider the practical implications of building a multi-accent speech recognition system able to process multiple accents of SAE. Our final system can be viewed as a single SAE recognition system for which the accent of the input speech need not be known.

Five accents of SAE are identified in the literature [6]: Afrikaans English, Black South African English, Cape Flats English, White South African English and Indian South African English. Explicit modelling and speech recognition of these five accents have not been considered in any other extensive study. This in itself serves as motivation for the research presented here. Our experiments are based on the African Speech Technology (AST) databases which include five English databases corresponding to the five SAE accents. Although these databases are small compared to those used in state-of-the-art systems, they are representative of an under-resourced environment. This serves as further motivation for our research; we investigate the behaviour of different modelling techniques in a representative under-resourced setting in which the presence of multiple accents further hampers the development of ASR technology. Because speech corpora are scarce and expensive to collect under such circumstances, we aim to determine which modelling approaches and strategies take best advantage of a limited corpus of accented speech.

1.2 Project Scope and Contributions

The evaluation of three acoustic modelling approaches (previously developed in [5]) forms the basis of this study: (i) accent-specific acoustic modelling in which separate model sets are trained for each accent, (ii) accent-independent acoustic modelling performed by straightforward pooling of data across accents, and (iii) multi-accent acoustic modelling. For the latter approach we extend the decision-tree clustering process normally used to construct tied-state hidden Markov models (HMMs) by allowing accent-based questions. This enables selective data sharing across accents. The question of whether data should be pooled or separated is not straightforward when one deals with multiple accents. A major aim of this study is therefore to determine which modelling approach takes best advantage of the available data.

Several authors have considered the traditional accent-specific and accent-independent acoustic modelling approaches (Section 2.3.1). The multi-accent acoustic modelling approach is similar to strategies evaluated in research discussed in Sections 2.4 and 2.5. However, to our knowledge this is the first time that multi-accent acoustic modelling is applied to multiple accents in a tied-state HMM system topology. Furthermore, this study is the first thorough evaluation and comparison of the three modelling approaches when applied to accents of SAE.

We first investigate the phone and word recognition performance of the three modelling approaches when employed in a system configuration where the accent of each test utterance is assumed to be known and each utterance is presented only to the matching model set. We refer to this configuration as 'oracle recognition'. By performing recognition in this way, acoustic modelling considerations are isolated from the effects caused by possible accent misclassifications. Oracle evaluation of the three acoustic modelling approaches when applied to only Afrikaans English and White South African English has been presented as a paper at *PRASA 2010* [7]. A more extensive evaluation of all five accents of SAE has been submitted as an article to the journal *Speech Communication*.

We also consider the practical scenario where the accent of test utterances is assumed to be unknown. 'Parallel recognition' refers to the configuration where an utterance is presented to a bank of recognisers running in parallel and the output with the highest associated likelihood is selected. By comparing oracle and parallel system performance, we analyse the effects of accent misclassifications in the accent identification process that occurs implicitly when parallel recognition is performed. Parts of this evaluation have been published in a paper presented in Florence, Italy at *Interspeech 2011* [8]. In that publication, however, only Afrikaans English, Black South African English and White South African English were considered. In Chapter 7 a more extensive evaluation of all five SAE accents is included, which we would like to publish in the future. Many authors fail to distinguish between the different recognition scenarios when they deal with multiple accents and a thorough analysis of the effects of misclassifications could not be found in the literature.

In our final experimental investigation we consider the unsupervised reclassification of training set accent labels. Accent labels are assigned by human annotators based on a speaker's mother-tongue or ethnicity, but these may be inappropriate for modelling purposes. By classifying the accent of each utterance in the training set by using first-pass acoustic models trained on the original databases and then retraining the models, reclassified acoustic models are obtained. Accent-specific and multi-accent acoustic modelling are employed in the reclassification process. We compare the performance of reclassified models to models trained on the original data as well as to accent-independent models. The evaluation and analysis of the proposed reclassification procedure have been presented as a paper at *PRASA 2011* [9].

1.3 Thesis Overview

The structure of this thesis is as follows. First, an overview of relevant literature is given in Chapter 2. Two main streams of research are encountered regarding accented speech recognition: research focussing on pronunciation modelling and research focussing on acoustic modelling. Noteworthy studies from both philosophies are discussed as well as relevant literature dealing with multilingual speech recognition. The chapter is concluded with a brief summary which indicates how the described literature relates to our research. Chapter 3 gives an overview of the AST databases on which our research is based and explains how these databases were divided into training and test sets. A brief description is also given of the SAE accents and how the accents are represented in the AST databases. The chapter is concluded with an analysis in which the similarity of the different SAE accents are estimated. In Chapter 4 we give a theoretical

overview of decision-tree state clustering, upon which the three acoustic modelling approaches considered in this research are based. This is followed by a detailed discussion of each of the three acoustic modelling approaches. In Chapter 5 the common setup of the experiments carried out as part of this research is described. An overview is given of language and pronunciation modelling and the system optimisation procedure. Chapters 6, 7 and 8 describe experiments in which the three acoustic modelling approaches were evaluated (Chapter 6), misclassifications occurring during parallel recognition were analysed (Chapter 7), and unsupervised accent reclassification of training set utterances was considered (Chapter 8). The thesis is concluded in Chapter 9 with a summary, conclusions and recommendations for future work.

Chapter 2

Literature Review

According to Van Compernolle [10], two main streams of research are encountered when literature dealing with speech recognition of multiple accents or dialects is considered. Some authors consider modelling accents as pronunciation variants which should be taken into account in the pronunciation dictionary employed by a speech recogniser. Other authors focus on tailoring acoustic models to accented speech. Research considering these two modelling philosophies is also commonly found in literature dealing with the recognition of non-native speech [11].

Research following both philosophies is described in this chapter. First, an overview of terminology and concepts is given. The modelling of accented pronunciation variants is discussed next, followed by a description of several acoustic modelling approaches used for accented speech. An overview of relevant research dealing with multilingual speech recognition is also included. Next, a discussion of the most relevant recent research in multidialectal acoustic modelling is presented. The chapter is concluded with a brief overview of literature dealing with different recognition configurations used for simultaneously recognising multiple accents.

2.1 Background and Terminology

2.1.1 Accents, Dialects, Varieties and Non-Native Speech

The terms 'accent', 'dialect', 'variety' and 'non-native speech' are sometimes used interchangeably, inconsistently or imprecisely in the literature, leading to some confusion. It is therefore useful to attempt to draw a distinction between these terms and give broad definitions. The following definitions are based on distinctions made by Crystal [12], Wells [13], and Uebler and Boros [14]:

Accent
: "The cumulative auditory effect of those features of pronunciation which identify where a person is from, regionally or socially" [12]. For example, speakers from London and speakers from Lancashire use different accents of British English [15]. Accents differ in terms of pronunciation but not in vocabulary or grammar.

Dialect
: "A regionally or socially distinctive variety of language, identified by a particular set of words and grammatical structures ... usually also associated with a distinctive pronunciation, or accent" [12]. For example, American English and British English are dialects of English, differing in pronunciation as well as vocabulary and grammar.

Variety
"A language variety may be determined by syntax, morphology, vocabulary and pronunciation, in any combination" [13]. For example, White South African English, Afrikaans English and Indian South African English are different varieties of South African English.

Non-native speech
Speech produced by a speaker using a language different from his or her first language and considered a "foreign language" in their education system. In [14], for example, German speech data from Italian mother-tongue speakers are considered non-native speech.

From these definitions we see that both accents and dialects can be considered varieties. Crystal continues by emphasising that 'accent' refers to pronunciation differences only, while 'dialect' refers to differences in grammar and vocabulary as well. However, the distinction between an accent and a dialect can become very vague in some cases [16]. Wells restricts himself to the use of the more neutral term 'variety' when it is necessary to avoid confusion. The term 'non-native speech' can also become cumbersome in some situtations, but is usually considered separately from native accented or dialectal speech [10, 17, 18]. Non-native speech is characterised by a much larger variability in proficiency levels between speakers than native speech, leading to additional challenges for speech recognition [10]. We will aim to keep to the definitions given above in the following review of relevant literature.

2.1.2 Accented Speech Recognition Scenarios

When reviewing literature dealing with speech recognition of accented speech, it is important to identify the recognition scenario that is being considered and how evaluation is performed. For example, the focus of model adaptation (Section 2.3.2) is most often the development of a single recognition system for a particular target accent. Evaluation would then involve recognising a test set from that target accent. In contrast, other research considers simultaneous recognition of several accents (see for example [4], described in Section 2.3.1). For this scenario, the test set consists of several accents and the accent of a test utterance is assumed to be unknown. Recognition can be performed by, for example, running several speech recognition systems tailored to different accents in parallel, which we refer to as 'parallel recognition'.

As a precursor to this last scenario, 'oracle recognition' can be performed in which test utterances are presented only to the recogniser tailored to the accent of that utterance. The accent of each test utterance is therefore assumed to be known during evaluation and this setup can therefore be considered a cheating experiment, depending on the application. By configuring the recognition setup in this way, modelling effects can be isolated since the effects of accent misclassifications are avoided. Parallel recognition is typically assumed to show deteriorated performance compared to oracle recognition as a result of these misclassifications.

In this chapter, literature focussing on several aspects of multi-accent speech recognition is described. For each case it is important to identify the recognition scenario that is being addressed and we attempt to indicate this in the text. In Section 2.6, an overview is given of literature which explicitly considers and compares some of these recognition scenarios.

2.2 Pronunciation Modelling of Accented Speech

According to the definition given in Section 2.1.1, the term 'accent' refers to features of *pronunciation* which indicate where a person is from, regionally or socially. Dialect is normally also associated with a distinctive pronunciation. As described in [15], techniques focussing solely on

acoustic modelling for multiple accents often make the simplifying assumption that the pronunciation defined in the pronunciation dictionary is invariant to accent and that accent variation is absorbed by the acoustic models. By reflecting on the definition of 'accent', some authors argue that this is a poor assumption and therefore focus on methods of explicitly modelling accents as pronunciation variants that should be taken into account in the pronunciation dictionary of a recogniser.

2.2.1 Research by Humphries et al.

Noteable research on the explicit modelling of pronunciation variants for accented speech has been conducted by Humphries et al. [15, 19, 20]. The focus of their research was the automatic generation of an accent-specific pronunciation dictionary for a target accent from a reference dictionary in another (seed) accent.[1] Their approach involves first obtaining phone-level transcriptions of some data from the target accent using a speech recogniser built for the seed accent. These transcriptions are then aligned (using a dynamic programming technique) to phone-level transcriptions derived from known word-level transcriptions using the reference pronunciation dictionary. From these alignments a list of context-dependent phone substitutions, insertions and deletions are generated which are representative of how the pronunciation of target accent speakers differ from the pronunciation assumed for seed accent speakers. The phone replacement rules are clustered in a binary decision-tree which considers phonetic context and the frequency of occurrence. The resulting trees can then be used to obtain phone replacement rules for arbitrary (and possibly unseen) contexts. Subsequently, these rules are used to generate an accent-specific pronunciation dictionary from a reference dictionary.

In [15] two accent regions in England were considered: London and South East (seed accent) with a training set of 8540 utterances from 100 speakers, and Lancashire and Yorkshire (target accent) with a training set of 7845 utterances from 93 speakers. The Lancashire and Yorkshire test set consisted of 1479 utterances from 17 speakers. All speakers were male and the database used was part of a telephone-based corpus of isolated words from a 2000 word vocabulary. A leaf probability threshold was used to control the number of pronunciations per word in the resulting dictionary. When a sufficient number of pronunciation variants were used, speech recognition word error rate (WER) decreased by almost 20% relative compared to that obtained when the reference dictionary was used. In other words, using London and South East acoustic models with a pronunciation dictionary adapted to the Lancashire and Yorkshire accent showed superior performance compared to a setup in which the London and South East reference dictionary was used. However, too many pronunciation variants were shown to introduce confusability which increased the WER. Only vowel substitutions were considered.

This work was continued in [19, 20] in which the synthesis of an American English pronunciation dictionary from a British English dictionary for large vocabulary continuous speech recognition (LVCSR) was attempted. Two scenarios were considered in [20]. In the first, a large corpus of American English speech data (7185 utterances) with known word-level transcriptions was assumed to be available for acoustic modelling purposes, but no pronunciation dictionary was available. British English acoustic models (trained on 7861 utterances from 92 speakers) and a British pronunciation dictionary were, however, assumed to be available. In the second scenario, a small amount of American English speech data (500 utterances from 10 speakers) was assumed to be available together with British acoustic models and a British dictionary. An American pronunciation dictionary was therefore assumed to be absent in both cases. For both scenarios a test set of 425 American utterances was used for evaluation. A single language model was

[1] In research dealing with modelling and recognition of multiple accents, it is often assumed that extensive resources are available for one accent, the seed accent, and that these can be used to tailor or adapt models to the target accent.

used and acoustic models consisted of cross-word triphone hidden Markov models (HMMs). In contrast to [15], where only vowel substitutions were considered, the dictionary synthesis procedure followed in these experiments catered for phone substitutions, deletions and insertions.

The first scenario was addressed by synthesising an American English pronunciation dictionary. This new dictionary was then used to derive phonetic transcriptions of the American English data in order to train American acoustic models. Subsequently, the American models and synthesised dictionary were used during recognition. This approach did not yield a statistically significant improvement over models simply trained on the American speech using phonetic transcriptions from the British dictionary (a WER of 14.9% compared to 14.1%). The second scenario was addressed by synthesising an American English dictionary and then performing speaker-independent MLLR adaptation (see Section 2.3.2) of the British models using the American accented speech, which was phonetically segmented using the synthesised dictionary. Using British acoustic models and a British dictionary resulted in a WER of 30.9%; using British acoustic models with the American synthesised dictionary gave 25.8% WER; using MLLR adapted acoustic models with a British dictionary yielded 24.7% WER; and the best WER of 21.2% was obtained using MLLR adapted acoustic models with the synthesised American dictionary.

From these results the following conclusions were drawn [20]:

> The results ... have shown how pronunciation modelling can be of significant benefit when an accent-specific recogniser is desired, but the amount of acoustic training data is limited, and phonological information non-existent. However, [it was] demonstrated that when a large amount of acoustic data is available, extra phonological information is of less value. This may perhaps be because when given enough acoustic data, pronunciation variability may be absorbed to a larger degree within the acoustic models.

2.2.2 Other Authors

Beringer et al. [21] considered pronunciation modelling for 12 dialects of German represented in the German VERBMOBIL database. This database consists of a total of 12 000 turns in the training set and 1800 turns in the test set with one turn having 22.8 words on average in a task reflecting official situations such as asking for flight departures or arranging a business meeting. A set of 12 dialect-specific pronunciation dictionaries were used (based on dialect-specific pronunciation transcriptions from the Munich VERBMOBIL group) with a single set of acoustic models. The 104 speakers in the test set were clustered into groups corresponding to the 12 dialect regions. Recognition of the test set involved using only the pronunciation dictionary corresponding to the group being recognised and, although several different approaches were considered, no significant improvements could be obtained compared to a baseline system using a German dictionary with canonical pronunciations.

In [22] Fung and Liu related their work to that of Humphries and Woodland [19]. They considered the construction of an accent-specific pronunciation dictionary for Cantonese accented English. However, instead of using a data-driven approach to determine phone replacement rules, phonological rules based on linguistic knowledge of Cantonese were used. By applying 28 phonetic rules to the BEEP dictionary designed for native English speakers, a Cantonese accented pronunciation dictionary was obtained. The use of this adapted dictionary gave a relative improvement of 13.5% in WER (from 30.8% to 26.6%) in the recognition of 4 Cantonese English speakers. This can be compared to the relative reduction of 19.7% in WER (from 30.9% to 24.8%) reported by Humphries and Woodland in [19]. The TIMIT corpus (6300 utterances from 630 American speakers) was used together with a corpus consisting of 800 utterances from

Table 2.1: WER for recognition of Japanese accented English following various accent-specific pronunciation modelling strategies as considered by Tomokiyo [23].

Modelling approach	WER (%)
Baseline	55.7
Phone mapping rules	52.8
Automatically generated	50.8
Hand-coded	50.6

both native English and Cantonese English speakers. The separation of training and test sets are not clearly indicated in [22].

Tomokiyo [23] compared different approaches in the creation of pronunciation dictionaries for Japanese accented English. A single set of acoustic models was used for all experiments. The strategies included using an automatically generated pronunciation dictionary (following an approach comparable to that of Humphries and Woodland [20]), using linguistic phoneme mapping rules (comparable to the work of Fung and Liu [22]) and using a hand-coded pronunciation dictionary containing frequently accented pronunciation variants for 100 common words. Results are presented in Table 2.1. It is evident that the automatically generated dictionary gives very similar results to the hand-coded one.

Similar techniques for explicitly modelling accented and non-native word pronunciation variants have also been considered by Huang et al. [24], for Mandarin speakers from Beijing (seed dialect) and Shanghai (target dialect); and Livescu et al. [25, 26], for non-native English variants.

2.2.3 Conclusions Regarding Pronunciation Modelling

Although various authors have obtained encouraging results by explicitly catering for accented pronunciations, the results obtained by Humphries and Woodland in [20] seem to suggest that accent-specific pronunciation modelling is less effective when similar amounts of training data from seed and target accents are available. In one scenario (described in Section 2.2.1) comparable amounts of American English data (7185 utterances) and British English data (7861 utterances) were available, but using a synthesised American English pronunciation dictionary with American English acoustic models did not provide significant improvements in recognition accuracy compared to a baseline system employing American English acoustic models with a British English dictionary.

Humphries and Woodland [20] suggest that pronunciation variants may be absorbed in the acoustic models when larger amounts of accent-specific data are available. Van Compernolle [10] states that "most authors [come] to the same conclusion that only the most pronounced variants are essential". He argues that pronunciation variants are mostly seen in short span phonemic strings rather than at word level. This would imply that pronunciation variants are learned implicitly when accent-specific context-dependent acoustic models are trained, assuming that sufficient acoustic training data are available.

2.3 Acoustic Modelling of Accented Speech

2.3.1 Traditional Approaches: Accent-Specific and Accent-Independent Acoustic Modelling

Perhaps the simplest approach to acoustic modelling when dealing with multiple accents is to train a single accent-independent acoustic model set by pooling accent-specific data across all accents considered. This results in a single acoustic model set for all the accents. An alternative to this would be to train a fully accent-specific system that allows no sharing between accents. This would result in a completely different set of acoustic models for each accent. A number of authors considered these two approaches for a variety of accents, dialects and non-native variants and a brief overview is given below. A summary of the literature described in this section is given in Table 2.2.

Van Compernolle et al. [2] compared the two approaches in a isolated digit recogniser for Dutch and Flemish, which are identically written languages but have large pronunciation differences. Telephone recordings were obtained from around 1000 speakers of which half were Dutch and half were Flemish. Each speaker spoke each digit in isolation. The available data consisted of 4804 Flemish and 3993 Dutch utterances for training and 735 Flemish and 985 Dutch utterances for testing purposes. Whole-word left-to-right HMMs were trained. Results indicated that using separate dialect-specific models for each dialect was superior to using models obtained by pooling the dialect-specific data. WERs improved from 8.71% to 4.49% for Flemish and from 7.72% to 5.28% for Dutch in oracle recognition experiments. When running Dutch and Flemish dialect-specific models in parallel, WERs of 4.90% and 5.99% were achieved on the Dutch and Flemish test sets, respectively. It was also shown that a division based on dialect was superior to a division based on gender when training multiple model sets.

In a continuous speech command-and-control application (approximately 200 words), Beattie et al. [27] compared different model sets for the recognition of three regional dialects of American English (Midwest, Southern and mainstream). A first experiment used a model set obtained by pooling the dialect-specific data. A second experiment used separate model sets for each gender. A third experiment employed separate models trained for each gender in each dialect. HMMs were used to model triphones and the same pronunciation dictionary was used for all the experiments. The results indicated that the gender- and dialect-independent models (trained on the pooled data) performed worst with 7.03% average WER, that gender-dependent models performed better (4.9%) and that gender- and dialect-specific models gave the best results (4.23%). A variation of parallel recognition was used for the evaluation. Training set size is not indicated in [27]. The test set consisted of about 210 sentences from 5 Southern speakers, 280 sentences from 4 Midwest speakers and 280 sentences from 4 mainstream speakers.

Teixeira et al. [3] considered isolated word recognition of non-native English spoken by native speakers of Danish, German, British English, Spanish, Italian and Portuguese. The corpus used in this research was collected from ten male speakers from each non-native accent, each speaker uttering a vocabulary of 200 isolated words twice. The corpus was split into a training set (60%) and test set (40%) with no overlap in speakers. Three strategies of training were considered: models were trained on British English only, models were trained on the pooled data and models were trained separately for each non-native accent. Both whole-word and sub-word (phone) models were evaluated. For the whole-word and phone models the models trained on British English only performed worst for all the non-native accents. In the evaluation of the whole-word models (i.e. word recognition), the models trained on the pooled data performed better than the separate accent-specific models. This was attributed to the increase in the amount of data used for training. In phone recognition experiments the separate models performed better than

Table 2.2: Literature comparing accent-specific and accent-independent acoustic modelling.

Authors	Accents	Task	Training corpus	Best approach
Van Compernolle et al. [2]	Dutch and Flemish.	Isolated digit recognition.	3993 Dutch and 4804 Flemish utterances.	Accent-specific modelling.
Beattie et al. [27]	Three dialects of American English.	Command and control (200 words).	Not indicated.	Gender- and dialect-specific modelling.
Teixeira et al. [3]	English by Danish, German, British English, Spanish, Italian and Portuguese speakers.	Isolated word recognition.	A vocabulary of 200 words uttered twice by ten male speakers from each accent.	Word recognition: accent-independent. Phone recognition: accent-specific.
Fischer et al. [28]	German and Austrian dialects.	LVCSR	90h German, 15h Austrian speech.	Accent-specific modelling.
Chengalvarayan [4]	American, Australian and British dialects of English.	Connected digit recognition.	7461 American, 5298 Australian and 2561 British digit strings.	Accent-independent modelling.

the models obtained by pooling the data. A single pronunciation dictionary was used for all experiments. For the evaluation of the models trained separately for each non-native accent, oracle recognition was performed.

In [28] Fischer et al. considered the improvement of a German LVCSR system for Austrian speakers. For the baseline German system, 90 hours of speech read by 700 native German speakers was used. The Austrian corpus used for training consisted of 15 hours of speech read by 100 Austrian speakers. Speech from 20 German and 20 Austrian speakers were used as test data. Context-dependent HMMs were trained. Models obtained by pooling data as well as accent-specific models were evaluated. Results indicated that the accuracy of Austrian speech recognised using the Austrian-trained models (12.24% WER) was comparable to the accuracy of German speech recognised using the German-trained models (13.16% WER). These accuracies were superior to the accuracy from the recogniser employing the accent-independent models (WERs of 13.72% and 15.61% for the German and Austrian test sets, respectively).

In [4] Chengalvarayan presented experimental results for connected digit recognition of American, Australian and British dialects of English. The corpora used contained read and spontaneous digit strings ranging from 1 to 16 digits. For training, 7461 strings from 5234 American speakers, 5298 strings from 700 Australian speakers and 2561 strings from 700 British speakers were used. The test set consisted of 2023, 848 and 505 valid digit strings from American, Australian and British speakers, respectively. A system developed using a single model set (context-dependent HMMs) obtained by pooling data across dialects outperformed a system employing several dialect-specific models in parallel.

Authors who followed similar approaches include Brousseau and Fox [29] who dealt with Canadian and European dialects of French, Kudo et al. [30] who considered two accents of Japanese, and Uebler and Boros [14] who considered recognition of non-native German.

From the findings of the literature surveyed above and summarised in Table 2.2 it seems that in most cases accent-specific modelling leads to superior speech recognition performance compared to accent-independent modelling. However, this is not always the case, as illustrated in [3] and [4], and the comparative merits of the accent-specific and accent-independent modelling

approaches appear to depend on factors such as abundance of training data, the type of task and the degree of similarity between the accents involved.

2.3.2 Acoustic Model Adaptation

A common technique employed when developing speaker-dependent speech recognition systems is to perform speaker adaptation. An initial set of models is trained using a large amount of data from various speakers. Transformation of the models is then performed based on a small amount of adaptation data from a single speaker. In this way a more accurate model of that particular speaker's speech is obtained without the detrimental effects caused by insufficient data. Adaptation can be either supervised, where the transcription of the adaptation data is known beforehand, or unsupervised, where the recogniser's output is used as the transcription for the adaptation data.

Common adaptation techniques include maximum a posteriori (MAP) adaptation [31, 32] and maximum likelihood linear regression (MLLR) adaptation [33]. In MLLR adaptation a single transformation function is estimated from the adaptation data and then applied to all the initial models. For MAP adaptation the model parameters are re-estimated individually by shifting the originally estimated mean values towards sample mean values calculated on the adaptation data. MLLR performs adaptation on all models while MAP does not perform adaptation for models where adaptation data are insufficient.

In cases where accent-specific data are insufficient to train accent-specific models, similar strategies as those for speaker adaptation can be followed. Initial models can be trained on a large amount of available data from one accent (the seed accent) and can then be adapted to another accent (the target accent) in two ways. In the one case speaker-dependent models can be obtained by adapting the initial models using limited amounts of enrolment data from a specific speaker with an accent. In the other case the models can be adapted using a small amount of accented data from various speakers, which will result in a set of speaker-independent models.

Beattie et al. [27] obtained speaker-dependent models using supervised adaptation for non-native speakers of American English and reduced the WER by more than a factor of 2 compared to native American English models. Zavaliagkos et al. [34] similarly obtained reduction in WER by factors of 2 to 4 for non-native speakers of American English by using both supervised and unsupervised adaptation to build speaker-dependent models from native models. Increased accuracies with speaker-dependent models were also obtained by Humphries and Woodland [19] who adapted British English models to American speakers using supervised MLLR, Tomokiyo [23] who obtained Japanese accented English models using supervised MLLR, and Huang et al. [24] who considered MLLR adaptation using Mandarin speakers from Beijing as the seed dialect and speakers from Shanghai as the target dialect.

More recently, Kirchhoff and Vergyri [35, 36] applied MAP and MLLR in an approach where acoustic training data were shared between Modern Standard Arabic (seed dialect) and Egyptian Conversational Arabic (target dialect) to obtain speaker-independent models. Speaker-independent adaptation has also been considered by Wang et al. [37] and Zhang et al. [38, 39]. The results obtained by these authors are described in more detail in Section 2.3.5, where different acoustic modelling and data sharing techniques are compared.

Diakoloukas et al. [40] investigated the development of a Swedish multidialectal speaker-independent speech recognition system for two dialects of Swedish in an air travel information task. Stockholm and Scanian dialects of Swedish were considered, the former being the seed dialect and the latter being the target dialect. An initial set of models was trained on approximately 21 000 sentences of Stockholm speech. These models achieved a WER of 8.9% when

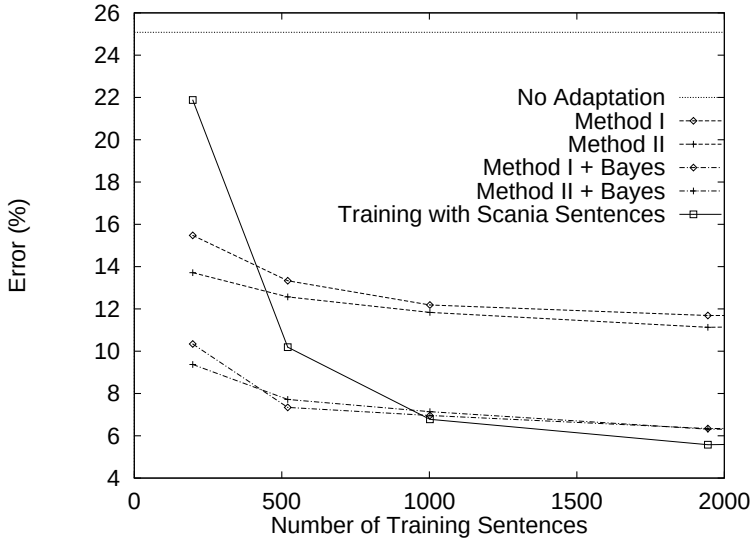


Figure 2.1: WER for recognition of Scanian speakers of Swedish. Stockholm models adapted to Scanian adaptation data are compared to models simply trained on Scanian data [40].

tested on Stockholm speakers but achieved only 25.1% WER when tested on the Scanian test set (8 speakers, each recording more than 40 sentences). These models were then adapted using various amounts of data from 32 Scanian speakers (there was no overlap with the speakers in the test set). Different adaptation techniques were considered. Scanian models were also trained from scratch using the same adaptation data to determine the benefit gained from adaptation. The results are presented in Figure 2.1, where 'Method I', 'Method II' and 'Bayes' refer to different adaptation approaches. It is clear that when larger amounts of training data were used (more than 1000 sentences) the models trained on only the Scanian dialect data outperformed the adapted models. In contrast to these results, Despres et al. [41] found that accent-independent models which had been MAP adapted with accented data outperformed both accent-specific and accent-independent models for Northern and Southern dialects of Dutch.

From the research presented in this section it is evident that acoustic model adaptation can be applied successfully in cases where limited amounts of accent-specific data are available for a target accent while a large amount of data is available for a seed accent. However, when larger amounts of target accent data are available, it is not clear whether adaptation or training from scratch is superior and differing results are reported for different adaptation configurations and accents, as illustrated in [40] and [41].

2.3.3 Multilingual Models for Non-Native Speech Recognition

Due to the increase in the use of speech recognition technology all over the world, multilingual speech recognition (described in more detail in Section 2.4) has been receiving increased attention over the last decade. Fischer et al. [42] argues that data collection tends to focus on the collection of native speech in many languages rather than the collection of non-native speech. It is also argued (e.g. [38, 39]) that speech from non-native speakers is greatly influenced by the speech sounds present in their native language. This naturally leads to the scenario where non-native speech recognition must be attempted where only native training data are available. Literature focussing on this scenario (modelling of non-native accented speech using native data) is discussed in this section, while relevant literature dealing explicitly with multilingual acoustic modelling (modelling of different languages) is discussed in Section 2.4.

Fischer et al. [42, 43, 44] considered the recognition of English digit strings spoken by speakers

from Spain, France, Germany and Italy. Data from 5 languages (Spanish, French, German, Italian and British English) were available with approximately 17 hours of data per language. In [42] a monolingual British English model set and 4 bilingual acoustic model sets (built by sharing data between British English and each of the 4 remaining languages) were created. Data were pooled for phones with the same SAMPA classification and bilingual word-internal triphone HMMs were obtained by growing binary decision-trees asking questions regarding phonetic context.[2] The monolingual model set was compared to the bilingual model sets for the recognition of 50 English digit sequences spoken by 10 speakers from each of the countries considered. The corresponding bilingual models performed better for all 4 groups of non-native speakers with an improvement over the monolingual models of up to 8% relative in WER. A similar procedure was followed in [44], except that data were pooled across all 5 languages resulting in multilingual word-internal triphone HMMs. Similar improvements over the native monolingual models were achieved.

Wang et al. [37] compared the use of bilingual acoustic models to adaptation and other data sharing techniques and an overview of this research is presented in Section 2.3.5. Zhang et al. [38, 39] used bilingual models in tandem with adaptation techniques for the recognition of non-native English from Mandarin speakers and this research is also described in more detail in Section 2.3.5. It is clear that the use of multilingual models for the recognition of non-native speech is mostly relevant in cases where non-native (accent-specific) training data are unavailable, which is also the conclusion drawn in [37].

2.3.4 Model Interpolation

In Section 2.3.1 research was described in which either accent-specific or accent-independent model sets were obtained from accented data. In Section 2.3.2 several acoustic model adaptation approaches were described. An alternative approach to these is to interpolate accent-specific models. Well-trained models can be obtained from seed accent data and then interpolated with models obtained using target accent data. This takes advantage of the larger amount of seed accent data, while incorporation of information from the target accent ensures a more accurate representation of the target accent. Several authors followed this approach, mostly for the interpolation of native and non-native models.

Livescu [25] defines interpolation as the weighted average of the probability density functions (PDFs) of several models to produce a single PDF. She considered the interpolation of models trained on native American English with models from non-native speakers. The PDF obtained by interpolating a native model and a non-native model can be defined as:

$$p(\mathbf{x}) = w_{\text{native}} \cdot p_{\text{native}}(\mathbf{x}) + w_{\text{non-native}} \cdot p_{\text{non-native}}(\mathbf{x}) \tag{2.1}$$

where $w_{\text{native}} + w_{\text{non-native}} = 1$ holds, w_{native} and $w_{\text{non-native}}$ are interpolation weights and $\mathbf{x}$ is an observed feature vector.

In Livescu's research the recognition of non-native English was considered in a telephone-based conversational information retrieval task. Native American data were divided into training, development and evaluation sets with 33 692, 730 and 1649 utterances, respectively. Non-native training, development and evaluation sets consisted of 2717, 609 and 1013 utterances, respectively. Models obtained by pooling the training data were compared to models trained on the non-native training data alone. The same pronunciation dictionary and language model were used in these experiments. WERs of 20.2% and 23.0% were obtained on the non-native development set using the pooled and non-native models, respectively. The corresponding WERs measured on the non-native evaluation set were 20.9% and 23.4%.

[2] Questions regarding language were not allowed in the decision-trees.

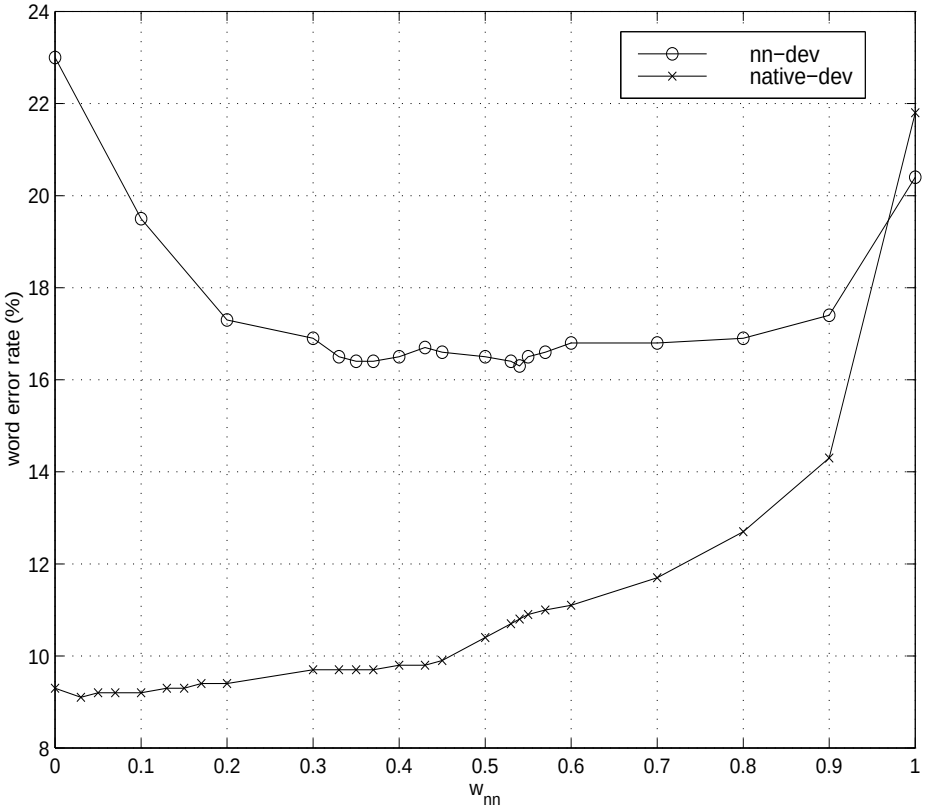


Figure 2.2: WER for recognition of native (native-dev) and non-native (nn-dev) American English development sets using interpolated models with varying interpolation weights [25].

Interpolation of models trained on native data alone with the models trained on non-native data was then considered. A series of interpolated models was created by varying the w_{native} and $w_{\text{non-native}}$ interpolation weights indicated in equation (2.1). The performance achieved on the native and non-native development sets is shown in Figure 2.2. The lowest WER obtained for the non-native development set was 16.3% (at $w_{\text{non-native}} = 0.54$) which is a 20% relative reduction in error rate compared to the non-native models and a 19.3% reduction compared to the WER obtained using the pooled models. Using these interpolation weights, the WER achieved on the non-native evaluation set was improved to 19.2%, which is an 11.5% relative improvement over the non-native models and an 8.1% improvement over the pooled models.

Liu and Fung [45, 46] also employed model interpolation in a larger data sharing scheme for Cantonese and Wu accented Mandarin (target accents) with standard Mandarin (seed accent) and obtained encouraging results. Results obtained by Tomokiyo [23] using model interpolation for the acoustic modelling of Japanese accented English are described in the next section.

2.3.5 Research Comparing Different Acoustic Modelling Approaches

As outlined in [14], one of the main difficulties when considering multilingual, multi-accent or multidialectal speech recognition is to decide when a speech unit in one language or accent should be considered equivalent to a speech unit in another language or accent. Practically speaking, this implies that a decision should be made about how and when data should be shared between acoustic models from different languages or accents.

Several approaches to address this issue have already been described in this chapter, such as obtaining accent-independent models by simply pooling data across accents (Section 2.3.1), having distinct accent-specific model sets (Section 2.3.1), performing adaptation on models from

Table 2.3: WER for recognition of German-accented English following various acoustic modelling strategies as considered by Wang et al. [37].

Acoustic models	Error rate (%)
Native models	49.3
Bilingual models	48.7
Non-native models	43.5
Pooled models	42.7
MAP adapted models	37.5
Interpolated models	36.0

a seed accent using adaptation data from a target accent (Section 2.3.2) and acoustic model interpolation (Section 2.3.4). Literature dealing with the use of these techniques in tandem as well as comparatively is discussed below.

Tomokiyo [23] considered acoustic modelling of Japanese accented English using limited amounts of non-native speech and models trained on native English for LVCSR. The size of the set used for training the native English models is not indicated in [23], but 2.8 hours of Japanese accented English from 15 speakers were used as non-native training data and 1.1 hours of speech from 10 Japanese speakers were used as a test set. Quintphone contextual models were trained and a single language model was used for all experiments. The WER for the test set using the native models was 90%. When supervised MLLR adaptation (50 utterances per speaker) was used, the WER reduced to 67.3%. The non-native training data (which were found to be insufficient to train non-native acoustic models from scratch) were used to perform additional training iterations on the native models. By using these models in tandem with MLLR adaptation, the average WER was reduced to 47.2%. Interpolating the trained non-native models with the native model set and again applying MLLR further reduced the error rate to 45.1%.

In [37] Wang et al. performed a thorough comparison of different acoustic modelling and data sharing techniques for recognition of non-native English from native German speakers. Native English from 2118 speakers (34 hours) was used to train native English models. The non-native training and test sets consisted of 52 minutes (64 speakers) and 36 minutes (40 speakers) of speech, respectively. The recordings were based on spontaneous face-to-face dialogues. Separate native and non-native model sets were trained, pooled models were obtained by combining the native and non-native training sets, speaker-independent MLLR and MAP adaptation using the accented data was performed on the native models, and acoustic model interpolation (very similar to the work of Livescu [25], described in Section 2.3.4) was considered. Recognition using bilingual models trained on native English and native German (pooled for phones with the same IPA classification) was also considered. This approach is similar to that followed by Fischer et al. [42, 43, 44] (see Section 2.3.3). Quintphone HMMs were trained and the same language model and vocabulary were used in all experiments. The results are presented in Table 2.3.

More recently Zhang et al. [38, 39] evaluated different data sharing approaches in research dealing with non-native English spoken by native Mandarin speakers. In [39] three training sets were used consisting of 500 hours of native Mandarin, 232 hours of native English (Wall Street Journal) and 20 hours of Mandarin-accented English. The test set contained 1568 English utterances from 4 female and 6 male native Mandarin speakers. Various acoustic modelling approaches using the native English and the Mandarin-accented English training sets for training cross-word triphone HMMs were considered. Native English models were trained, models were trained on pooled native and non-native speech, and MLLR and MAP adaptation of the native English models were performed using the accented data. Results are presented in Table 2.4. Although models trained on non-native speech alone were not considered in [39], the authors

Table 2.4: Phrase error rate for recognition of Mandarin-accented English following various acoustic modelling strategies as considered by Zhang et al. [39].

Acoustic models	Error rate (%)
Native models	46.9
Pooled models	39.3
MLLR adapted models	35.2
MAP adapted models	34.3
Bilingual model approach	31.6

did consider non-native models in earlier research [38]. Results indicated that such models yield similar performance to models obtained by pooling training data.

Sharing between English and native Mandarin models (trained on the native Mandarin training set) was also considered. This is similar to the work by Fischer et al. [42, 43, 44] described in Section 2.3.3. Similarities between states from MAP adapted English models and native Mandarin models were compared and observation PDFs of similar states were then combined in a manner similar to model interpolation (Section 2.3.4). From the results in Table 2.4 it is evident that this approach (the final line in Table 2.4) leads to a significant improvement compared to the MAP adapted models. As described in [39], the effectiveness of this bilingual modelling approach depends on the proficiency of the non-native speakers. For speakers just beginning to learn a foreign language a tendency might exist to substitute sounds from their mother-tongue for foreign sounds they cannot produce. This might not be the case for more proficient non-native speakers. The test set used for this research specifically contained words and phrases which are hard for Mandarin speakers to pronounce and this may explain the large improvement obtained by including information from the native Mandarin acoustic models.

From the results obtained by Wang et al. [37] (Table 2.3) and Zhang et al. [39] (Table 2.4), very similar conclusions can be drawn for the two different cases of non-native speech recognition. In both cases the native models perform worst. The pooled models yield an improved accuracy and adaptation results in still further increases in accuracy. In the research presented by Wang et al. it seems that the bilingual models perform only slightly better than the native models, while the bilingual modelling approach performs best of all in the results obtained by Zhang et al. However, the approach followed by Zhang et al. is fundamentally different because MAP adapted models (adapted using non-native Mandarin-accented English) were combined with similar native Mandarin acoustic models in a weighted manner. In the approach followed by Wang et al. no non-native accented speech was used – the sharing of only native English and native German data was considered. The conclusion drawn by Wang et al. is that the inclusion of native German speech does not lead to significantly improved performance if non-native speech is not also included. This is not contradictory to the results obtained by Zhang et al. As mentioned earlier, the use of multilingual models is, however, greatly influenced by the proficiency of the non-native speakers considered, which makes it difficult to generalise trends.

2.3.6 Conclusions Regarding Acoustic Modelling

In light of the summary of Section 2.3.1 given in Table 2.2 we conclude that the comparative benefits of accent-specific and accent-independent acoustic modelling appear to depend on factors such as the amount of training data, the type of task and the degree of similarity between accents. Results reported in literature dealing with acoustic model adaptation (Section 2.3.2) indicate that when only a small amount of accented data is available, adaptation can be successfully employed in order to achieve improved recognition performance. However, it remains

unclear whether adaptation or training acoustic models from scratch are more advantageous when larger amounts of accented speech data are available. The use of multilingual models (Section 2.3.3), model interpolation (Section 2.3.4) and studies comparing several modelling approaches (Section 2.3.5) also indicate that the best acoustic modelling approach depends on the task and the accents involved and is furthermore greatly dictated by the character and amount of the acoustic training data available.

2.4 Multilingual Speech Recognition

As mentioned in Section 2.3.3, the implementation of speech recognition technology all over the world has led to research dealing with multilingual speech recognition. 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. Techniques for sharing data across languages are therefore often relevant to research dealing with modelling of multiple accents or dialects.

Multilingual speech recognition has received some attention over the last decade, most notably from Schultz and Waibel [47, 48, 49, 50]. Their research considered LVCSR systems of languages spoken in different countries and forming part of the GlobalPhone speech corpus. Experiments were based on context-dependent HMMs in a tied-mixture system topology. 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. In [48, 50] three approaches to multilingual acoustic modelling were developed by applying different decision-tree clustering methodologies:

1. Separate language-specific acoustic model sets can be obtained for the different languages considered by not allowing any sharing of data between languages. In this case, separate decision-trees are grown for each language, employing only questions relating to phonetic context. This can be seen as an approach that puts language information above phonetic context information. It was referred to as ML-sep in [50].

2. A language-independent model set can be obtained by pooling language-specific data across all languages considered. A single set of decision-trees is grown for all languages; these only employ questions relating to phonetic context as no information regarding language is retained. By applying this approach phonetic information is regarded as most important. It was referred to as ML-mix in [50].

3. Multilingual acoustic models can be obtained in an approach similar to that followed to obtain the language-independent model set, except that language information is preserved in the tree-based clustering process. By retaining language information, questions regarding language can be employed in the decision-tree clustering process. A data-driven decision is thus made whether language information is more or less important than phonetic context information. This approach was referred to as ML-tag in [50].

Five languages were considered: Croatian, Japanese, Korean, Spanish and Turkish. The authors were unable to improve on the performance of separate monolingual systems (ML-sep) using the described sharing approaches (ML-mix and ML-tag). Of the two approaches allowing sharing between languages, ML-tag was found to be superior to ML-mix. The authors did, however, show that these modelling approaches could be effectively applied in the development of recognition systems for an unseen language. The ML-sep and ML-mix approaches respectively coincide with the traditional accent-specific and accent-independent acoustic modelling strategies discussed in Section 2.3.1. Recent research applying similar approaches to acoustic modelling of multiple dialects is described in the next section.

Multilingual speech recognition has particular relevance in the South African context where there are eleven official languages and most of the population speak two or more languages. In [5] Niesler addressed multilingual acoustic modelling of four South African languages: Afrikaans, English, Xhosa and Zulu. Experiments were also based on the AST databases (Section 3.1). Similar techniques to those proposed by Schultz and Waibel were employed in the development of tied-state multilingual triphone HMMs by applying different decision-tree state clustering approaches. Language-specific, language-independent and multilingual acoustic models were considered which, respectively, resemble the ML-sep, ML-mix and ML-tag approaches implemented by Schultz and Waibel, except that these were applied in a tied-state system topology as opposed to a tied-mixture topology. While Schultz and Waibel were unable to improve on the performance of separate monolingual systems, Niesler showed modest average performance improvements (approximately 0.1% absolute) over language-specific and language-independent systems using multilingual HMMs in phone recognition experiments. The improvements were consistent over a range of model sizes considered. A detailed description of tied-state HMMs systems, decision-tree state clustering and the different acoustic modelling approaches based on different strategies of decision-tree clustering is given in Chapter 4.

2.5 Multidialectal Speech Recognition

More recently, Caballero et al. [51, 52] considered five dialects of Spanish spoken in Spain and Latin America (Argentina, Caribbean, Colombia and Mexico) and evaluated different methods of sharing speech data between dialects. Different approaches to multidialectal acoustic modelling based on decision-tree clustering algorithms using tied-mixture systems, as employed by Schultz and Waibel for multiple languages (see the preceding section), were compared. Two different decision-tree structures were considered for the clustering process:

1. The typical structure used when performing decision-tree clustering of context-dependent units is a multiroot (MR) structure where a separate tree (root) is grown for each basephone. Similar context-dependent units which are found in different trees can therefore not be clustered.

2. When using a one-root (OR) tree structure, a single tree is grown for all the context-dependent units. This structure allows parameters to be shared between context-dependent units which would traditionally be found in separate trees. Units with different basephones can therefore be clustered.

Using these two tree structures, four approaches to acoustic modelling of multiple dialects were considered:

1. In the GPS-MR approach, a global phone set (GPS) based on the SAMPA alphabet is defined. No distinction is made between corresponding phones from different dialects, i.e. dialect-specific data are pooled across dialects. A decision-tree clustering algorithm using a multiroot structure and employing only questions relating to context are used to obtain a dialect-independent model set. This approach corresponds to the ML-mix approach used by Schultz and Waibel and discussed in the previous section.

2. The GPS-OR approach is similar to the GPS-MR approach, except that a one-root tree structure is employed in the clustering process. Questions regarding the basephone are also employed. This allows similar context-dependent units with different basephones to be clustered, something which is not possible when using a multiroot tree structure.

3. Dialect-context-dependent (DCD) models are clustered using a multiroot structure in the DCD-MR approach. Questions relating to context and dialect are employed. This allows

similar speech units from different dialects with the same basephones to be clustered. This technique corresponds to Schultz and Waibel's ML-tag approach described in the previous section.

4. Similar to the DCD-MR approach, the DCD-OR approach performs clustering of dialect-context-dependent speech units. However, in this case a one-root tree structure is employed. The question set consists of questions regarding context, the basephone and the dialect of a context-dependent unit. This allows similar context-dependent units from different dialects with different basephones to be clustered.

Experiments were based on databases recorded in Argentina, the Caribbean (Venezuela), Colombia, Mexico and Spain. The database from Spain contained speech from 3500 speakers in the training set and 500 speakers in the test set while the databases from Latin America each contained speech from 800 speakers in the corresponding training set and 200 speakers in the corresponding test set. The number of utterances in each set is shown in Table 2.5.

Each test utterance consisted of only one word, which implies that isolated word recognition was performed. From [52] it seems that the same unigram language model (or possibly a zerogram language model) and the same pronunciation dictionary were used for all experiments. The vocabulary used was identical for all dialects and consisted of about 4500 words which included all the words appearing in the test set. As a baseline, a monodialectal recogniser (employing a dialect-specific model set) was trained for each dialect using a multiroot decision-tree structure. This corresponds to the ML-sep approach described in the previous section. The performance of systems employing the different acoustic modelling approaches is presented in Table 2.6. Oracle recognition, where the correct dialect is assumed to be known, was performed in all experiments.

All systems showed improved performance in comparison to the baseline recognisers. GPS-MR and GPS-OR resulted in similar improvements. The DCD-MR system further improved on the GPS-based systems and the DCD-OR system performed best. The use of the one-root tree structure led to modest improvements, both where dialect-independent models were used (i.e. the GPS-based systems), as well as where multidialectal acoustic modelling (i.e. the DCD-based systems) was considered. In summary, it seems that selective cross-dialect sharing which is enabled when the decision-tree clustering process is extended by allowing questions relating to dialect to be asked (the DCD-based models) is superior to traditional accent-specific (baseline)

Table 2.5: Number of training and test set utterances for the different dialects considered in the research presented by Caballero et al. [52].

Dataset	Argentina	Caribbean	Colombia	Mexico	Spain	Total
Training	9568	9303	8874	11 506	40 936	80 187
Test	722	686	640	624	718	3390

Table 2.6: WERs (%) for the multidialectal recognition systems based on various acoustic modelling approaches considered by Caballero et al. [52].

Dialect	GPS-MR	GPS-OR	DCD-MR	DCD-OR	Baseline
Argentina	8.31	7.76	6.37	6.23	7.34
Caribbean	6.27	6.27	6.41	6.41	6.71
Colombia	8.28	8.28	7.97	7.81	9.22
Mexico	8.01	8.17	9.62	8.65	10.10
Spain	4.74	4.60	4.46	4.04	3.62
Average	7.12	7.02	6.97	6.63	7.40

and accent-independent (GPS-based) modelling for these dialects when applied in a tied-mixture topology.

2.6 Simultaneous Recognition of Multiple Accents

As described in Section 2.1.2, several recognition scenarios and configurations can be considered when evaluating systems designed for simultaneous recognition of multiple accents. In that section, parallel recognition was described in which a bank of accent-specific recognisers is run in parallel and the output with the highest associated likelihood is selected (illustrated in Figure 2.3). Accent identification (AID) is thus performed implicitly during recognition. Alternatively, oracle recognition can be performed in which test utterances are presented to matching accent-specific recognisers. Acoustic modelling effects are thus isolated since the effects of accent misclassifications are avoided. This scenario is illustrated in Figure 2.4. Other approaches when attempting simultaneous recognition of several accents include using a single accent-independent recogniser (Figure 2.5) or explicitly preceding accent-specific speech recognition with AID (Figure 2.6).

In the preceding sections of this chapter we attempted to indicate which of these scenarios were applied in each of the described studies. Several authors performed oracle recognition experiments, including Caballero et al. [51, 52] (see Section 2.5) and many of the authors referenced in Section 2.3.1. Accent-independent acoustic modelling was also considered in many of the studies described in Section 2.3.1. Below, a brief outline is given of literature in which some of these recognition configurations for the simultaneous recognition of multiple accents were explicitly considered and compared.

As already described in Section 2.3.1, Van Compernolle et al. [2] considered the effects of accent misclassifications by comparing oracle and parallel recognition in isolated digit recognition of Dutch and Flemish. Oracle and parallel recognition configurations respectively achieved WERs of 4.49% and 4.90% for Flemish and 5.28% and 5.99% for Dutch. In Section 2.3.1 we also discussed a study by Chengalvarayan [4] in which parallel and accent-independent recognition for connected digit recognition of American, Australian and British dialects of English were compared. Chengalvarayan found that accent-independent models performed best.

In a study similar to [3] described in Section 2.3.1, Teixeira et al. [53] performed AID followed by accent-specific speech recognition for non-native English accents from six European countries. The corpus used consisted of 200 isolated English words spoken twice by 20 speakers from each non-native accent. Of this data, 60% were used for training and the remaining 40% for testing. By performing AID for each test utterance using accent-specific phone HMMs, the authors obtained an average identification accuracy of 65.48%. When performing identification followed by accent-specific speech recognition, word recognition accuracies were found to be similar to those obtained with an oracle system (the former achieved a word recognition accuracy of 80.36% and the latter 80.78%). Both these approaches were, however, outperformed by an accent-independent system trained on the pooled data, which gave an accuracy of 85.63%.

In [54] Faria describes AID followed by speech recognition when distinguishing between native American English and non-native English (speech from Indian, Chinese, Russian, Spanish, German, French and other speakers). A classifier based on Gaussian mixture models (GMMs) was used to distinguish between native and non-native speech. Based on this distinction, the appropriate accent-specific acoustic models and integrated language model were selected during recognition. Compared to the case where native acoustic models and a native language model were employed (yielding an average WER of 57.20%), recognition accuracy improved to 53.55% WER when AID followed by recognition was performed. An oracle system led to an even lower

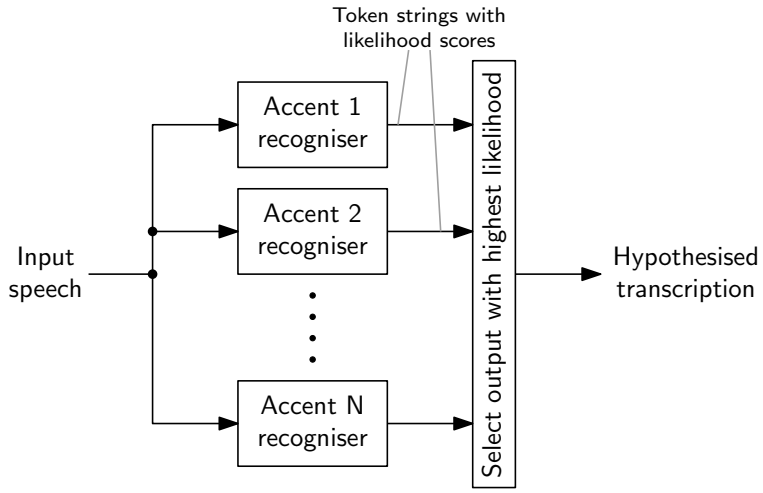


Figure 2.3: A speech recognition system employing multiple accent-specific recognisers in parallel for simultaneous recognition of several accents.

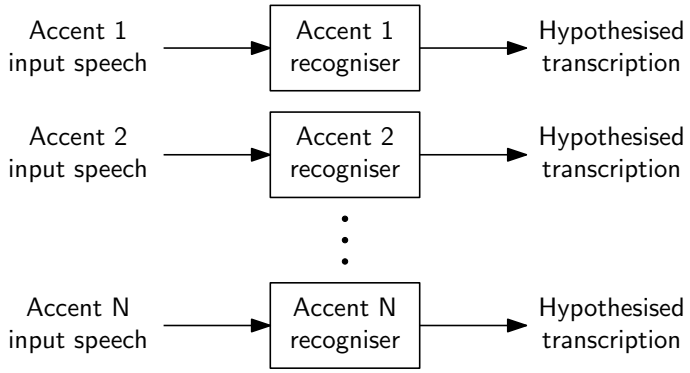


Figure 2.4: Oracle speech recognition in which test utterances are presented only to the accent-specific recognition system matching the accent of that utterance.

Figure 2.5: Simultaneous speech recognition of several accents using a single accent-independent system.

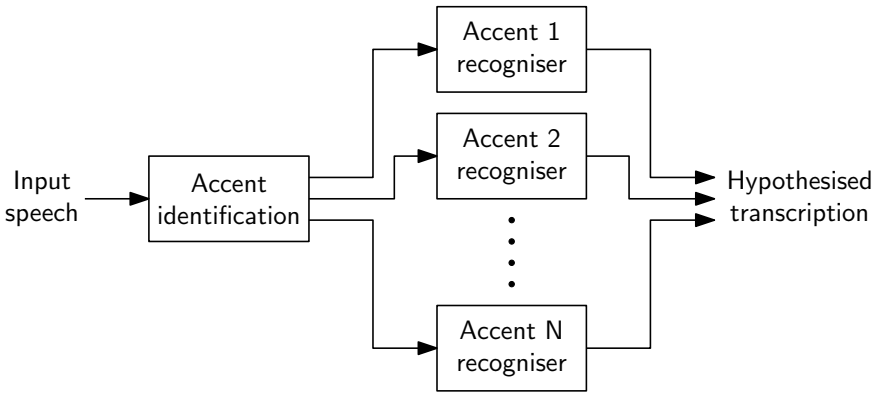


Figure 2.6: A multi-accent speech recognition system in which accent-specific recognition is preceded with accent identification.

WER of 51.70%. The GMM classifier obtained a classification accuracy of 69.5%. Native and non-native training sets both consisted of approximately 60 hours of speech from 374 speakers, while test sets consisted of approximately 17 hours of speech from 100 speakers. Faria notes that, in retrospect, accent-independent acoustic and language models trained on all the data should also have been considered.

2.7 Summary and Conclusions

In this chapter we surveyed the literature dealing with modelling and recognition of accented speech. We noted that two streams of research can be identified, namely research focussing on pronunciation modelling (Section 2.2) and research focussing on acoustic modelling (Section 2.3). We also highlighted relevant research focussing on multilingual speech recognition (Section 2.4) and recent research dealing with multidialectal acoustic modelling (Section 2.5). Finally we described different recognition configurations which can be used when processing speech in multiple accents (Section 2.6).

Research by Humphries et al. [20] led us to conclude in Section 2.2.3 that explicit modelling of accented variants in the pronunciation dictionary is less advantageous when larger amounts of accent-specific data are available. This scenario is also only relevant in situations where accent-specific pronunciations are not available for some of the accents considered. Accent-specific pronunciation dictionaries have been developed for each of the five accents of South African English as part of the AST project (see Section 5.2). The decision was therefore made to focus on acoustic modelling in our study of speech recognition of the South African English accents.

In Section 2.3 we noted that the comparative merits of the traditional accent-specific and accent-independent acoustic modelling approaches as well as the relative merits of several adaptation and interpolation techniques appear to depend on factors such as the abundance, character and availability of acoustic training data; on the type of task; on the recognition setup; and on the similarity of the accents involved. The question of how and whether acoustic training data should be shared across accents or whether accents should be modelled separately is dependent on all these factors. In Section 2.5 we described experiments carried out by Caballero et al. [51, 52] based on techniques previously applied to multiple languages, in which this training data partitioning can be achieved automatically. A data-driven decision is made to decide whether data from different dialects should be shared or should be separated. Tied-mixture systems based on this approach outperformed systems based on both dialect-specific and dialect-independent acoustic modelling for dialects of Spanish. Niesler [5] followed a similar approach for multilingual speech recognition and also obtained encouraging results (Section 2.4).

The findings of Caballero et al. and the work previously conducted by Niesler at Stellenbosch University led to a decision to focus on a similar comparison of the traditional accent-specific and accent-independent modelling approaches, and to compare these with a multi-accent acoustic modelling approach for the five accents of South African English. We described different recognition scenarios in Section 2.6 and, although both Niesler and Caballero et al. only considered oracle recognition, we attempt to evaluate both oracle and parallel recognition and to analyse the effects of the associated accent misclassifications.

Chapter 3

Speech Databases

In this chapter we give an overview of the speech databases used and the accents of South African English (SAE) considered in this research. First, a brief description is given of the format of the speech databases. Subsequently, each of the five accents of SAE identified in the literature [6] and represented in the databases is described. The separation of the speech data into training, development and evaluation sets is discussed next, followed by a description of the phone set used for the phonetic transcriptions of the speech data. The chapter is concluded with an analysis of acoustic similarity between the different accents of SAE. Because it is not always obvious whether we are dealing with accents or dialects when considering varieties of SAE (see Section 2.1.1), we shall consistently use the term 'accent' to avoid confusion.

3.1 The AST Databases

Our experiments were based on the African Speech Technology (AST) databases [55]. As part of the AST project, a total of eleven databases were collected for five languages spoken in South Africa: Xhosa, Southern Sotho, Zulu, English and Afrikaans. For Afrikaans, three databases were compiled corresponding to three varieties of Afrikaans, while for English, five accented databases were developed. 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 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. All speech audio was sampled at 8 kHz and stored with 16-bit precision.

As noted above, speech databases of five English accents were compiled. These correspond to the following groups of speakers [55]:

1. White Afrikaans second language English speakers (AE database).
2. Black second language English speakers (BE database).
3. First and second language coloured speakers of English (CE database).
4. White mother-tongue English speakers (EE database).
5. Indian/Asian mother-tongue English speakers (IE database).

Within the AST project, the assignment of a speaker's accent was thus guided by the speaker's first language and race. Respectively, the above databases roughly correspond to the following varieties of English described in [6]: Afrikaans English, Black South African English, Cape Flats

English, White South African English and Indian South African English. It is important to note that although the labels used to differentiate between these accents are not intended to reflect the Apartheid classifications, there exists an undeniable correlation between the different varieties of English used in South Africa and the different ethnic groups. To some degree, these labels therefore reflect the influence of South Africa's legacy on the development of English and its varieties. In Table 3.1 an indication is given of the proportion of the South African population using each of these accents. It is evident from these statistics that non-mother-tongue variants of English (spoken by white Afrikaans, black and some coloured speakers) are used by the overwhelming majority of the population. In the following section a brief description is given of each accent and how it is represented in the AST databases.

Table 3.1: Percentage of the population falling into specific speaker groups, loosely indicating the proportion of speakers of a corresponding SAE accent [1]. 'Other' refers to speakers not falling into one of the relevant groups, for example a white person using Xhosa as a first language.

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

3.2 Accents of South African English

3.2.1 White 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 [1]. 'White South African English' refers to the first language English spoken by white South Africans, chiefly of British descent. The old term, 'South African English', which was once used to refer to this accent, is now used to refer to all the accents of English spoken in South Africa [56]. The term 'Standard South African English' is also sometimes used when referring to White South African English. This accent is used by approximately 3.8% of the South African population (Table 3.1).

According to [56], White South African English can further be divided into three groups based on social variation: 'Cultivated', associated with the upper class; 'General', associated with the middle class; and 'Broad', associated with the working class. When the specific social variant is not explicitly stated, White South African English often actually refers to General White South African English. Broad White South African English is closely associated with the Afrikaans English accent, used as a second language by speakers of Afrikaans descent. Most of the AST EE database (speech from white mother-tongue English speakers) can be considered General White South African English and we will use the abbreviation EE to refer to this accent.

When considering the phonology, morphology and syntax of White South African English, as described in [56, 57], Bowerman notes the influence of Afrikaans on White South African English as an important feature. He also writes, however, that some features previously accepted as

Afrikaans influences may be attributed to input from Settler English.[1] Nevertheless, ample evidence exists to support the influence of Afrikaans on White South African English.

3.2.2 Afrikaans English

'Afrikaans English' refers to the accent used by white South African second language English speakers of Afrikaans descent. Afrikaans is a Germanic language and its origins lie in 17th century Dutch which was brought to South Africa by settlers from the Netherlands. The language 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 3.1, white Afrikaans speakers comprise approximately 5.7% of the South African population.

As mentioned in the previous section, Broad White South African English is closely associated with the Afrikaans English accent. For example, Afrikaans English is not considered a separate variety in [6], but rather discussed implicitly as part of White South African English. The influence of Afrikaans on White South African English is undeniable and there should therefore exist some degree of similarity between Afrikaans English and White South African English. The AST AE database (speech from white Afrikaans second language English speakers) corresponds to the Afrikaans English accent and we shall use the abbreviation AE to refer to this accent.

3.2.3 Black South African English

The term 'Black South African English' applies to the non-mother-tongue English spoken by black South Africans. Mesthrie [58] writes that although some black South Africans now speak English as a first language, these speakers should not necessarily be considered speakers of Black South African English. 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 3.1), it is not surprising that Black South African English has become prominent in government, commerce and the media since 1994 [59]. 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). We will use the abbreviation BE to refer to this accent.

3.2.4 Cape Flats English

'Caps Flats English' has its roots in the 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. In one of these neighbourhoods, District Six, the languages spoken included Yiddish, Russian, Polish, Afrikaans, English, Zulu, Xhosa and Sotho, and these all eventually contributed to Cape Flats English. 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 [60]. The term 'coloured' ('bruinmens' or previously 'kleurling' in Afrikaans) refers to the mixed-race ethnic group most closely associated with the Cape Flats English accent today. As indicated in Table 3.1, this accent is used by approximately 8.8% of the South African population. The

[1] Settler English was spoken specifically by English settlers arriving in 1820 and 1821 as part of a settlement programme which Britian launched after the Cape was established as a British colony in 1815. Although they homogeneously spoke English as a first language, several regional dialects were used.

diverse origins of these people can be traced back to Europe, Indonesia, Madagascar, Malaysia, Mozambique, Mauritius, Saint Helena and Southern Africa.

Cape Flats English is also sometimes called 'Coloured English' but this term becomes problematic because not all people considered coloured speak this dialect, and because the ethnic classification of coloured itself could be considered somewhat controversial. Today, many coloured speakers use a dialect of Afrikaans as a home language, although English is also considered a first language in many homes. In fact, McCormick writes that most speakers of Cape Flats English also use a dialect of Afrikaans and "are likely to switch between Afrikaans and Cape Flats English, even within the same conversation" [61]. The connection between Cape Flats English and Afrikaans English, both being closely associated with Broad White South African English (see Section 3.2.1), is also emphasised by Finn [60]. The Cape Flats English variant is associated with the CE database from the AST project, which includes speech gathered from either mother-tongue English or mother-tongue Afrikaans coloured speakers. The abbreviation CE will therefore be used to refer to the Cape Flats English accent.

3.2.5 Indian South African English

Indian languages were brought to South Africa by labourers who were recruited from India after the abolition of slavery in European colonies in the 19^{th} century. These Indian languages have existed in South Africa since 1860, mainly in Natal (today KwaZulu-Natal). 'Indian South African English' presents an interesting sociolinguistic case: the variety shifted from being associated with second language speakers (originally as a lingua franca) to being a first language, despite the Apartheid policy (1948-1991) preventing contact between Indian children and first language English speakers [62]. Today, the majority of South African Indians use English as a first language. According to [1], approximately 2.5% of the South African population are considered Indian or Asian and 94% speak English as a first language, as indicated in Table 3.1.

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 [62, 63]. The influence of other SAE accents, especially those spoken in KwaZulu-Natal (mainly White South African English and Black South African English), are also emphasised. The AST IE database contains speech gathered from predominantly Indian mother-tongue English speakers. We shall use the abbreviation IE to refer to the Indian South African English accent represented in this database.

3.3 Training and Test Sets

The five English AST databases were all divided into training, development and evaluation sets, as indicated in Tables 3.2, 3.3 and 3.4, 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. For all the experiments conducted as part of this research, 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. In addition to using these sets for acoustic modelling, their word-level transcriptions were used for language modelling purposes.

Table 3.2: Training set partition for each English AST 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	9878	245	178 954	47 279
IE	7.21	15 073	295	218 372	57 253
Total	31.78	54 078	1240	911 061	241 064

Table 3.3: Development set partition for each English AST 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 3.4: Evaluation set partition for each English AST 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

3.4 Phone Set

As part of the AST project phonetic transcriptions of the five English databases were compiled by linguistic experts using a large IPA-based phone set, similar to that used in [5]. Since certain phones occurred only in some of the databases and with very low frequency, phones were mapped to a smaller set of 50 phones common to all five accents. Examples of these phones include clicks in the names of black speakers, such as the dental click [|] which occurred only 5 times in the training corpus in names such as 'Mcira' and 'Nocwaka'. It was mapped to the voiceless velar plosive [k]. All the phone mappings are listed in Table B.2. Fewer than 1.4% of all the phone tokens were affected by this process. The same IPA-based phone set was therefore used for the transcriptions of all five databases. This final phone set after mapping is listed in Table B.1.[2]

Although an analysis of phone set overlap is not relevant for the transcriptions obtained after mapping, an early indication of accent similarity can be obtained by considering the transcriptions prior to mapping. Table 3.5 shows the extent to which the original phone set of each accent

[2]These phone mappings were applied prior to the involvement of the author in this research.

Table 3.5: Degree to which the phone types of each accent cover the training set phone types and tokens of every other accent, prior to the phone mapping indicated in Table B.2.

Phone types of	Covers % of training set phone types/tokens in				
	AE	BE	CE	EE	IE
AE	100/100	59.4/99.6	88.1/100	93.1/100	84.1/100
BE	96.6/100	100/100	94.0/100	98.3/100	92.8/100
CE	100/100	65.6/99.8	100/100	98.3/100	89.9/100
EE	91.5/99.9	59.4/99.5	85.1/99.9	100/100	84.1/100
IE	98.3/100	66.6/99.8	92.5/100	100/100	100/100

covers the phone types and tokens found in the other accents' training sets prior to mapping. Phone types refer to the number of different phones that occur in the data, while phone tokens indicate their total number. For example, 91.5% of the phone types in the AE training set are present among the EE phone types, while 99.9% of the phone tokens in the AE training set are present among the EE phone types. The analysis indicates that most of the original AE and EE phone types are covered by the other accents. The CE and IE types are covered less well and the BE phone types are the most poorly covered by the other accents. The analysis also indicates that in all cases (even BE) a very high proportion of phone tokens are covered by the other accents and this further justifies the phone mapping that was performed.

3.5 Estimation of Accent Similarity

To obtain some initial intuition regarding the relative acoustic similarity between the five SAE accents, we considered a computational method allowing the similarity between accents to be estimated. This can be achieved by determining the similarity between the probability density functions (PDFs) associated with two sets of HMMs trained on different accents.[3] Several such measures have been proposed in the literature, including the Kullback-Leibler divergence [64, 65] and the Bhattacharyya bound [66]. 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 [67, 68, 69, 70]. 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.

3.5.1 The Bhattacharyya Bound

The following theoretical development is taken in part from [65, pp. 39–41] and [66, pp. 97–99] and is similar to that presented in [71, pp. 13–16]. Assume we have to assign some observed value x to one of two classes C_1 or C_2 which are characterised by the joint PDFs $p(x, C_1)$ and $p(x, C_2)$ as illustrated in Figure 3.1. If we choose the decision boundary at $x = x_b$, the probability of making a classification error is

$$\begin{aligned} P(\text{error}) &= P(x \in C_1 \text{ but assigned to } C_2) + P(x \in C_2 \text{ but assigned to } C_1) \\ &= \int_{x_b}^{\infty} p(x, C_1)\,\mathrm{d}x + \int_{-\infty}^{x_b} p(x, C_2)\,\mathrm{d}x \end{aligned} \tag{3.1}$$

which is the sum of the coloured areas in Figure 3.1.

[3] The training procedure used to obtain these HMMs will be described later in Section 5.3.1. However, it is useful to consider the accent similarity reflected by these models here.

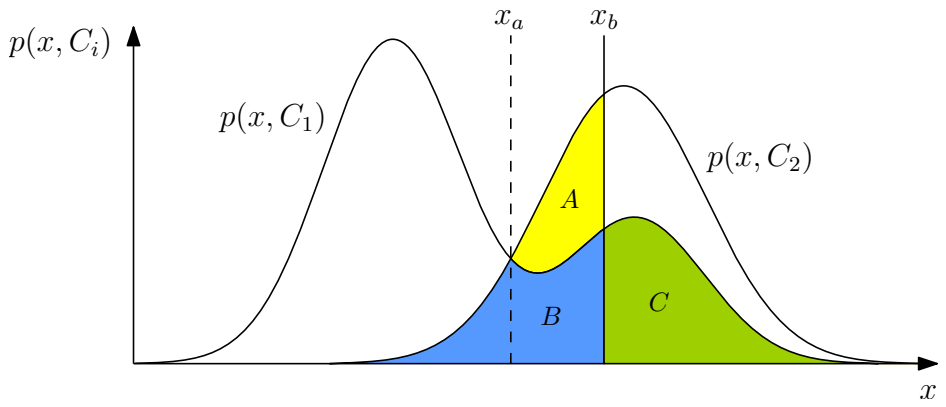


Figure 3.1: A representation of two decision boundaries (x_a and x_b) when considering classification of x as belonging to one of two classes C_1 or C_2 which is characterised by the PDFs $p(x, C_1)$ and $p(x, C_2)$.

It is apparent from Figure 3.1 that the minimum error would result when choosing the decision boundary at $x = x_a$ because the yellow coloured area A would disappear. The resulting minimum error (which is simply the sum of the coloured areas B and C) is referred to as the Bayes error:

$$\begin{aligned}\varepsilon &= \int_{x_a}^{\infty} p(x, C_1)\,\mathrm{d}x + \int_{-\infty}^{x_a} p(x, C_2)\,\mathrm{d}x \\ &= \int_{-\infty}^{\infty} \min\left[p(x, C_1), p(x, C_2)\right]\,\mathrm{d}x \end{aligned} \tag{3.2}$$

In the multivariate case, equation (3.2) can be written as

$$\begin{aligned}\varepsilon &= \int_{\mathbf{x}\in\mathbb{R}^n} \min\left[p(\mathbf{x}, C_1), p(\mathbf{x}, C_2)\right]\,\mathrm{d}\mathbf{x} \\ &= \int_{\mathbf{x}\in\mathbb{R}^n} \min\left[P_1\, p_1(\mathbf{x}), P_2\, p_2(\mathbf{x})\right]\,\mathrm{d}\mathbf{x} \end{aligned} \tag{3.3}$$

with $p_i(\mathbf{x}) = p(\mathbf{x}|C_i)$, $P_i = P(C_i)$ and integration taking place over the whole vector space $\mathbf{x} \in \mathbb{R}^n$ with n the dimensionality of $\mathbf{x}$.

The intuition behind the Bhattacharyya bound is that if two PDFs $p_1(\mathbf{x})$ and $p_2(\mathbf{x})$ are very similar, the Bayes error in equation (3.3) would be higher than for two very different PDFs because it would be harder to distinguish between the two classes. Unfortunately, it is not possible to obtain a closed-form solution for equation (3.3) in general. However, by assuming Gaussian PDFs, a closed-form solution for an upper bound to equation (3.3) can be found.

By using the identity (the proof is given in [66, p. 98]):

$$\min[a, b] \leq a^s b^{1-s} \text{ for } 0 \leq s \leq 1 \tag{3.4}$$

with $a, b \geq 0$, an upper bound to equation (3.3) can be determined. If we do not insist on the optimisation of s and choose $s = 1/2$, we obtain the Bhattacharyya bound:

$$\varepsilon_u = \sqrt{P_1\, P_2} \int \sqrt{p_1(\mathbf{x})\, p_2(\mathbf{x})}\,\mathrm{d}\mathbf{x} \tag{3.5}$$

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 [66, p. 99]:

$$\varepsilon_u = \sqrt{P_1\, P_2}\; e^{-B} \tag{3.6}$$

where

$$B = \frac{1}{8}(\mu_2 - \mu_1)^{\mathrm{T}} \left[\frac{\Sigma_1 + \Sigma_2}{2}\right]^{-1} (\mu_2 - \mu_1) + \frac{1}{2} \ln \frac{\left|\frac{\Sigma_1 + \Sigma_2}{2}\right|}{\sqrt{|\Sigma_1|\,|\Sigma_2|}} \tag{3.7}$$

The term B is known as the Bhattacharyya distance. When we assume the prior probabilities to be equal, as suggested in [67, 69], ε_u is bounded $0 \leq \varepsilon_u \leq 1/2$ with $\varepsilon_u = 1/2$ when the PDFs are identical and ε_u approach zero when they are very different.

3.5.2 Similarity Between Accent Pairs

In order to develop some intuition regarding the five accents of SAE, we used the Bhattacharyya bound to compute the degree of similarity between each pair of accents. As part of the experiments performed in this research, three-state left-to-right single-mixture monophone HMMs with diagonal covariance matrices were trained for each accent individually (see Section 5.3.1). The single-mixture monophone HMMs were used since the Bhattacharyya bound cannot easily be determined for Gaussian mixture PDFs. 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 was computed, where each individual similarity was weighted by the frequency of occurrence of the phone in the training set. This final figure gives an indication of the similarity between two accents.

The approach described above was applied to the five SAE accents and the results are presented in Table 3.6. It is evident that the highest degree of similarity exists between the AE, CE and EE accents. The maximum similarity (0.4030) is observed between AE and CE, and the second highest (0.3929) between AE and EE. 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 3.6 appear to indicate that AE, CE and EE are quite similar, with IE lying further away and BE being the most dissimilar from the others. When the same analysis was considered but only the vowel monophone HMMs were used, similar trends were observed.

Table 3.6: 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

3.6 Summary

In this chapter we gave an overview of the speech databases and accents of SAE considered in this research. In Section 3.1 we described the AST databases on which this research is based. The five English AST databases correspond to the five accents of SAE discussed in Section 3.2. In Section 3.3 we indicated how these five databases were divided into training and test sets. A description of the phone set used for the phonetic transcriptions of the databases was given in Section 3.4. Finally, an analysis was presented in Section 3.5 in which we attempted to estimate the similarity of the different SAE accents. This analysis indicated that the AE, CE and EE accents appear to be quite similar while the BE and IE accents appear to be more dissimilar.

Chapter 4

Acoustic Modelling

In this chapter we describe the three acoustic modelling approaches that form the basis of our research. The different acoustic modelling approaches are based on different strategies for decision-tree state clustering – a process which forms part of the standard procedure for training tied-state triphone HMMs. First we discuss the purpose of decision-tree state clustering and how it forms part of the tied-state triphone HMM training procedure. Next we describe the decision-tree construction process, followed by the different strategies for state clustering which correspond to the three acoustic modelling approaches. We conclude the chapter with a description of how the different clustering strategies are applied in the model training procedure as a whole.

4.1 Decision-Tree State Clustering

4.1.1 Modelling Context Dependency

Since human speech is constrained by the limitations of the physical articulators producing the speech sounds, a conceptually isolated speech unit is usually influenced by the preceding and/or following speech units. This effect, referred to as 'co-articulation', is therefore directly related to the context in which a particular speech unit occurs. In general, improved speech recognition performance is achieved when context-dependent speech units are used to explicitly model the effects of co-articulation [72, 73].

Context-dependent phonetic modelling is performed by considering each phone in the context in which it occurs. This can be achieved at several levels, including the use of monophones, where one model is used to represent a particular phone in any context; biphones, where each model represents a phone with a specific left or right context; and triphones, where each model represents a phone with specific left and right contexts.[1] In this research we have made use of cross-word triphones, meaning that the triphone contexts can cross word boundaries. An example of a triphone would be [j]−[i]+[k] which indicates the phone [i] (the basephone) with a left context of [j] and a right context of [k], i.e. [i] preceded by [j] and followed by [k].

Although we introduce an additional level of complexity by employing context-dependent models, the amount of training data for a particular context-dependent phone will be significantly less than for its context-independent counterpart and improved performance is therefore not guaranteed. The amount of training data should thus be sufficient for the reliable estimation of the model parameters. In this research a set of 50 phones with additional silence and speaker

[1] Some researches use wider contexts still, e.g. quintphones [23, 37].

models were used resulting in a total of 135 200 possible triphone models.[2] Many of these triphones never occur in the training data and even when the phones do occur, there may not be enough training data for reliable parameter estimation. Decision-tree state clustering provides a solution to this data sparseness problem. Similar triphone HMM states are clustered to ensure that parameter estimation is based on an adequate amount of training data. Furthermore, the decision-trees can be used to synthesise models for unseen context-dependent phones which may be required during recognition.

4.1.2 Tied-State Triphone HMM Systems

This section presents an overview of the process used to build tied-state HMM systems, with particular emphasis on the decision-tree state clustering process. A triphone-based system is considered, but the same procedure is followed when developing a tied-state system based on any other context-dependent speech unit. The content of this section is based on [74] and [75, Ch. 10].

The development of a tied-state HMM system is normally initiated by training a set of single-mixture monophone HMMs. Subsequently, these monophone HMMs are cloned to obtain seed HMMs for all triphones that occur in the training set; each seed triphone HMM is cloned from the corresponding monophone HMM. These models are then retrained, using embedded Baum-Welch re-estimation, to obtain a set of single-mixture triphone HMMs. Usually the transition probabilities of all triphones with the same basephone are tied. The states of these single-mixture triphone HMMs are then clustered and tied using decision-tree state clustering.

The clustering process is normally initiated by pooling the data of corresponding states from all triphones with the same basephone within a single cluster. This is done for all triphones 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 triphone is a vowel or whether the right context is a silence. There are in general many such questions and each potential question results in a split which results in an increase in training set likelihood. For each cluster the optimal question (leading to the largest likelihood increase) is determined. In this way clusters are subdivided repeatedly until either the increase in likelihood or the number of observation vectors associated with a resulting cluster (i.e. the cluster occupancy count) falls below a certain predefined threshold.

The result is a phonetically-motivated binary decision-tree where the leaf nodes represent clusters of triphone HMM states for which data should be pooled. An example of such a decision-tree is shown in Figure 4.1. In this figure, for example, the leaf node A is a set of all triphones with basephone [i] for which the left context is a voiced vowel and the right context is the phone [m]. The advantage of this approach is that the amount of data available for training the observation probability density function (PDF) parameters of a particular cluster of states can be controlled by varying the minimum cluster occupancy. This ensures that model parameters are estimated on a sufficient amount of training data. Furthermore, each state of a triphone not seen in the training set can be associated with a leaf node in the decision-trees. This allows the synthesis of triphone HMMs that are required during recognition but are not present in the training set.

Clustering is normally followed by further iterations of embedded Baum-Welch re-estimation, after which the number of Gaussian mixtures per state is gradually increased. Each mixture increase is again followed by further iterations of re-estimation. This process of increasing mixtures and re-estimation is repeated until the desired number of mixture components is reached. The result is a set of tied-state triphone HMMs with multiple Gaussian mixtures per state.

[2] $52 \times 50 \times 52 = 135\,200$

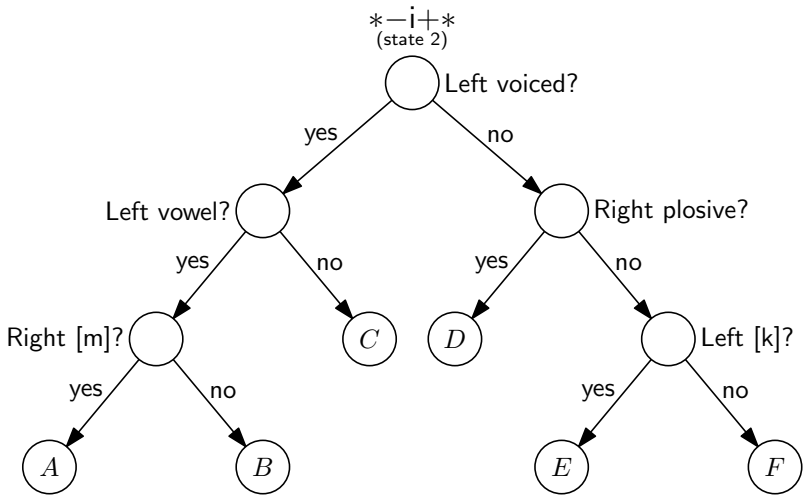


Figure 4.1: An example of a phonetic binary decision-tree where the second state of all triphones with the basephone [i] are clustered. The leaf nodes indicate clusters of triphone states which should be tied.

4.1.3 Phonetic Decision-Trees Construction

The discussion presented in this section is based on [72, 73, 74]. From the description in the preceding section it is evident that a number of values need to be calculated in order to construct the phonetic decision-trees. The problems associated with the calculation of these values can be summarised as follows:

1. Given a cluster of states, we must be able to calculate the mean vector and covariance matrix for the observation PDF associated with the whole cluster.
2. Given a cluster of states, the training set likelihood of the observation vectors associated with that cluster must be computable.
3. Given multiple clusters of states, we should be able to calculate the training set likelihood associated with the combination of these clusters. This is required in order to determine the optimal question with which to split a particular cluster.

To ensure that the decision-tree clustering process is computationally tractable, the above values need to be computable from only the mean vectors, covariance matrices and state occupancy counts associated with the individual states of the respective clusters. In other words, the values must be computable without requiring recourse to the observation vectors themselves.

In order to proceed we need to make the following assumptions [72, pp. 37–38]:

- The assignment of observation vectors to states $\gamma_s(\mathbf{o}_f)$ is not altered by the clustering process, where formally $\gamma_s(\mathbf{o}_f)$ is the a posteriori probability of the observation vector $\mathbf{o}_f$ being generated by HMM state s. In other words, we assume that the state alignment is not changed by the clustering process.
- The contribution of the transition probabilities in the calculation of the training set likelihood can be ignored. Although the transition probabilities do contribute to the total likelihood, the transition probabilities will not change if the assignment of observation vectors to states does not change (the previous assumption).
- In the following derivation we assume a "hard" assignment of observation vectors to states, as would be used in the Viterbi HMM training procedure. Odell however states in [72, p. 38] that this assumption is a good approximation of the probabilistic case in which "soft" assignments are made, as in the Baum-Welch re-estimation procedure.

Problem 1: Estimating the PDF Parameters

To ensure that the decision-tree state clustering process is computationally tractable, the mean vector and covariance matrix characterising the observation PDF of a whole cluster should be computable from the means, covariance matrices and state occupancy counts for the individual states populating that cluster. Let us consider the simple case of a set (or cluster) of HMM states $\mathbf{S}$ consisting of two states only, i.e. $\mathbf{S} = \{s_1, s_2\}$. If we assume that the observation vectors for the set of training frames $\mathbf{F}_1$ and the observation vectors for the set of training frames $\mathbf{F}_2$ are assigned to states s_1 and s_2, respectively, then the mean of the observation PDF of s_1 can be calculated as

$$\mu_{s_1} = \frac{1}{N_1} \sum_{f \in \mathbf{F_1}} \mathbf{o}_f \tag{4.1}$$

where

$$N_1 = \sum_{f \in \mathbf{F}} \gamma_{s_1}(\mathbf{o}_f) \tag{4.2}$$

Frames $\mathbf{F}$ are all the frames in the training set and $\gamma_{s_i}(\mathbf{o}_f)$ is the a posteriori probability of the observation vector $\mathbf{o}_f$ being generated by HMM state s_i. A similar set of equations holds for s_2. The mean of the cluster $\mathbf{S}$ can then be calculated as

$$\begin{aligned}
\mu(\mathbf{S}) &= \frac{1}{N} \sum_{f \in \mathbf{F}} \mathbf{o}_f \\
&= \frac{1}{N_1 + N_2} \left[\sum_{f \in \mathbf{F_1}} \mathbf{o}_f + \sum_{f \in \mathbf{F_2}} \mathbf{o}_f \right] \\
&= \frac{N_1 \cdot \mu_{s_1} + N_2 \cdot \mu_{s_2}}{N_1 + N_2} \\
&= \frac{\mu_{s_1} \cdot \sum\limits_{f \in \mathbf{F}} \gamma_{s_1}(\mathbf{o}_f) + \mu_{s_2} \cdot \sum\limits_{f \in \mathbf{F}} \gamma_{s_2}(\mathbf{o}_f)}{\sum\limits_{f \in \mathbf{F}} \gamma_{s_1}(\mathbf{o}_f) + \sum\limits_{f \in \mathbf{F}} \gamma_{s_2}(\mathbf{o}_f)}
\end{aligned} \tag{4.3}$$

Generalising equation (4.3), it can be shown that the mean of a cluster $\mathbf{S}$ populated by an arbitrary number of states can be calculated as

$$\mu(\mathbf{S}) = \frac{\sum\limits_{s \in \mathbf{S}} \mu_s \cdot \sum\limits_{f \in \mathbf{F}} \gamma_s(\mathbf{o}_f)}{\sum\limits_{s \in \mathbf{S}} \sum\limits_{f \in \mathbf{F}} \gamma_s(\mathbf{o}_f)} \tag{4.4}$$

Equation (4.4) is equivalent to equation (4.22) given in [73, p. 65]. The mean of the observation PDF for a cluster of HMM states is thus calculable from the means of the observation PDFs and the state occupancy counts of the individual states in that cluster.

A similar approach can be followed to obtain an expression for the covariance matrix $\Sigma(\mathbf{S})$ for a cluster of HMM states. Suppose again we have two states $\mathbf{S} = \{s_1, s_2\}$ and that the set of observation vectors for training frames $\mathbf{F}_1$ and $\mathbf{F}_2$ are assigned to states s_1 and s_2, respectively. The covariance matrix for the observation PDF of the first state s_1 can then be calculated as

$$\Sigma_{s_1} = \frac{1}{N_1} \sum_{f \in \mathbf{F}_1} (\mathbf{o}_f - \mu_{s_1})(\mathbf{o}_f - \mu_{s_1})^{\mathrm{T}} \tag{4.5}$$

with a similar equation for the covariance matrix Σ_{s_2} of state s_2. The covariance matrix of the PDF describing the cluster $\mathbf{S}$ then is

$$\begin{aligned}\Sigma(\mathbf{S}) &= \frac{1}{N}\sum_{f\in\mathbf{F}}(\mathbf{o}_f-\mu(\mathbf{S}))(\mathbf{o}_f-\mu(\mathbf{S}))^{\mathrm{T}} \\ &= \frac{1}{N_1+N_2}\left[\sum_{f\in\mathbf{F}_1}(\mathbf{o}_f-\mu(\mathbf{S}))(\mathbf{o}_f-\mu(\mathbf{S}))^{\mathrm{T}}+\sum_{f\in\mathbf{F}_2}(\mathbf{o}_f-\mu(\mathbf{S}))(\mathbf{o}_f-\mu(\mathbf{S}))^{\mathrm{T}}\right] \end{aligned} \tag{4.6}$$

In Appendix A.1 it is shown that it follows from equations (4.1), (4.2) and (4.5) that

$$\sum_{f\in\mathbf{F}_1}(\mathbf{o}_f-\mu(\mathbf{S}))(\mathbf{o}_f-\mu(\mathbf{S}))^{\mathrm{T}} = N_1\cdot\left[\Sigma_{s_1}+(\mu_{s_1}-\mu(\mathbf{S}))(\mu_{s_1}-\mu(\mathbf{S}))^{\mathrm{T}}\right] \tag{4.7}$$

and by using this result, equation (4.6) can be written as

$$\Sigma(\mathbf{S}) = \frac{N_1\cdot\left[\Sigma_{s_1}+(\mu_{s_1}-\mu(\mathbf{S}))(\mu_{s_1}-\mu(\mathbf{S}))^{\mathrm{T}}\right]+N_2\cdot\left[\Sigma_{s_2}+(\mu_{s_2}-\mu(\mathbf{S}))(\mu_{s_2}-\mu(\mathbf{S}))^{\mathrm{T}}\right]}{N_1+N_2} \tag{4.8}$$

It follows from equation (4.8) that, for a cluster $\mathbf{S}$ consisting of an arbitrary number of states, the covariance matrix of the PDF for the cluster can be calculated as

$$\Sigma(\mathbf{S}) = \frac{\sum_{s\in\mathbf{S}}\sum_{f\in\mathbf{F}}\gamma_s(\mathbf{o}_f)\cdot\left[\Sigma_s+(\mu_s-\mu(\mathbf{S}))(\mu_s-\mu(\mathbf{S}))^{\mathrm{T}}\right]}{\sum_{s\in\mathbf{S}}\sum_{f\in\mathbf{F}}\gamma_s(\mathbf{o}_f)} \tag{4.9}$$

Equation (4.9) is equivalent to equation (4.23) given in [73, p. 65]. Only the means, covariance matrices and state occupancy counts of the individual states in the cluster are thus required to calculate the covariance matrix for the whole cluster.

Problem 2: Log Likelihood of a Cluster

Let $\mathbf{S}$ again denote a set of HMM states and let $L(\mathbf{S})$ be the log likelihood of the training observation vectors assigned to the states in $\mathbf{S}$, under the assumption that all states in $\mathbf{S}$ are tied with a common mean $\mu(\mathbf{S})$ and covariance matrix $\Sigma(\mathbf{S})$ and that the transition probabilities have a negligible effect on the log likelihood and can therefore be ignored. The log likelihood that the observation vectors were generated by the states $\mathbf{S}$ can then be calculated as

$$\begin{aligned}L(\mathbf{S}) &= \ln\prod_{f\in\mathbf{F}}p(\mathbf{o}_f|\mathbf{S}) \\ &= \sum_{f\in\mathbf{F}}\ln\left[\mathcal{N}(\mathbf{o}_f|\mu(\mathbf{S}),\Sigma(\mathbf{S}))\right]\end{aligned} \tag{4.10}$$

where $\mathbf{F}$ is the set of training frames for which the observation vectors are assigned to the states in $\mathbf{S}$, i.e. $\mathbf{F}=\{f : \mathbf{o}_f \text{ is generated by states in } \mathbf{S}\}$. The observation PDFs are assumed to be single-mixture Gaussian PDFs:

$$\mathcal{N}(\mathbf{o}_f|\mu(\mathbf{S}),\Sigma(\mathbf{S})) = \frac{1}{\sqrt{(2\pi)^n|\Sigma(\mathbf{S})|}}\exp\left\{-\frac{1}{2}(\mathbf{o}_f-\mu(\mathbf{S}))^{\mathrm{T}}\Sigma^{-1}(\mathbf{S})(\mathbf{o}_f-\mu(\mathbf{S}))\right\} \tag{4.11}$$

From equation (4.11), equation (4.10) can then be written as

$$L(\mathbf{S}) = -\frac{1}{2}\sum_{f\in\mathbf{F}}\left\{\ln[(2\pi)^n|\Sigma(\mathbf{S})|]+(\mathbf{o}_f-\mu(\mathbf{S}))^{\mathrm{T}}\Sigma^{-1}(\mathbf{S})(\mathbf{o}_f-\mu(\mathbf{S}))\right\} \tag{4.12}$$

The covariance matrix of the cluster of states $\mathbf{S}$ can be calculated as

$$\Sigma(\mathbf{S}) = \frac{1}{N} \sum_{f \in \mathbf{F}} (\mathbf{o}_f - \mu(\mathbf{S}))(\mathbf{o}_f - \mu(\mathbf{S}))^{\mathrm{T}} \tag{4.13}$$

where N is the number of frames in $\mathbf{F}$ and can be calculated as

$$N = \sum_{s \in \mathbf{S}} \sum_{f \in \mathbf{F}} \gamma_s(\mathbf{o}_f) \tag{4.14}$$

By cross-multiplication, equation (4.13) becomes

$$\mathbf{I} \cdot N = \sum_{f \in \mathbf{F}} \Sigma^{-1}(\mathbf{S}) \cdot (\mathbf{o}_f - \mu(\mathbf{S}))(\mathbf{o}_f - \mu(\mathbf{S}))^{\mathrm{T}} \tag{4.15}$$

In [73, p. 62] the matrix identity

$$\mathbf{x}^{\mathrm{T}} \mathbf{A} \mathbf{x} = \mathrm{tr}(\mathbf{A} \mathbf{x} \mathbf{x}^{\mathrm{T}}) \tag{4.16}$$

is given, where $\mathbf{x}$ is an $n \times 1$ vector and $\mathbf{A}$ is an $n \times n$ matrix and tr denotes the trace of a matrix. By taking the trace of both sides of equation (4.15) and then applying equation (4.16), we obtain

$$\begin{aligned} n \cdot N &= \mathrm{tr} \left\{ \sum_{f \in \mathbf{F}} \Sigma^{-1}(\mathbf{S}) \cdot (\mathbf{o}_f - \mu(\mathbf{S}))(\mathbf{o}_f - \mu(\mathbf{S}))^{\mathrm{T}} \right\} \\ &= \sum_{f \in \mathbf{F}} \mathrm{tr} \left\{ \Sigma^{-1}(\mathbf{S}) \cdot (\mathbf{o}_f - \mu(\mathbf{S}))(\mathbf{o}_f - \mu(\mathbf{S}))^{\mathrm{T}} \right\} \\ &= \sum_{f \in \mathbf{F}} (\mathbf{o}_f - \mu(\mathbf{S}))^{\mathrm{T}} \Sigma^{-1}(\mathbf{S}) (\mathbf{o}_f - \mu(\mathbf{S})) \end{aligned} \tag{4.17}$$

where n is the dimensionality of the observation vectors.

By substituting equation (4.17) into equation (4.12), we obtain the result

$$\begin{aligned} L(\mathbf{S}) &= -\frac{1}{2} \sum_{f \in \mathbf{F}} \ln[(2\pi)^n |\Sigma(\mathbf{S})|] - \frac{1}{2} \cdot n \cdot N \\ &= -\frac{1}{2} \left\{ \ln[(2\pi)^n |\Sigma(\mathbf{S})|] + n \right\} N \\ &= -\frac{1}{2} \left\{ \ln[(2\pi)^n |\Sigma(\mathbf{S})|] + n \right\} \sum_{s \in \mathbf{S}} \sum_{f \in \mathbf{F}} \gamma_s(\mathbf{o}_f) \end{aligned} \tag{4.18}$$

which is identical to the result obtained in [74] where a slightly different derivation was used. Thus, to determine the log likelihood of a cluster of states, only the covariance matrix $\Sigma(\mathbf{S})$ and the total state occupancy of the cluster $\sum_{s \in \mathbf{S}} \sum_{f \in \mathbf{F}} \gamma_s(\mathbf{o}_f)$ are required. Since $\gamma_s(\mathbf{o}_f) = 0$ for $s \in \mathbf{S}$ when $f \notin \mathbf{F}$, the set of frames $\mathbf{F}$ in the final line of equation (4.18) can be taken as all the frames in the training set instead of only the frames for which the observation vectors are generated by $\mathbf{S}$.[3]

Problem 3: Likelihood of a Combination of Clusters and Splitting Clusters

Assume that $\mathbf{S}_1$ and $\mathbf{S}_2$ are two distinct clusters ($\mathbf{S}_1 \cap \mathbf{S}_2 = \varnothing$) with all the states in each cluster tied and that the two clusters generate the observation vectors for frames $\mathbf{F}_1$ and $\mathbf{F}_2$,

[3] This is also the case for equation (4.14).

respectively. Then the log likelihood of the set of states $\mathbf{S} = \mathbf{S}_1 \cup \mathbf{S}_2$ generating the observation vectors for frames $\mathbf{F} = \mathbf{F}_1 \cup \mathbf{F}_2$ is

$$\begin{aligned} L(\mathbf{S}_1 \cup \mathbf{S}_2) &= \ln \prod_{f \in \mathbf{F}} p(\mathbf{o}_f | \mathbf{S}) \\ &= \ln \left[\prod_{f_1 \in \mathbf{F_1}} p(\mathbf{o}_{f_1} | \mathbf{S}_1) \prod_{f_2 \in \mathbf{F_2}} p(\mathbf{o}_{f_2} | \mathbf{S}_2) \right] \\ &= \ln \prod_{f_1 \in \mathbf{F_1}} p(\mathbf{o}_{f_1} | \mathbf{S}_1) + \ln \prod_{f_2 \in \mathbf{F_2}} p(\mathbf{o}_{f_2} | \mathbf{S}_2) \\ &= L(\mathbf{S}_1) + L(\mathbf{S}_2) \end{aligned} \tag{4.19}$$

Using equation (4.19), the optimal question for splitting a particular node can be determined. Suppose question q splits the node with states $\mathbf{S}$ into two nodes with states $\mathbf{S}_1(q)$ and $\mathbf{S}_2(q)$, respectively. From equation (4.19), the increase in log likelihood resulting from the split is then given by

$$\begin{aligned} \Delta L_q &= L(\mathbf{S}_1(q) \cup \mathbf{S}_2(q)) - L(\mathbf{S}) \\ &= L(\mathbf{S}_1(q)) + L(\mathbf{S}_2(q)) - L(\mathbf{S}) \end{aligned} \tag{4.20}$$

The question q^* which maximises equation (4.20) is selected as the optimal question to split the node. Only if ΔL_{q*} exceeds the minimum log likelihood improvement and the total state occupancy of both $\mathbf{S}_1(q^*)$ and $\mathbf{S}_2(q^*)$ exceed the minimum state occupancy, is the node actually split.

Summary of the Decision-Tree State Clustering Algorithm

The decision-tree state clustering algorithm is summarised by the following pseudocode:

```
pool states from all triphones with the same basephone in the root node
calculate Σ(S) and L(S) for the states in the root node using (4.9) and (4.18)
repeat {
    for each leaf node with states S {
        for each question q {
            determine the sets S_1(q) and S_2(q) into which states S are split
            calculate Σ(S_1(q)), Σ(S_2(q)), L(S_1(q)), L(S_2(q))
            calculate ΔL_q as in equation (4.20)
        }
        determine the question q* giving the maximum ΔL_q
        if ΔL_q* > minimum increase and state occupancy > minimum {
            split leaf node using question q*
            store Σ(S_1(q*)), Σ(S_2(q*)), L(S_1(q*)), L(S_2(q*)) for new leaf nodes
        }
    }
} until no leaf nodes are split
```

Initially the leaf node in line 4 is also the root node. The algorithm iteratively splits all leaf nodes until the stopping criteria is met.

4.2 Acoustic Modelling Approaches

In this research we apply three different variants of the decision-tree state clustering algorithm described in the preceding section. The three approaches are based on different ways of combining and/or separating training data between different accents. By applying these decision-tree strategies in the tied-state triphone HMM system training procedure, three distinct acoustic model sets are obtained. We apply these approaches to the five accents of SAE.

Similar acoustic modelling approaches have been considered by several other authors. As described in Section 2.4, Niesler [5] applied the same three approaches to multilingual acoustic modelling of four South African languages in a tied-state system. Schultz and Waibel applied similar approaches to acoustic modelling of 10 languages, but in a tied-mixture topology [48, 50]. Similarly, Caballero et al. considered tied-mixture systems based on the three approaches for multidialectel acoustic modelling of dialects of Spanish spoken in Spain and Latin America [51, 52]. Their approach was described in Section 2.5. Various other authors (Section 2.3.1) have also considered accent-specific and accent-independent acoustic modelling, the first two approaches considered in this research. The three acoustic modelling approaches are described below.

4.2.1 Accent-Specific Acoustic Modelling

As a first approach, the construction of accent-specific acoustic model sets where no sharing of data is allowed between accents is considered. Corresponding states from all triphones with the same basephone are clustered separately for each accent, resulting in separate decision-trees for each accent. The decision-tree clustering process employs only questions relating to phonetic context. The structure of the resulting acoustic models after performing decision-tree state clustering and mixing up the triphone HMMs is illustrated in Figure 4.2 for an AE and an EE 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 about accent is therefore considered more important than information about phonetic context. This approach

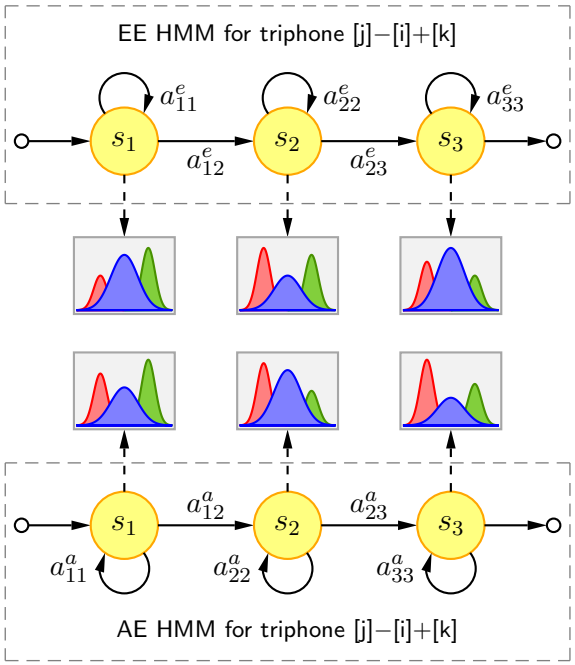


Figure 4.2: Accent-specific HMMs for corresponding AE and EE triphones.

corresponds to the ML-sep approach used by Schultz and Waibel (Section 2.4) and the baseline approach used by Caballero et al. (Section 2.5).

4.2.2 Accent-Independent Acoustic Modelling

For the second approach, a single accent-independent model set is obtained by pooling accent-specific data across accents for phones with the same IPA classification. A single set of decision-trees is constructed for all accents considered and employs only questions relating to phonetic context. Information regarding phonetic context is thus regarded as more important than information regarding accent. Figure 4.3 illustrates the resulting acoustic models when accent-independent modelling is applied to the AE and EE accents. Both triphone HMMs share the same Gaussian mixture observation PDFs as well as transition probabilities. If more accents are considered, for example the acoustic modelling of AE, BE and EE, the corresponding triphone HMMs from all three accents would share the same Gaussian mixture observation PDFs and transition probabilities. Such pooled models are often employed in multi-accent ASR (see Sections 2.3.1 and 2.6) and therefore represent an important baseline. This approach corresponds to the ML-mix approach used by Schultz and Waibel (Section 2.4) and the GPS-MR approach used by Caballero et al. (Section 2.5).

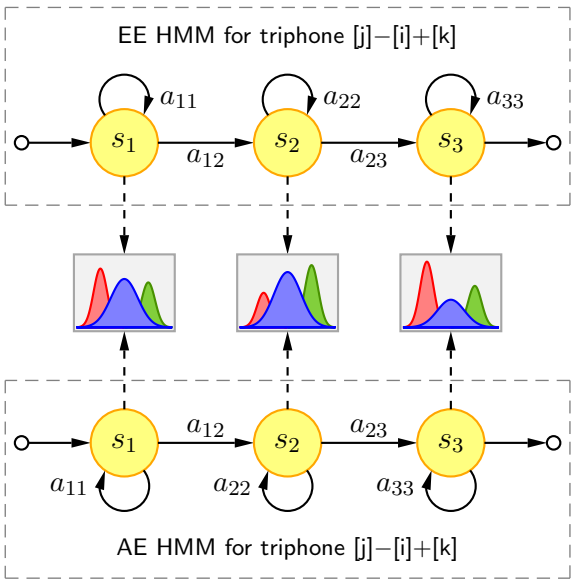


Figure 4.3: Accent-independent HMMs for corresponding AE and EE triphones.

4.2.3 Multi-Accent Acoustic Modelling

The third and final approach involves 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 contexts, but also the accent of the basephone. The HMM states of two or more triphones with the same basephone 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. This approach corresponds to the ML-tag approach

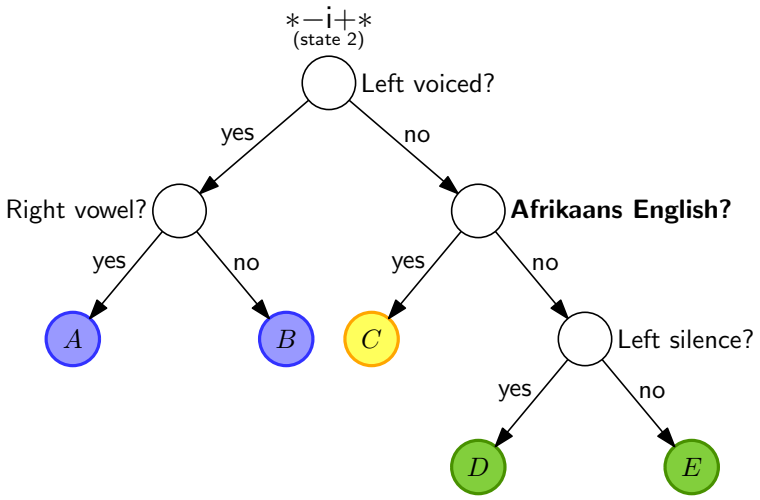


Figure 4.4: An example of a multi-accent decision-tree, taking into account the accent of the clustered triphone HMM states as well as the phonetic character of the left and right contexts.

used by Schultz and Waibel (Section 2.4) and the DCD-MR approach used by Caballero et al. (Section 2.5).

An example of a decision-tree constructed in this manner is given in Figure 4.4. Suppose that the tree clusters states from the AE and EE accents. Then the yellow leaf node C will indicate triphone HMM states that should be modelled separately for AE, the green leaf nodes D and E will indicate states that should be modelled separately for EE, and the blue leaf nodes A and B will indicate states that should be tied and modelled together for the two accents.

The structure of the multi-accent acoustic models resulting from this approach is illustrated in Figure 4.5, again for the AE and EE accents. Here the centre state of the triphone [j]-[i]+[k] is tied across the EE and AE accents while the first and last states are modelled separately. As for the accent-independent acoustic models, the transition probabilities of all triphones with the same basephone are tied across accents.

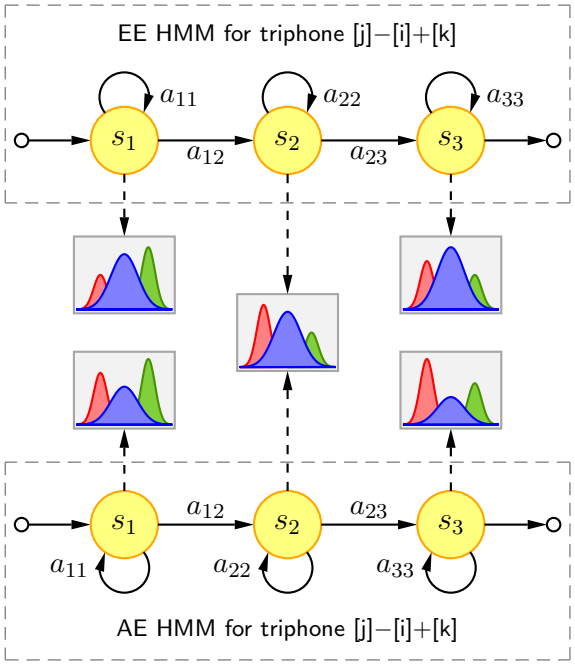


Figure 4.5: Multi-accent HMMs for corresponding AE and EE triphones.

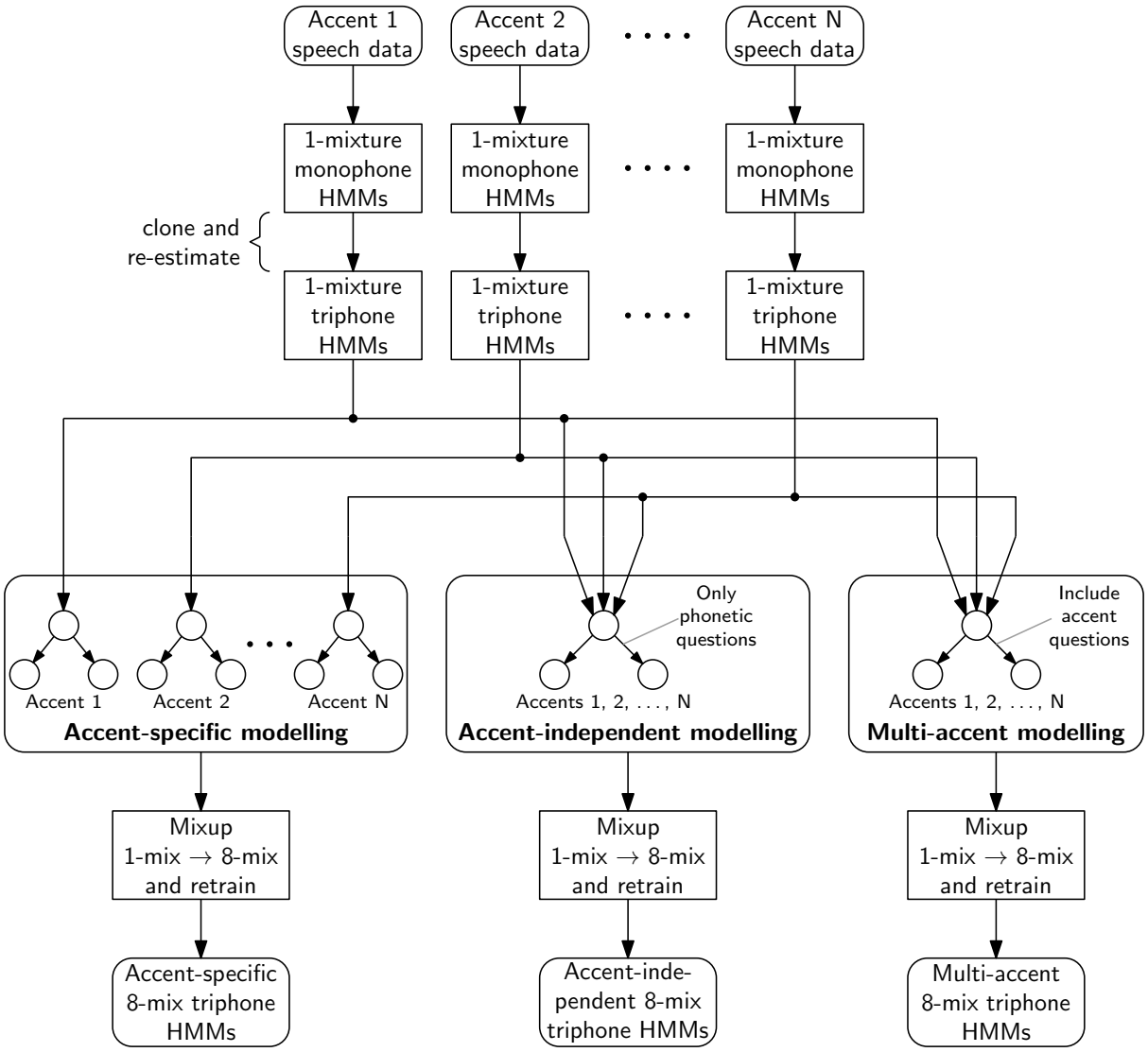


Figure 4.6: The complete acoustic model training process for the accent-specific, accent-independent and multi-accent acoustic modelling approaches applied to an arbitrary number of accents.

4.2.4 Model Training

The complete acoustic model training procedure involves applying each of the three decision-tree clustering strategies described above to the standard tied-state triphone HMM training methodology described in Section 4.1.2. This process is illustrated in Figure 4.6 for all three modelling approaches when applied to an arbitrary number of accents. For each accent, a set of accent-specific single-mixture monophone HMMs is trained. These are cloned and re-estimated to obtain single-mixture triphone HMMs which are then clustered using the three approaches. After clustering, the number of Gaussian mixtures is increased to finally yield accent-specific, accent-independent and multi-accent triphone HMM sets with multi-mixture (in our case eight-mixture) output PDFs.

4.3 Summary

In this chapter we described the purpose of decision-tree state clustering (Section 4.1.1) and how it is applied in the procedure for training tied-state triphone HMMs (Section 4.1.2). We explained the phonetic decision-tree construction process and derived the necessary equations from first principles in Section 4.1.3. Next, accent-specific, accent-independent and multi-accent acoustic modelling were described (Section 4.2). These form the foundation of the experimental investigations in this study. The three approaches are based on different strategies of decision-tree state clustering and in Section 4.2.4 we explained how the different strategies are applied in the training of tied-state multi-mixture triphone HMMs.

Chapter 5

Experimental Setup

Apart from performing experiments in which all five accents of SAE were modelled, several subsets of the five accents were also considered in order to investigate model behaviour for accents which differ in their degree of similarity. For each of the experiments that was carried out, the precise experimental setup was determined by the accent combination and the focus of that particular experiment. In this chapter we provide an overview of the common setup and system configuration followed for all the experiments conducted as part of this research. First, an overview is given of the language models which we employed. Next, pronunciation dictionaries are described. The chapter is concluded with general remarks regarding system configuration which include a description of the acoustic model training process, parameter optimisation, and the different recognition scenarios which we considered for recognition of the SAE accents.

5.1 Language Models

Speech recognition performance was assessed in terms of both phone and word recognition accuracies. For the phone recognition experiments, separate accent-specific phone backoff bigram language models (LMs) [76] were trained for each accent individually using the corresponding training set transcriptions and the SRILM toolkit [77]. For the word recognition experiments, the same tools were used to train accent-independent bigram LMs on the combined set of training transcriptions of all five accents in the AST databases (approximately 240k words, Table 3.2). Initial word recognition experiments described in Section C.2 had indicated that such accent-independent word LMs significantly outperformed accent-specific models trained individually on the training set transcriptions of each accent. In phone recognition experiments the opposite was observed, an effect that we ascribe to the larger size of the phone LM training set and the observation that, unlike the word sequences, the phone sequences are clearly accent-specific. Except for some experiments presented in Chapter 7, accent-specific LMs were therefore used for phone recognition experiments while accent-independent LMs trained on all the English AST data were used in word recognition experiments. Absolute discounting was used for the estimation of LM probabilities [78].

Phone and word LM perplexities are shown in Tables 5.1 and 5.2, respectively. The 'unigram types' and 'bigram types' columns refer to the number of unique phone or word unigrams and bigrams occurring in the training set transcriptions of each accent. The vocabulary of the accent-independent word LM was taken from the combined training set transcriptions of all five SAE accents. Out-of-vocabulary (OOV) rates are also indicated in Table 5.2. In the experiments in which only a subset of the five SAE accents was considered, the vocabulary was taken from the combined training set transcriptions of only the particular accents involved. However, LMs were still trained on the combined set of training transcriptions of all five accents in the AST

Table 5.1: Accent-specific phone bigram language model perplexities measured on the evaluation sets.

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

Table 5.2: Accent-independent word bigram language model perplexities and OOV rates measured on the evaluation sets.

Accent	Unigram types	Bigram types	Perplexity	OOV rate
AE	3700	11 580	24.07	1.82%
BE	2850	9639	27.87	2.84%
CE	3090	10 641	27.45	1.40%
EE	3413	10 451	24.90	1.08%
IE	3679	11 677	25.55	1.73%

databases. The perplexities of these LMs and the associated OOV rates are given in the sections in which the corresponding experiments are described.

5.2 Pronunciation Dictionaries

As part of the AST Project, five separate accent-specific English pronunciation dictionaries (PDs) were compiled by human annotators, corresponding to the five English-accented AST databases described in Chapter 3. 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. The AE, BE, CE, EE and IE accent-specific PDs respectively contain pronunciations for 3700, 2850, 3090, 3413 and 3679 words and on average 1.26, 1.33, 1.28, 1.09 and 1.16 pronunciations per word.

In order to obtain an indication of the similarity of the five accent-specific dictionaries, we considered the pair-wise phone alignment of corresponding pronunciations. For each pair of PDs, 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

Table 5.3: 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

other. If multiple pronunciations occurred for a particular word, the minimum distance was taken as the representative distance for that word. The average Levenshtein distance between each pair of PDs is given in Table 5.3. 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. The corresponding figure for EE and IE is just 0.60. The Levenshtein distance has also been used in other studies to estimate accent and dialect similarity, for example in [79].

Except for some experiments presented in Chapter 7, a single PD obtained by pooling the accent-specific PDs was used for all word recognition experiments. The PD obtained by pooling all five accent-specific PDs contains pronunciations for 7256 words and on average 2.10 pronunciations per word. The alternative is to use the accent-specific dictionaries, which differ in their vocabularies and lead to higher OOV rates on the test sets. We evaluated the use of accent-specific PDs in experiments presented in Section C.3 and found that the accent-independent PD yields better performance. In most of the experiments presented in this research we therefore restricted ourselves to the use of a single accent-independent PD. In experiments where accent-independent word LMs and PDs are used, any observed performance differences can be attributed to differences in the acoustic models, which are the focus of most of the experiments in this research. For experiments in which a subset of the SAE accents was considered, dictionaries corresponding to the vocabulary of the accent-independent LMs described above were needed. They were obtained by pooling the accent-specific PDs of only the particular accents involved.

5.3 System Configuration and Setup

5.3.1 Common Setup

Speech recognition systems were developed using the HTK tools [75]. The following setup was used for all systems irrespective of the accents considered, the acoustic modelling approach under evaluation, or the recognition configuration followed. Speech audio data were parameterised as 13 Mel-frequency cepstral coefficients (MFCCs) with their first- and second-order derivatives to obtain 39-dimensional feature vectors. Cepstral mean normalisation was applied on a per-utterance basis. The complete parameterised training set was used to obtain initial three-state left-to-right single-mixture monophone HMMs with diagonal covariance matrices by means of a flat start. These models were subsequently retrained individually for each accent using embedded Baum-Welch re-estimation. The resulting accent-specific monophone models were cloned and re-estimated to obtain initial accent-specific cross-word triphone models which were subsequently clustered using decision-tree state clustering. 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. This training methodology is the standard approach for training tied-state triphone HMMs and was discussed in Section 4.1.2.

5.3.2 The Three Acoustic Modelling Approaches

As described in Section 4.2, the distinction between the three acoustic modelling approaches considered in this research is based on different methods of decision-tree state clustering. The different modelling approaches were described in detail in Section 4.2. For each of these approaches the training procedure described above (Section 5.3.1) was followed and the appropri-

ate decision-tree state clustering strategy was employed, as illustrated in Figure 4.6. In this figure the three modelling approaches are applied to an arbitrary number of accents.

5.3.3 System Optimisation

Initial parameter optimisation on the development set indicated that recognition performance measured separately for each accent and each acoustic modelling approach is very robust towards 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, the WIP and LMS values were fixed across accents and acoustic modelling approaches for all experiments. For phone recognition, values of WIP $= -5$ and LMS $= 10$ were selected, while values of WIP $= -20$ and LMS $= 15$ were selected for word recognition experiments.

The initial development set optimisation did, however, indicate that recognition performance is sensitive to the number of independent parameters (which is proportional to the number of physical states) used by the acoustic model sets. 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 in an oracle recognition setup (see the next section). However, for a particular acoustic modelling approach, the same threshold value was used regardless of the accent. In some of the experiments that we present in Chapter 6, performance was measured on the evaluation set using the whole range of trained acoustic models in order to compare the relative performance of the different acoustic modelling approaches across different system sizes. For these cases, the performance of the systems optimised on the development set are indicated using circular markers (see for example Figure 6.1). The minimum cluster occupancy was set to 100 frames for all experiments, as was done in [75, p. 41].

5.3.4 Recognition Configurations

As mentioned in Sections 2.1.2 and 2.6, it is important to distinguish between the different recognition configurations which can be followed in order to evaluate a system designed for the simultaneous recognition of multiple accents. We consider two scenarios in this research. Oracle recognition involves presenting each test utterance to the accent-specific speech recogniser matching the accent of the utterance, as illustrated in Figure 2.4. In contrast, parallel recognition involves presenting each test utterance to a bank of accent-specific recognisers and selecting the output with the highest associated likelihood, as illustrated in Figure 2.3. Several speech recognition experiments are described in the following chapters and in each case we indicate which recognition configuration was followed. In Chapter 7 we present a set of experiments in which oracle and parallel recognition are compared in order to determine the effects of accent misclassifications which occur when performing parallel recognition.

5.4 Summary

In this chapter we described the language models (Section 5.1), the pronunciation dictionaries (Section 5.2), as well as the general system configuration and setup (Section 5.3) common to the experiments conducted as part of this research. Various combinations of the five SAE accents are considered in this study and experiments focussed on different aspects of modelling and speech recognition of the SAE accents. The aim of this chapter was to provide an overview

of the common experimental methodology employed. More details of precise system setup and configuration are provided as the experiments are presented in the following chapters.

Chapter 6

Acoustic Modelling Experiments

In this chapter we consider acoustic modelling for different combinations 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) accents. Four accent combinations are considered:

1. AE and EE
2. BE and EE
3. AE, BE and EE
4. AE, BE, CE, EE and IE

These are, respectively, referred to as AE+EE, BE+EE, AE+BE+EE and the five-accent combination. The similarity analysis presented in Section 3.5 indicated that AE+EE represents a pair of relatively similar accents, while the BE+EE combination represents a pair of relatively dissimilar accents. By considering these two accent pairs first, the behaviour of the three acoustic modelling approaches can be compared for accent combinations which differ clearly in their degree of similarity. The AE+BE+EE combination follows naturally from these two pair-wise modelling experiments and the five-accent combination, which includes all major South African English (SAE) accents, is very relevant from a practical system development perspective.

For each of the accent combinations, the three acoustic modelling approaches described in Section 4.2 are applied and subsequently evaluated by means of both phone and word recognition experiments. An oracle system configuration, in which each test utterance is only presented to the model set matching the accent of that utterance, is employed in all cases (see Sections 2.1.2, 2.6 and 5.3.4). By configuring the recognition setup in this way, the relative merits of the three acoustic modelling approaches can be compared in isolation; in particular, the effects of accent misclassifications on speech recognition performance are avoided. The fundamental aim of this chapter is to identify the acoustic modelling approach which takes best advantage of the available training data.

Multilingual acoustic modelling of four South African languages was considered in [5] (see Section 2.4). In that work experiments were also based on the AST databases, although in that case for different languages and not accents. In terms of phone recognition accuracies, modest improvements were achieved by multilingual acoustic models relative to language-specific and language-independent acoustic models. The language-independent models performed worst. While there are fundamental differences between the multilingual and multi-accent cases, similar databases were used and hence the results are comparable to some degree. In this chapter we will therefore often relate the results and trends we observe to those presented in [5].

Sections 6.2, 6.3, 6.5 and 6.6 describe experimental evaluations for the AE+EE, BE+EE, AE+BE+EE and five-accent combinations, respectively. Analogous experiments and the same experimental methodology are applied throughout, leading to some degree of repetition in the presented material.

6.1 System Optimisation and Evaluation

As described in Section 5.3.3, initial parameter optimisation on the development set indicated that recognition performance is sensitive to the number of independent parameters (which is proportional to the number of physical states) used by the acoustic model set. In order to compare the three acoustic modelling approaches thoroughly, and since the optimal size of an acoustic model set is not apparent beforehand, it was decided to train several sets of HMMs with different numbers of physical states. This was done by varying the likelihood improvement threshold used during decision-tree state clustering. Figure 6.1, for instance, shows performance curves for the three acoustic modelling approaches measured on the AE+EE evaluation sets using systems with sizes ranging from approximately 1000 to 10 000 physical states. By presenting evaluation results over a range of system sizes, the performance of systems employing different acoustic models with the same or similar number of independent parameters can be compared. In addition to evaluating the performance over the range of systems, we highlight the performance of the systems for which model set size was optimised on the development set. Evaluation set performance of these systems is indicated with circular markers in Figure 6.1, for instance.

6.2 Acoustic Modelling of AE and EE

The three acoustic modelling approaches described in Section 4.2 were applied to the combination of the AE and the EE training sets described in Section 3.3. In this section we present phone and word recognition results obtained in an oracle recognition configuration.

6.2.1 Language Models and Pronunciation Dictionaries

Separate accent-specific phone backoff bigram language models (LMs) were used in phone recognition experiments while accent-independent word backoff bigram LMs trained on the combined set of training transcriptions of all five accents were used in word recognition experiments (see Section 5.1). Perplexities for these LMs are shown in Table 6.1. The phone LMs are identical to those described in Section 5.1, while the AE+EE word LM differs from the five-accent word LM described in Section 5.1 only in its vocabulary. An accent-independent pronunciation dictionary (PD) was obtained by pooling the pronunciations from the AE and EE accent-specific PDs described in Section 5.2. The resulting dictionary contains pronunciations for 4933 words and on average 1.47 pronunciations per word. Out-of-vocabulary (OOV) rates are shown in Table 6.1.

Table 6.1: Phone and word bigram language model (LM) perplexities and OOV rates measured on the evaluation sets of the AE and EE accents.

Accent	Accent-specific phone LM		AE+EE accent-independent word LM		
	Bigram types	Perplexity	Bigram types	Perplexity	OOV rate (%)
AE	1891	14.40	11 580	23.48	3.06
EE	1542	12.64	10 451	23.85	2.12

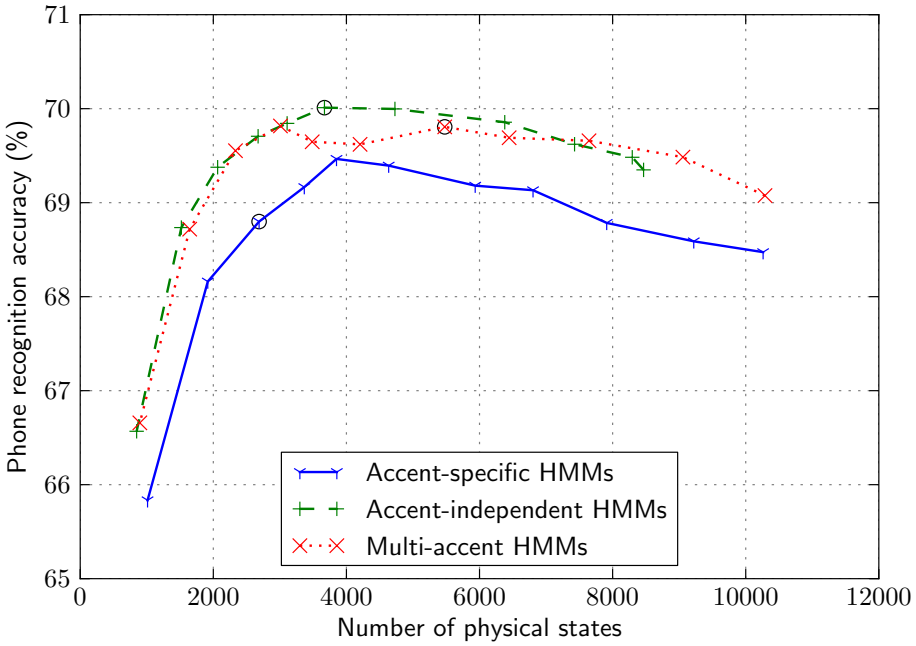


Figure 6.1: AE+EE average evaluation set phone accuracies of accent-specific, accent-independent and multi-accent systems as a function of the total number of distinct HMM states. Circular markers indicate the systems delivering optimal performance on the development set.

6.2.2 Phone Recognition Experiments

Figure 6.1 shows the average evaluation set phone recognition accuracy using eight-mixture triphone HMMs. These were obtained by application of the three acoustic modelling approaches. Recognition performance is shown for a range of model sizes and the systems leading to optimal performance on the development set are indicated by circular markers. For each acoustic modelling approach a single curve indicating the average accuracy for the two 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 set is therefore directly proportional to the total number of independent parameters. Hence, systems containing the same number of independent parameters are aligned vertically in Figure 6.1.

The results presented in Figure 6.1 show that multi-accent acoustic modelling and accent-independent modelling both yield consistently superior performance compared to accent-specific modelling over the range of models considered. Accent-independent modelling appears to yield slightly superior performance compared to multi-accent acoustic modelling. Table 6.2 summarises the evaluation set performance and the number of states for the systems optimised on the development set. Accent-independent modelling yields best performance followed by multi-accent modelling and then accent-specific modelling. Bootstrap confidence interval estimation [80] was used to calculate the statistical significance levels of the improvements for these systems relative to each other, as indicated in Table 6.6. The improvements shown by the accent-independent models over the multi-accent models are statistically significant only at the 76% level. In contrast, the improvements shown by both accent-independent and multi-accent models over accent-specific models are statistically significant at the 99.9% level.

Since the triphone clustering process optimises the overall training set likelihood, improvements for each individual accent are not guaranteed when multi-accent acoustic modelling is per-

Table 6.2: The number of states and evaluation set phone accuracies for the AE+EE systems delivering optimal performance on the development set and identified by the circular markers in Figure 6.1.

Model set	No. of states	Accuracy (%)
Accent-specific	2688	68.80
Accent-independent	3670	70.01
Multi-accent	5477	69.81

Table 6.3: Evaluation set phone accuracies (%) for each accent for the AE+EE systems delivering optimal performance on the development set and identified by the circular markers in Figure 6.1.

Model set	AE	EE	Average
Accent-specific	64.48	72.89	68.80
Accent-independent	65.81	73.99	70.01
Multi-accent	66.15	73.27	69.81

formed. The per-accent phone accuracies for the systems optimised on the development set are listed in Table 6.3. These results indicate that multi-accent acoustic modelling improves the phone recognition accuracy relative to the remaining two approaches for AE, while accent-independent acoustic modelling leads to the best performance for EE. The relative per-accent performance is consistent with the perplexities given in Table 6.1: all three systems show superior performance for the EE accent, which has the lower phone perplexity. For both accents the accent-independent system outperforms the accent-specific system.

6.2.3 Word Recognition Experiments

Figure 6.2 shows the average evaluation set word recognition accuracy using eight-mixture triphone HMMs. These were obtained by application of the three acoustic modelling approaches. Once again, performance is shown for a range of acoustic models, with the systems leading to optimal performance on the development set identified by circular markers. For each acoustic modelling approach a single curve indicating the average accuracy between the two accents is shown. The number of physical states is calculated as described in the previous section.

Figure 6.2 indicates that, over the range of models considered, accent-specific modelling performs worst while accent-independent and multi-accent modelling yield similar performance. The evaluation set performance of the systems optimised on the development set is summarised in Table 6.4. The statistical significance levels associated with the relative improvements between these systems are given in Table 6.6. It is evident that the accent-specific modelling approach is outperformed by both accent-independent and multi-accent acoustic modelling. The improvements of the accent-independent and multi-accent systems over the accent-specific system are statistically significant at the 96% and 97% levels, respectively. Both Tables 6.4 and 6.6 indicate that there is very little to distinguish between the accent-independent and multi-accent modelling approaches and that the difference in performance is not statistically significant.

Table 6.5 presents the word accuracies separately for each accent for the systems optimised on the development set. As in the phone recognition experiments (Table 6.3), multi-accent acoustic modelling outperforms the remaining two approaches for AE, while accent-independent acoustic modelling yields best performance for EE. For both accents the accent-independent system outperforms the accent-specific system. Again, the relative per-accent performance is consistent with the associated perplexities: in all three cases performance is slightly better for the AE accent, which has the lower word perplexity (Table 6.1). It is interesting that word

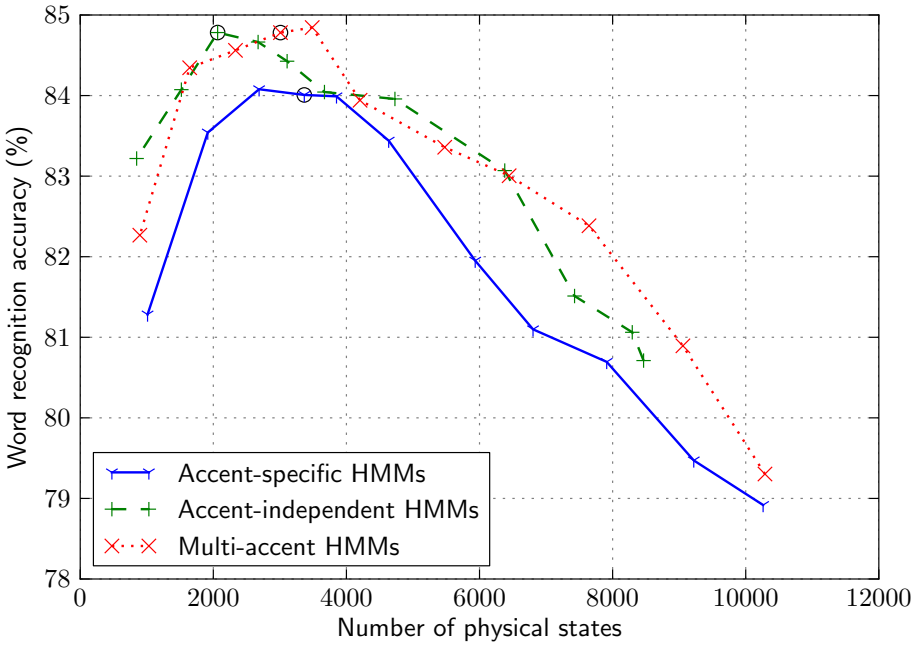


Figure 6.2: AE+EE average evaluation set word accuracies of accent-specific, accent-independent and multi-accent systems as a function of the total number of distinct HMM states. Circular markers indicate the systems delivering optimal performance on the development set.

Table 6.4: The number of states and evaluation set word accuracies for the AE+EE systems delivering optimal performance on the development set and identified by the circular markers in Figure 6.2.

Model set	No. of states	Accuracy (%)
Accent-specific	3367	84.01
Accent-independent	2065	84.78
Multi-accent	3009	84.78

Table 6.5: Evaluation set word accuracies (%) for each accent for the AE+EE systems delivering optimal performance on the development set and identified by the circular markers in Figure 6.2.

Model set	AE	EE	Average
Accent-specific	84.52	83.52	84.01
Accent-independent	84.90	84.67	84.78
Multi-accent	85.14	84.44	84.78

Table 6.6: Statistical significance levels (%) of improvements for pair-wise comparisons of the AE+EE systems delivering optimal performance on the development set and described in Tables 6.2 and 6.4, calculated by using bootstrap confidence interval estimation [80].

Comparison	Phone recognition	Word recognition
Accent-specific vs. accent-independent	99.9	96
Accent-specific vs. multi-accent	99.9	97
Multi-accent vs. accent-independent	76	50

recognition indicates slightly better performance for the second language AE compared to the first language EE, which is in contrast to the phone recognition accuracies given in Table 6.3.

6.2.4 Analysis of the 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 6.3 analyses the decision-trees of the multi-accent system delivering optimal word accuracy on the development set (3009 states, Table 6.4). The figure shows that approximately 22% of all questions at the root nodes are accent-based and that this proportion drops to 16% and 11% for the roots' children and grandchildren, respectively. This is in contrast to the multilingual case presented in [5] where the percentage of language-based questions dropped from more than 45% at the root nodes to less than 5% at the 10^{th} level of depth. This indicates that, compared to the multilingual case, separation of the accents close to the root node is performed far less often for the AE and EE accents. This further emphasises that pooling seems to be the best approach to follow for this accent pair.

The relatively small influence of the accent-based questions 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 clustering process. Figure 6.4 illustrates this improvement as a function of the 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 within the decision-trees. In Figure 6.4, approximately 22% of the total increase at the root nodes is afforded by accent-based questions. This proportion is much lower than the corresponding figure of 74% for multilingual systems [5]. Figures 6.3 and 6.4 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 development set phone accuracy as well as for the multi-accent system with the largest number of parameters, revealed similar trends.

In summary, the above analysis indicates that early partitioning of models into accent-based groups is not necessarily performed or advantageous for the two accents considered. This concurs with the phone and word recognition results presented in Sections 6.2.2 and 6.2.3, respectively, which indicated that accent-independent modelling is superior to accent-specific modelling. In contrast, it was found that for the multilingual case presented in [5], language-specific modelling outperformed language-independent modelling and early partitioning of models into language-based groups was observed.

6.2.5 Analysis of Cross-Accent Data Sharing

In order to determine to what extent data sharing ultimately takes place between AE and EE, we considered the proportion of decision-tree leaf nodes (which correspond to the state clusters) that are populated by states from both accents. A 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 6.5 illustrates how these proportions change as a function of the total number of clustered states in the system.

From Figure 6.5 it is apparent that, as the number of clustered states is increased, the proportion of clusters combining data from 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 one might expect. It is interesting to note that, although our findings suggest that multi-accent and accent-independent systems give similar performance and that

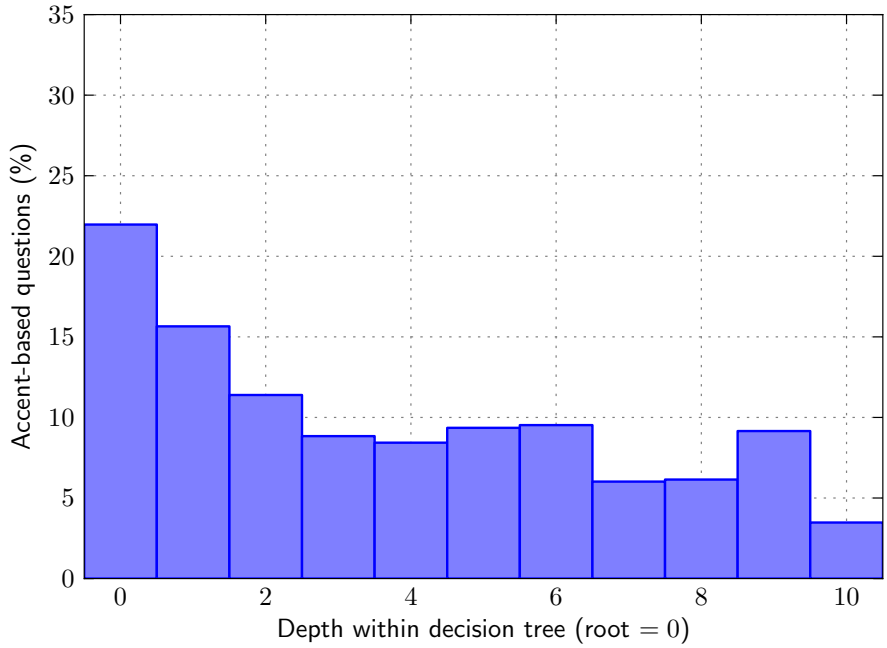


Figure 6.3: Analysis showing the percentage of questions that are accent-based at various depths within the decision-trees for the AE+EE multi-accent system with optimal development set word accuracy.

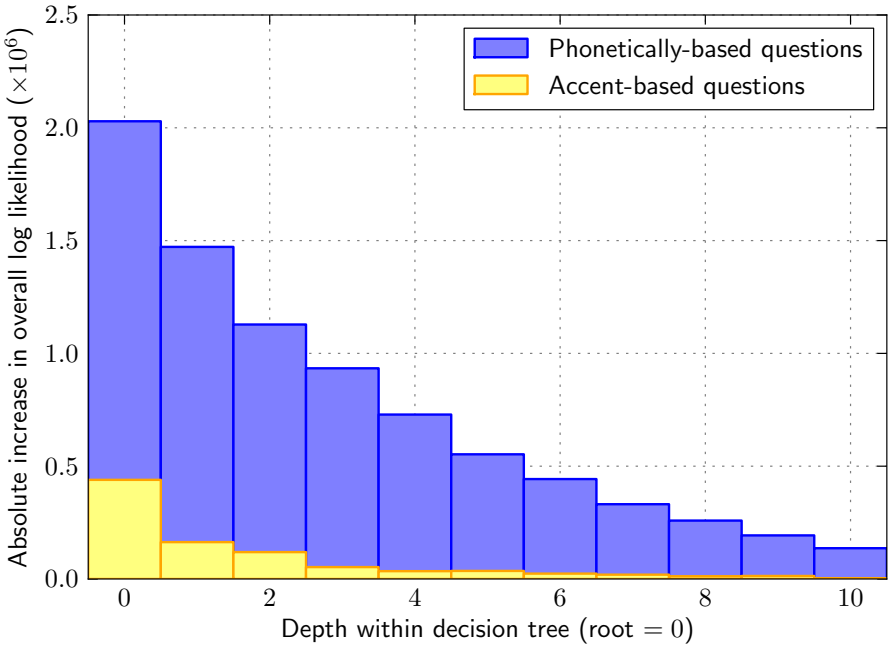


Figure 6.4: Analysis showing the contribution made to the increase in overall log likelihood by the accent-based and phonetically-based questions, respectively, within the decision-trees for the AE+EE multi-accent system with optimal development set word accuracy.

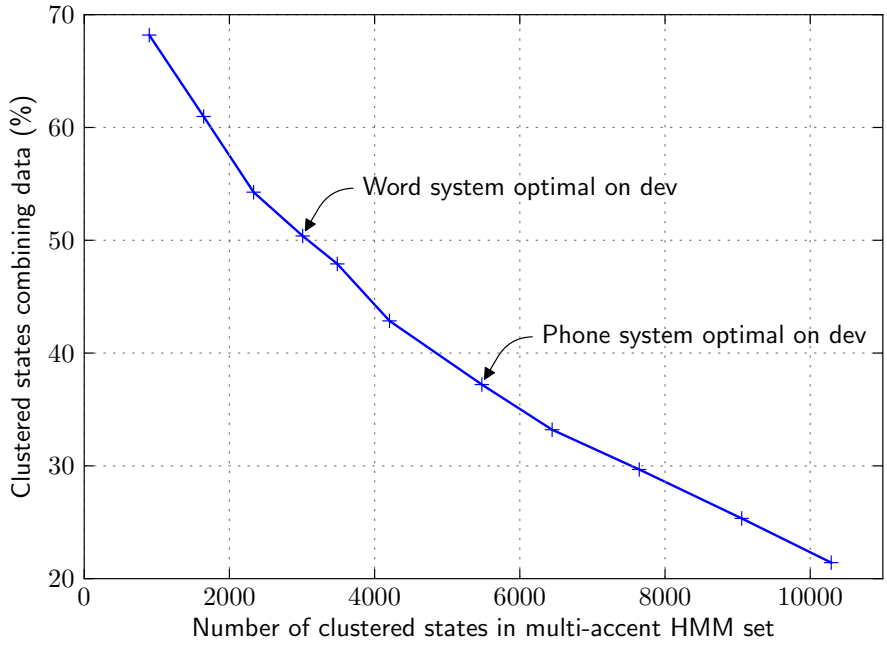


Figure 6.5: Proportion of state clusters combining data from both the AE and EE accents.

accent-specific modelling performs worst, the multi-accent system with optimal development set word accuracy (3009 states) models only approximately 50% of state clusters together for both accents. For the multi-accent system with optimal development set phone accuracy (5477 states) this figure is approximately 37%. Thus, although accent-independent modelling is advantageous 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 [5], only 20% of state clusters contained more than one language, emphasising that the multi-accent case is much more prone to sharing.

6.3 Acoustic Modelling of BE and EE

The three acoustic modelling approaches described in Section 4.2 were applied to the combination of the BE and the EE training sets described in Section 3.3. In this section we present phone and word recognition results obtained in an oracle recognition configuration.

6.3.1 Language Models and Pronunciation Dictionaries

Separate accent-specific phone backoff bigram LMs were used in phone recognition experiments while accent-independent word backoff bigram LMs trained on the combined set of training transcriptions of all five accents were used in word recognition experiments. Perplexities for these LMs are shown in Table 6.7. The phone LMs are identical to those described in Section 5.1, while the BE+EE word LM differs from the five-accent word LM described in Section 5.1 only in its vocabulary. An accent-independent PD was obtained by pooling pronunciations from the BE and EE accent-specific PDs described in Section 5.2. The resulting dictionary contains pronunciations for 4576 words and on average 1.57 pronunciations per word. OOV rates are shown in Table 6.7.

Table 6.7: Phone and word bigram language model (LM) perplexities and OOV rates measured on the evaluation sets of the BE and EE accents.

Accent	Accent-specific phone LM		BE+EE accent-independent word LM		
	Bigram types	Perplexity	Bigram types	Perplexity	OOV rate (%)
BE	1761	15.44	9639	26.74	3.87
EE	1542	12.64	10 451	24.04	3.04

6.3.2 Phone Recognition Experiments

Figure 6.6 shows the average evaluation set phone recognition using eight-mixture triphone HMMs. The results indicate that, over the range of models considered, accent-independent modelling performs worst. Although the multi-accent systems clearly outperform the accent-specific systems at some system sizes, accent-specific and multi-accent acoustic modelling yield similar performance over the range of models considered without one approach being clearly superior to the other. The evaluation set recognition accuracy for systems optimised on the development set is summarised in Table 6.8. The accent-specific and multi-accent systems yield similar results and both outperform the accent-independent system. These improvements have been found to be statistically significant at the 99.9% level, as indicated in Table 6.12. The significance levels in Table 6.12 also indicate that the improvement of the multi-accent system over the accent-specific system is not significant.

The per-accent phone accuracies for the systems optimised on the development set are presented in Table 6.9. These results indicate that multi-accent acoustic modelling improves the phone recognition accuracy relative to the other two approaches for BE, while accent-specific acoustic modelling leads to the best performance for EE. The relative per-accent performance is also consistent with the perplexities given in Table 6.7: all three systems show superior performance for the EE accent, which has the lower phone perplexity. For both accents the accent-specific

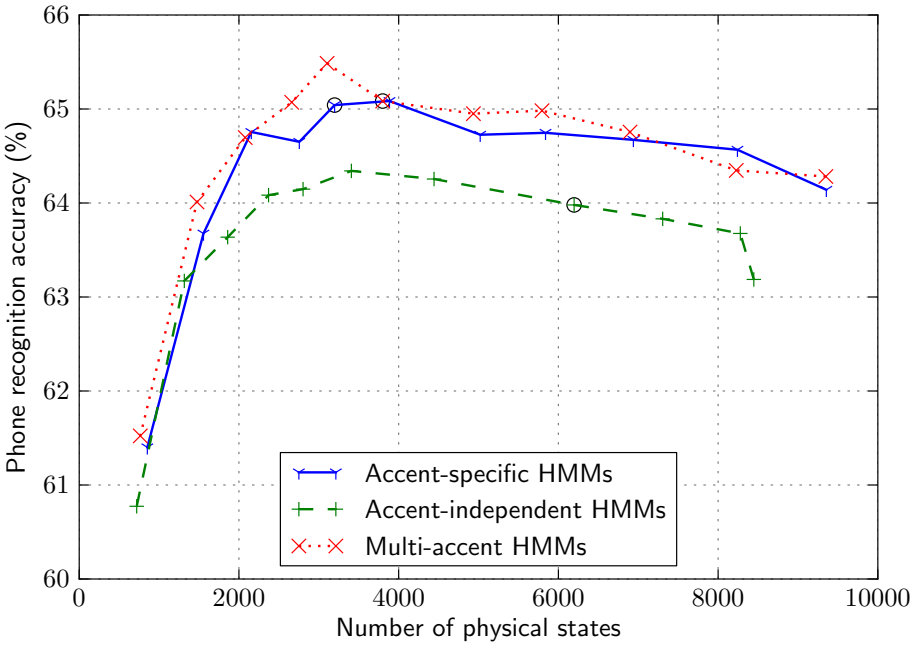


Figure 6.6: BE+EE average evaluation set phone accuracies of accent-specific, accent-independent and multi-accent systems as a function of the total number of distinct HMM states. Circular markers indicate the systems delivering optimal performance on the development set.

Table 6.8: The number of states and evaluation set phone accuracies for the BE+EE systems delivering optimal performance on the development set and identified by the circular markers in Figure 6.6.

Model set	No. of states	Accuracy (%)
Accent-specific	3198	65.04
Accent-independent	6197	63.98
Multi-accent	3800	65.08

Table 6.9: Evaluation set phone accuracies (%) for each accent for the BE+EE systems delivering optimal performance on the development set and identified by the circular markers in Figure 6.6.

Model set	BE	EE	Average
Accent-specific	56.64	73.38	65.04
Accent-independent	56.10	71.80	63.98
Multi-accent	57.23	72.88	65.08

system outperforms the accent-independent system.

6.3.3 Word Recognition Experiments

Figure 6.7 shows the average evaluation set word recognition accuracy using eight-mixture triphone HMMs. The results indicate that, over the range of models considered, accent-independent modelling performs worst. Multi-accent modelling leads to consistently improved performance relative to accent-specific modelling except for one system (3198 states). The evaluation set performance for the systems optimised on the development set is summarised in Table 6.10. Multi-accent modelling yields best performance followed by accent-specific modelling and then accent-independent modelling. As indicated in Table 6.12, the improvements of the multi-accent system compared to the accent-specific and accent-independent systems were found to be statistically significant at the 93% and 99.9% levels, respectively.

Table 6.11 presents the word accuracies separately for each accent for the systems optimised on the development set. For both accents, multi-accent acoustic models yield the best performance followed by accent-specific and then accent-independent modelling. As in the phone recognition experiments, the relative per-accent performance is consistent with the perplexities given in Table 6.7: all three systems perform better on the EE accent which has the lower word perplexity.

While the results for the AE+EE accent pair in Section 6.2 suggest that simply pooling training data across accents is advantageous, the phone and word recognition results for the BE+EE combination presented in this section clearly indicate that this is not always the case. Although pooling leads to more training data for a single model set, it is clear that the comparative merits of the two approaches (accent-specific and accent-independent modelling) depend on both the abundance of training data as well as the degree of similarity between accents.

6.3.4 Analysis of the Decision-Trees

Figure 6.8 analyses the decision-trees of the multi-accent system delivering optimal word accuracy on the development set (2661 states, Table 6.10). Figure 6.8 indicates that 31% of all questions at the root nodes are accent-based and that this proportion drops to approximately 20% and 14% for the roots' children and grandchildren, respectively. This is in contrast to the AE+EE case for which the same analysis is shown in Figure 6.3. In that figure, approximately

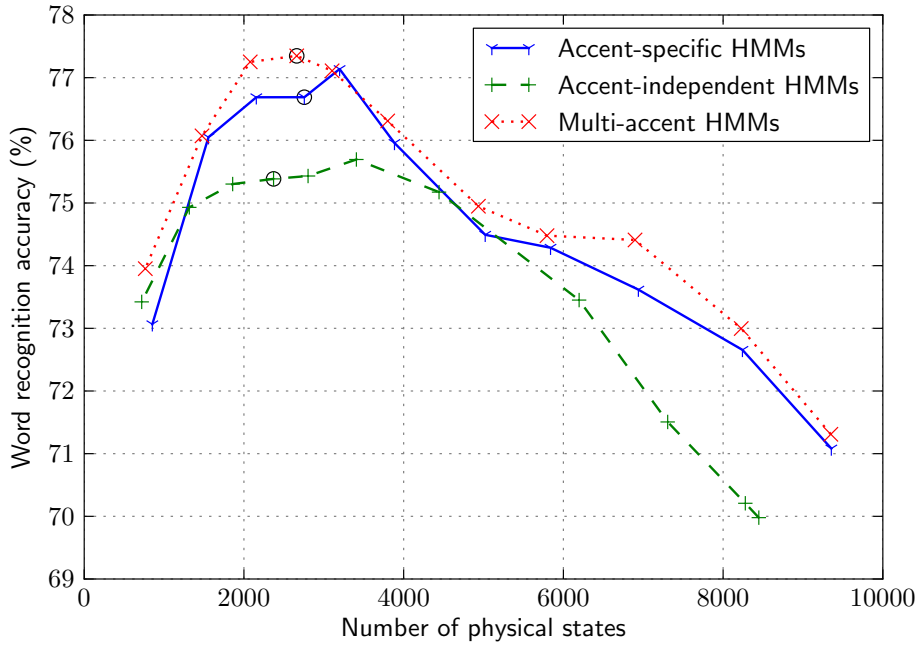


Figure 6.7: BE+EE average evaluation set word accuracies of accent-specific, accent-independent and multi-accent systems as a function of the total number of distinct HMM states. Circular markers indicate the systems delivering optimal performance on the development set.

Table 6.10: The number of states and evaluation set word accuracies for the BE+EE systems delivering optimal performance on the development set and identified by the circular markers in Figure 6.7.

Model set	No. of states	Accuracy (%)
Accent-specific	2756	76.69
Accent-independent	2371	75.38
Multi-accent	2661	77.35

Table 6.11: Evaluation set word accuracies (%) for each accent for the BE+EE systems delivering optimal performance on the development set and identified by the circular markers in Figure 6.7.

Model set	BE	EE	Average
Accent-specific	72.65	80.78	76.69
Accent-independent	70.87	79.96	75.38
Multi-accent	73.29	81.46	77.35

Table 6.12: Statistical significance levels (%) of improvements for pair-wise comparisons of the BE+EE systems delivering optimal performance on the development set and described in Tables 6.8 and 6.10, calculated by using bootstrap confidence interval estimation [80].

Comparison	Phone recognition	Word recognition
Accent-independent vs. accent-specific	99.9	99
Accent-specific vs. multi-accent	56	93
Accent-independent vs. multi-accent	99.9	99.9

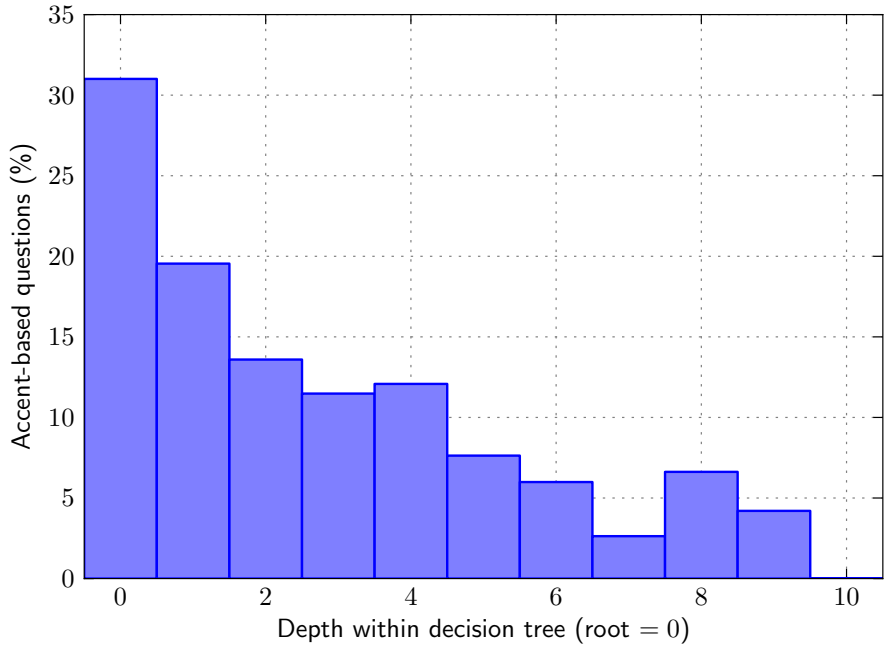


Figure 6.8: Analysis showing the percentage of questions that are accent-based at various depths within the decision-trees for the BE+EE multi-accent system with optimal development set word accuracy.

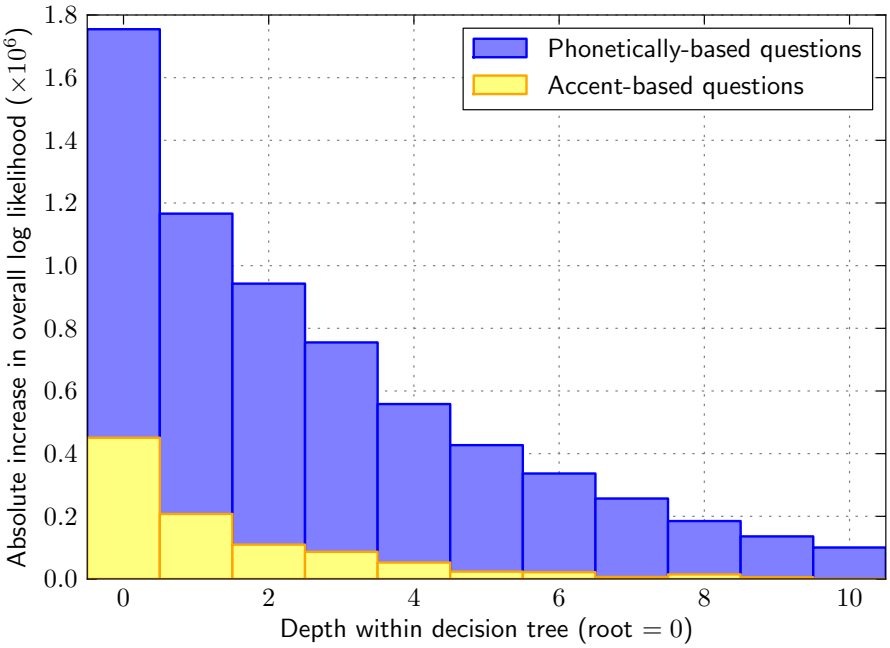


Figure 6.9: Analysis showing the contribution made to the increase in overall log likelihood by the accent-based and phonetically-based questions, respectively, within the decision-trees for the BE+EE multi-accent system with optimal development set word accuracy.

22% of all root node questions are accent-based and this proportion drops to 16% and 11% for the roots' children and grandchildren, respectively. It is therefore evident that the accent-based questions are asked less often and are more evenly distributed through the different depths of the AE+EE decision-trees when compared with the BE+EE decision-trees. This indicates that early partitioning of models into accent-based groups is more advantageous for BE+EE while separation of models is performed less often for AE+EE. This is also reflected by the BE+EE phone and word recognition results in which accent-specific modelling consistently outperforms accent-independent modelling. In contrast, the AE+EE experiments indicated that the accent-independent models were superior to the accent-specific models. The BE+EE combination is closer to the multilingual case where the percentage of language-based questions dropped from more than 45% at the root nodes to less than 5% at the 10^{th} level of depth [5].

The contribution to the log likelihood improvement made by the accent-based and phonetically-based questions, respectively, during the decision-tree growing process of the optimal word recognition multi-accent system are shown in Figure 6.9 as a function of the depth within the decision-trees. As for AE+EE in Figure 6.4, phonetically-based questions make a much larger contribution to the log likelihood improvement than the accent-based questions throughout the various depths of the decision-trees. The contribution to the likelihood increase is, however, slightly higher for the BE+EE combination. The analyses in Figures 6.4 and 6.9 are both in contrast to the multilingual case, where approximately 74% of the total log likelihood improvement is due to language-based questions at the root nodes and phonetically-based questions make up the greatest contribution thereafter [5]. Thus, even for BE+EE, where separation of data is preferred to combining data, early separation of models into accent-based groups is not performed to the same degree as in multilingual models. Figures 6.8 and 6.9 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 development set phone accuracy as well as for the multi-accent system with the largest number of parameters, revealed similar trends.

6.3.5 Analysis of Cross-Accent Data Sharing

In order to determine to what extent data sharing ultimately takes place between BE and EE, we considered the proportion of decision-tree leaf nodes (which correspond to the state clusters) that are populated by states from both accents. Figure 6.10 illustrates how these proportions change as a function of the total number of clustered states in the system. As for AE+EE, it is apparent that the proportion of clusters consisting of both accents decreases as the number of clustered states is increased. 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. It is interesting to note that, although our findings suggest that multi-accent and accent-specific systems give comparable results, the multi-accent system delivering optimal development set word accuracy (2661 states) models approximately 33% of state clusters together for both accents. This figure is approximately 27% for the multi-accent system delivering optimal development set phone accuracy (3800 states). For the optimal multilingual system in [5], only 20% of state clusters contained more than one language, again emphasising that the multi-accent case is more prone to sharing even when accents are relatively different.

When Figure 6.10 for BE+EE and Figure 6.5 for AE+EE are compared, it is evident that BE+EE is much more inclined to separate the triphone HMM states of the two accents. This agrees with the phone and word recognition results which indicated that accent-specific modelling was superior for BE+EE while accent accent-independent modelling was found to be superior for the AE+EE accent pair.

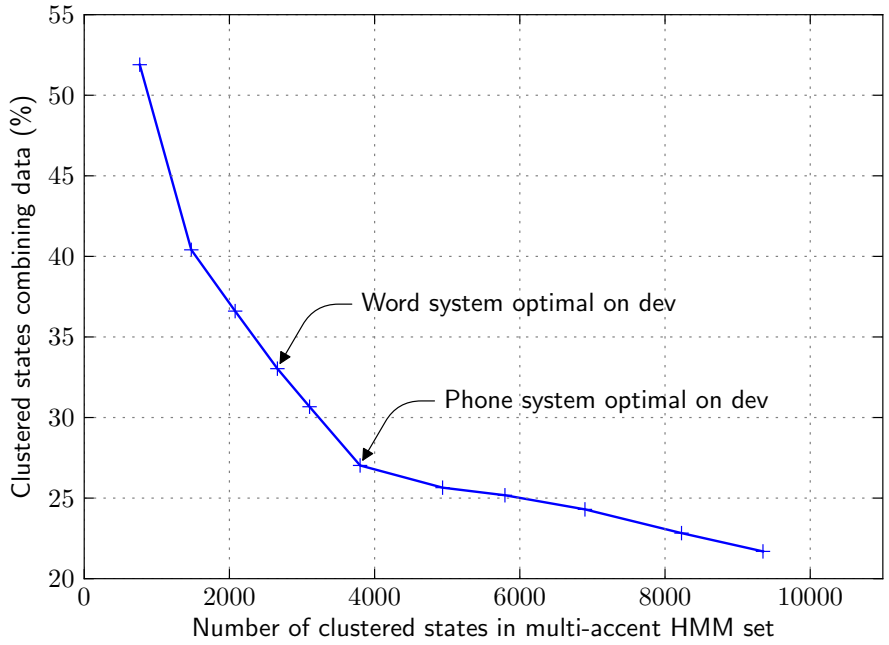


Figure 6.10: Proportion of state clusters combining data from both the BE and EE accents.

6.4 Summary and Comparison of AE+EE and BE+EE Modelling

The phone and word recognition experiments presented in Section 6.2 indicated that accent-specific modelling was outperformed by accent-independent acoustic modelling for the AE+EE accent pair. Systems employing multi-accent acoustic models achieved slightly poorer performance compared to accent-independent modelling but consistently outperformed accent-specific systems. In contrast, the phone and word recognition experiments for the BE+EE accent pair presented in Section 6.3 indicated that accent-specific modelling is superior to accent-independent acoustic modelling. For this accent pair, multi-accent acoustic models showed similar or improved performance compared to accent-specific modelling. Based on these results, we speculate that the differences in the comparative performance of the accent-specific and accent-independent modelling approaches can be attributed to the similarity of the accents involved. This is supported by the literature (see Section 2.3.1) which indicates that the comparative merits of these two approaches appear to depend on the amount of training data, the type of task, as well as the degree of similarity between the accents involved.

As part of the similarity analysis presented in Section 3.5, the average pair-wise Bhattacharyya distances between the AE, BE and EE accents were calculated. These distances are shown to scale in Figure 6.11. To lend further support to our observations above, the figure illustrates that AE+EE, which favours accent-independent modelling, represents two relatively similar accents, while BE+EE, which favours accent-specific modelling, represents a pair of relatively dissimilar accents. In general, the English proficiency of AE speakers is high, which may contribute to the similarity of the AE and EE accents and thus explain why accent-independent modelling is advantageous [81]. Furthermore, as mentioned in Section 3.2.1, Bowerman [56, 57] also notes the influence of Afrikaans on the development of White South African English.

Analyses indicated that although multi-accent modelling of AE+EE is more prone to sharing data, a significant number of clusters are still modelled separately for each accent. Figure 6.5 shows that 50% of the state clusters are separated for the optimal AE+EE word recognition multi-accent system (3009 states), without impairing performance compared to accent-

independent modelling (Table 6.4). For the BE+EE accent pair, on the other hand, multi-accent modelling is more inclined to separate modelling. However, many clusters are still modelled together for both accents: 33% of the clusters for the optimal word recognition system employing 2661 states (Figure 6.10), again without impairing performance (Table 6.10). Based on the above observations, we speculate that multi-accent acoustic modelling will yield improvements over both accent-specific and accent-independent modelling when applied to more than two accents. This hypothesis is tested experimentally in the next section.

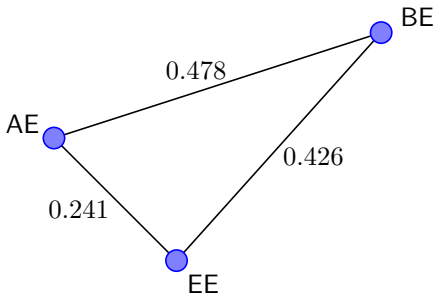


Figure 6.11: Average Bhattacharyya distances between Afrikaans English (AE), Black South African English (BE) and White South African English (EE).

6.5 Acoustic Modelling of AE, BE and EE

The three acoustic modelling approaches described in Section 4.2 were applied to the combination of the AE, BE and EE training sets described in Section 3.3. In this section we present phone and word recognition results obtained in an oracle recognition configuration.

6.5.1 Language Models and Pronunciation Dictionaries

Separate accent-specific phone backoff bigram LMs were used in phone recognition experiments while accent-independent word backoff bigram LMs trained on the combined set of training transcriptions of all five accents were used in word recognition experiments. Perplexities for these LMs are shown in Table 6.13. The phone LMs are identical to those described in Section 5.1; they have also been used in the AE+EE and the BE+EE experiments described in Sections 6.2 and 6.3, respectively. The AE+BE+EE word LM and the five-accent word LM described in Section 5.1 differ only in their vocabularies. An accent-independent PD was obtained by pooling pronunciations from the AE, BE and EE accent-specific PDs described in Section 5.2. The resulting dictionary contains pronunciations for 5753 words and on average 1.83 pronunciations per word. OOV rates are shown in Table 6.13.

Table 6.13: Phone and word bigram language model (LM) perplexities and OOV rates measured on the evaluation sets of the AE, BE and EE accents.

Accent	Accent-specific phone LMs		AE+BE+EE accent-independent word LM		
	Bigram types	Perplexity	Bigram types	Perplexity	OOV rate (%)
AE	1891	14.40	11 580	24.21	2.51
BE	1761	15.44	9639	27.69	3.16
EE	1542	12.64	10 451	24.65	1.73

6.5.2 Phone Recognition Experiments

Figure 6.12 shows the average evaluation set phone recognition accuracy using eight-mixture triphone HMMs. The results indicate that accent-specific modelling appears to outperform accent-independent modelling for most system sizes. Multi-accent acoustic modelling, however, consistently outperforms the other two approaches. The evaluation set performance of the systems optimised on the development set is summarised in Table 6.14. The improvement of the multi-accent system relative to the accent-specific system is statistically significant only at the 72% level, while the improvement over the accent-independent system is significant at the 99% level, as indicated in Table 6.18.

Table 6.15 shows the phone recognition performance measured separately on the evaluation set

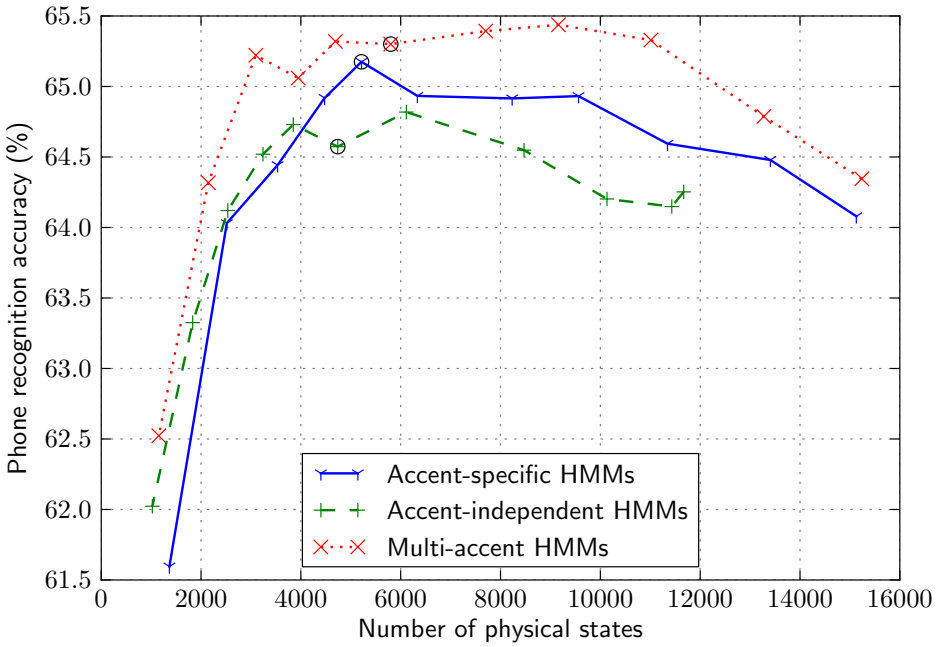


Figure 6.12: AE+BE+EE average evaluation set phone accuracies of accent-specific, accent-independent and multi-accent systems as a function of the total number of distinct HMM states. Circular markers indicate the systems delivering optimal performance on the development set.

Table 6.14: The number of states and evaluation set phone accuracies for the AE+BE+EE systems delivering optimal performance on the development set and identified by the circular markers in Figure 6.12.

Model set	No. of states	Accuracy (%)
Accent-specific	5215	65.17
Accent-independent	4739	64.57
Multi-accent	5801	65.30

Table 6.15: Evaluation set phone accuracies (%) for each accent for the AE+BE+EE systems delivering optimal performance on the development set and identified by the circular markers in Figure 6.12.

Model set	AE	BE	EE	Average
Accent-specific	65.74	56.57	73.18	65.17
Accent-independent	65.35	55.14	73.20	64.82
Multi-accent	65.92	56.72	73.23	65.44

of each accent for the systems optimised on the development set. For AE and BE, accent-specific modelling outperforms accent-independent modelling. For EE, the two approaches yield very similar results. Multi-accent acoustic modelling yields the best performance for all three accents. The results in Table 6.15 are consistent with the perplexities presented in Table 6.13: best performance is achieved on the EE evaluation set which has the lowest phone perplexity, while the lowest accuracies are achieved on BE which has the highest perplexity.

6.5.3 Word Recognition Experiments

Figure 6.13 shows the average evaluation set word recognition accuracy using eight-mixture triphone HMMs. The results indicate that, over the range of models considered, multi-accent modelling consistently outperforms both accent-specific and accent-independent acoustic modelling. The evaluation set performance for the systems optimised on the development set is summarised in Table 6.16. Table 6.18 shows that the improvements of the multi-accent system compared to the accent-specific and accent-independent systems were calculated to be statistically significant at the 99% and 99.9% levels, respectively.

Table 6.17 presents the word accuracies separately for each accent for the systems optimised on the development set. As with the phone recognition results, the results in this table correspond to the word LM perplexities given in Table 6.13. Except for the accent-specific system, best performance is achieved for AE, which also has the lowest perplexity, while performance is poorest on BE, which has the highest perplexity. Although improvements for the individual accents are not guaranteed by the multi-accent acoustic modelling approach, multi-accent modelling does yield the best per-accent performance for all three accents.

In general, similar trends are observed in both the phone and word recognition experiments. Results indicate that multi-accent acoustic modelling is superior when compared to both accent-specific as well as accent-independent modelling for the three accents considered. Furthermore, accent-specific models usually outperform accent-independent models, although this behaviour is not consistent over the whole range of models considered. In [5] a similar approach to multilingual acoustic modelling was also shown to outperform language-specific and language-independent modelling.

6.5.4 Analysis of the Decision-Trees

Figure 6.14 analyses the decision-trees of the multi-accent system delivering optimal word accuracy on the development set (3948 states, Table 6.16). The figure shows that approximately 45% of all questions at the root nodes are accent-based and that this proportion drops to 25% and 18% for the roots' children and grandchildren, respectively. The same analysis for the AE+EE combination, where accent-independent modelling was shown to be superior to accent-specific modelling, was shown in Figure 6.3. This analysis indicated that approximately 22% of all root node questions are accent-based and that this proportion drops to 16% and 11% for the roots' children and grandchildren, respectively. The analysis of the BE+EE combination, where accent-specific modelling was shown to be superior compared to accent-independent modelling, was presented in Figure 6.8. It showed that approximately 31% of all questions at the root nodes are accent-based and that this proportion drops to 20% and 14% for the roots' children and grandchildren, respectively.

It is apparent that the AE+BE+EE combination is more prone to early partitioning of models into accent-based groups than the AE+EE and BE+EE combinations. This is also similar to the multilingual case presented in [5], where the percentage of language-based questions dropped from more than 45% at the root nodes to less than 5% at the 10^{th} level of depth. It is, however,

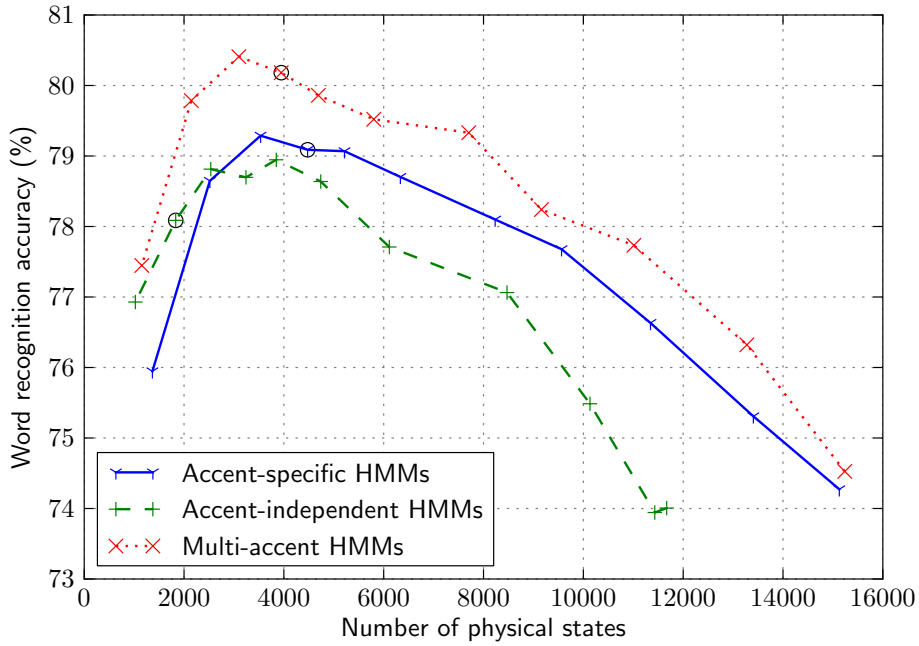


Figure 6.13: AE+BE+EE average evaluation set word accuracies of accent-specific, accent-independent and multi-accent systems as a function of the total number of distinct HMM states. Circular markers indicate the systems delivering optimal performance on the development set.

Table 6.16: The number of states and evaluation set word accuracies for the AE+BE+EE systems delivering optimal performance on the development set and identified by the circular markers in Figure 6.13.

Model set	No. of states	Accuracy (%)
Accent-specific	4476	79.09
Accent-independent	1833	78.09
Multi-accent	3948	80.18

Table 6.17: Evaluation set word accuracies (%) for each accent for the AE+BE+EE systems delivering optimal performance on the development set and identified by the circular markers in Figure 6.13.

Model set	AE	BE	EE	Average
Accent-specific	82.56	71.61	83.10	79.09
Accent-independent	83.66	68.10	82.90	78.95
Multi-accent	84.07	72.81	83.95	80.41

Table 6.18: Statistical significance levels (%) of improvements for pair-wise comparisons of the AE+BE+EE systems delivering optimal performance on the development set and described in Tables 6.14 and 6.16, calculated by using bootstrap confidence interval estimation [80].

Comparison	Phone recognition	Word recognition
Accent-independent vs. accent-specific	99	99
Accent-specific vs. multi-accent	72	99
Accent-independent vs. multi-accent	99	99.9

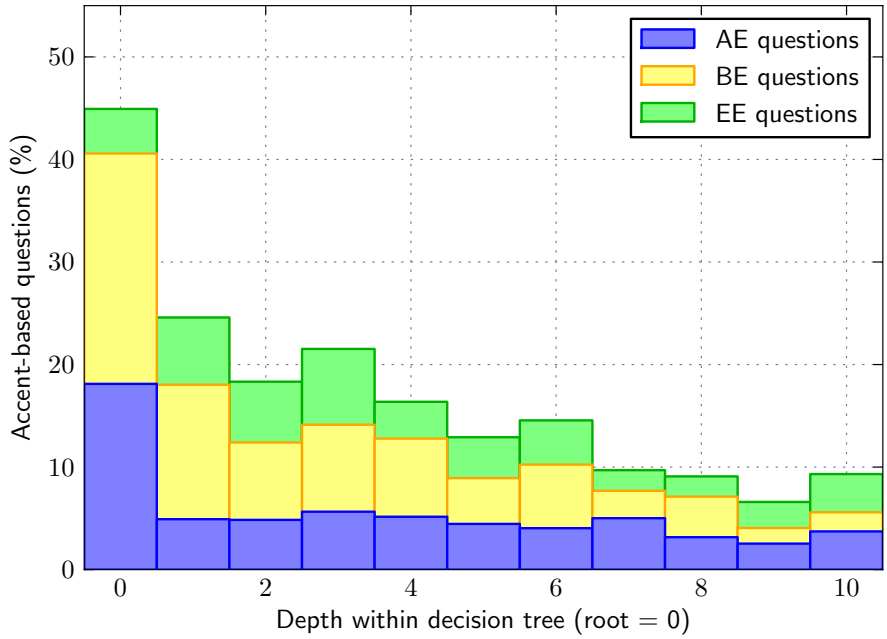


Figure 6.14: Analysis showing the percentage of questions that are accent-based at various depths within the decision-trees for the AE+BE+EE multi-accent system with optimal development set word accuracy. The questions enquire whether the accent of a basephone is AE, BE, or EE. The individual proportions of these questions are also shown.

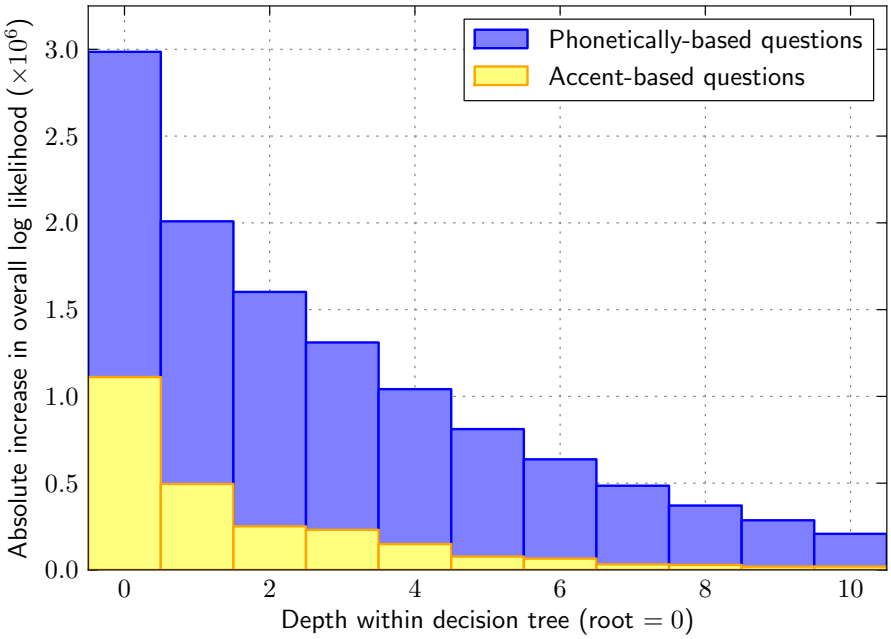


Figure 6.15: Analysis showing the contribution made to the increase in overall log likelihood by the accent-based and phonetically-based questions, respectively, within the decision-trees for the AE+BE+EE multi-accent system with optimal development set word accuracy.

evident from Figure 6.14 that most of the accent-based questions near and including the root of the decision-trees ask whether the basephone belongs to the BE accent (22% of all questions at the root nodes, dropping to 13% and 8% for the roots' children and grandchildren, respectively). The BE question has a twofold effect: firstly it creates a cluster consisting of only BE models and secondly it groups together AE and EE models. In Section 6.2 we concluded that the clustering of AE and EE models is advantageous, while we concluded in Section 6.3 that the separation of BE and EE models is superior to pooled modelling. Figure 6.14 also indicates a high proportion of questions inquiring about the AE accent (approximately 18% of all questions at the root nodes), while very few EE questions are asked. This may suggest that EE is shared most often with the other accents while AE is shared less often and BE is modelled separately most often. This is supported by the analysis given in the following section.

The contribution to the log likelihood improvement made by the accent-based and phonetically-based questions, respectively, during the decision-tree growing process is shown in Figure 6.15 as a function of the depth within the decision-trees. As with AE+EE in Figure 6.4 and BE+EE in Figure 6.9, phonetically-based questions make a larger contribution to the log likelihood improvement than the accent-based questions throughout the decision-trees. However, whereas 22% and 26% respectively of the total increase in likelihood was afforded by accent-based questions at the root nodes for the AE+EE and BE+EE combinations, approximately 37% of the total increase at the root nodes is afforded by accent-based questions for AE+BE+EE. Although this is lower than the 74% seen for multilingual acoustic modelling in [5], the effect of the partitioning into accent-based groups in the decision-trees is more pronounced in this case than for the AE+EE and BE+EE combinations. Figures 6.14 and 6.15 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 development set phone accuracy, as well as for the multi-accent system with the largest number of parameters, revealed similar trends.

6.5.5 Analysis of Cross-Accent Data Sharing

In order to determine to what extent and for which accents data sharing ultimately takes place between AE, EE and EE, we considered the proportion of decision-tree leaf nodes (which correspond to the state clusters) that are populated by states from one, two or all three accents. Figure 6.16 illustrates how these proportions change as a function of total number of clustered states in a system. It is apparent that, as the number of clustered states is increased, the proportion of clusters consisting of a single accent also increases. 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 states from two or all three accents shows a decrease as the number of clustered states increases. For the multi-accent system delivering optimal development set phone accuracy (5801 states, Table 6.14; indicated with the dashed vertical line in Figure 6.16), approximately 36% of the state clusters contain a mixture of accents. For the multi-accent system delivering optimal development set word accuracy (3948 states, Table 6.16; indicated with the dot-dashed vertical line in Figure 6.16), this proportion is approximately 43%. This demonstrates that a significant degree of sharing is taking place across accents. For the optimal multilingual system in [5], only 20% of state clusters contained more than one language, again emphasising that the multi-accent case is much more prone to sharing.

In order to determine which accents are shared most often by the clustering process, Figure 6.17 analyses the proportion of state clusters consisting of groups of accents. It is evident that the largest proportion of two-accent clusters results from the combination of AE and EE. This agrees with the results presented in Section 6.2 which suggested that accent-independent acoustic modelling of AE and EE outperforms accent-specific acoustic modelling. EE and BE is the second

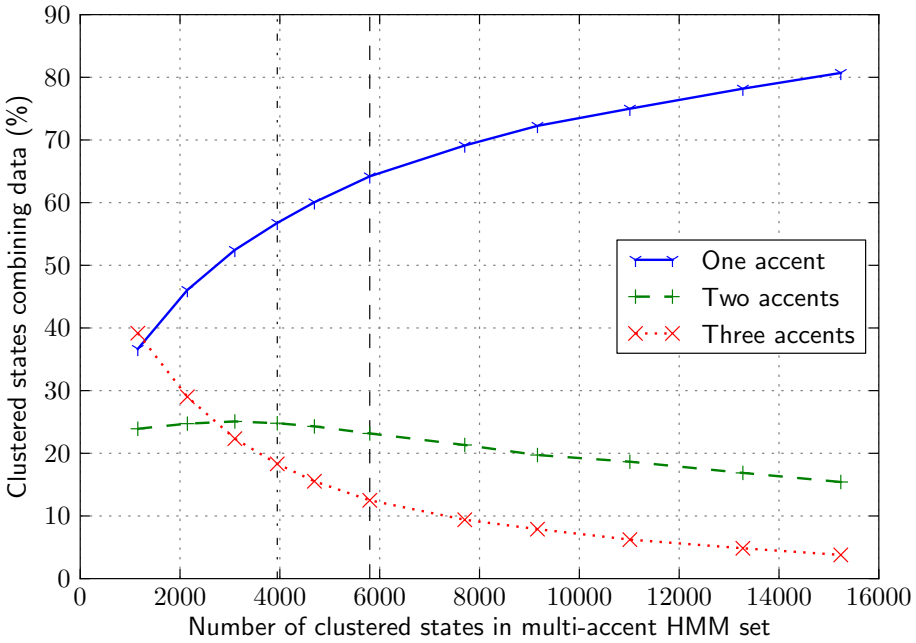


Figure 6.16: Proportion of state clusters combining data from one, two or three accents. The dashed and dot-dashed vertical lines indicate the number of states for respectively the phone and the word recognition systems with optimal performance on the development set.

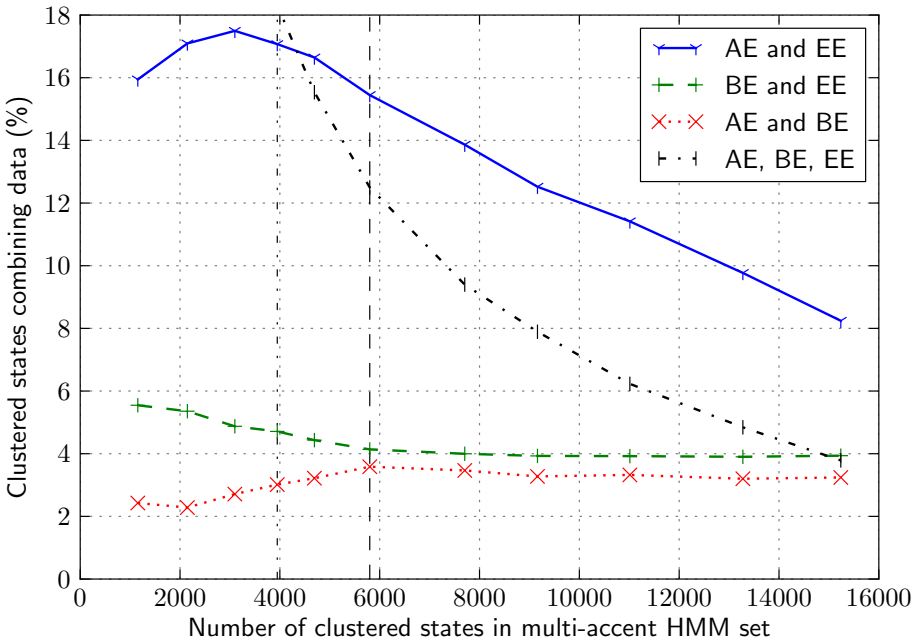


Figure 6.17: Proportion of state clusters combining data from combinations of two or all three accents, with the latter also indicated in Figure 6.16 and shown only partially here. The dashed and dot-dashed vertical lines indicate the number of states for, respectively, the phone and the word recognition systems with optimal performance on the development set.

most frequent combination, and AE and BE the least frequent. Both these combinations are, however, far less common than the AE and EE clusters. The decision-tree analysis presented here corresponds directly to the analysis of accent similarity described in Section 3.5 and illustrated for AE, BE and EE in Figure 6.11. This indicates that, of the three accents, AE and EE are most similar, while AE and BE are most dissimilar. It is not surprising that the similarity analysis and the decision-tree analysis indicate the same trends since the former is based on the monophone HMMs used to initialise the triphone HMMs which are clustered (Section 5.3.1).

In order to determine which accent is separated most often from the other two accents during the clustering process, Figure 6.18 analyses the proportion of state clusters consisting of only one accent. BE is most often found in single-accent clusters, followed by AE and then EE. This indicates that EE is most often involved in sharing and is therefore found least frequently in single-accent clusters, which agrees with the analysis presented in Figure 6.17.

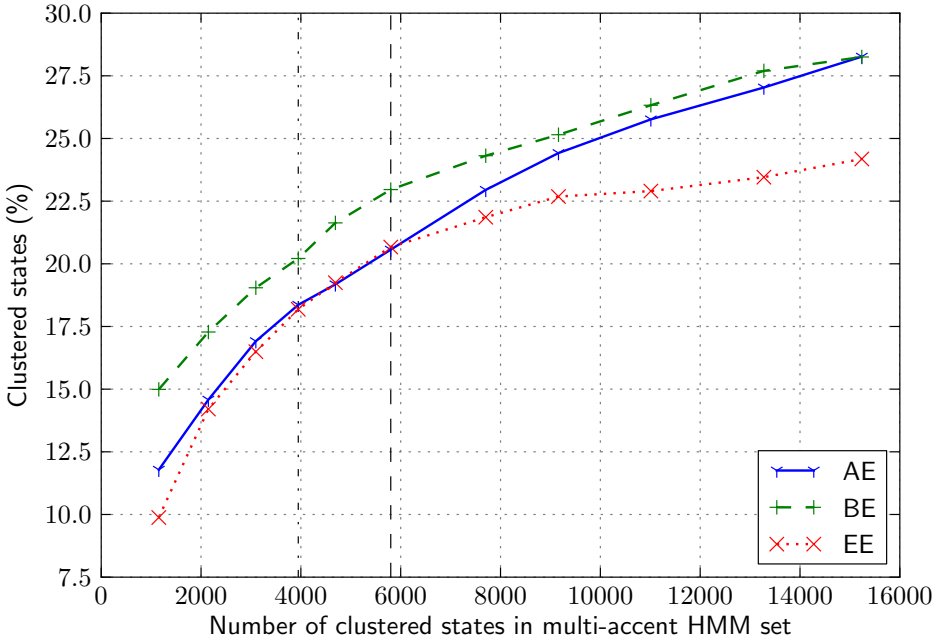


Figure 6.18: Proportion of state clusters containing data from just one accent. The dashed and dot-dashed vertical lines indicate the number of states for, respectively, the phone and the word recognition systems with optimal performance on the development set.

6.5.6 Summary for AE+BE+EE Acoustic Modelling

Multi-accent systems achieved consistent improvements over accent-specific as well as accent-independent systems in both phone and word recognition experiments. Further analysis showed that the decision-trees cluster a relatively high proportion of states together for AE and EE, whilst a relatively high proportion of states are modelled separately for BE. This is consistent with the observations for the separate AE+EE and BE+EE combinations made in the preceding sections. Analysis of the AE+BE+EE decision-trees showed that a considerable proportion of state clusters contain more than one accent, from which we conclude that accent sharing occurs regularly in the multi-accent acoustic models. Compared to the multilingual case in [5], our results and analyses appear to indicate a larger degree of similarity between corresponding phones in different accents than in different languages (as one might expect) and that the multi-accent modelling approach are able to exploit these similarities better and ultimately yield a larger gain in performance compared to the multilingual scenario.

6.6 Acoustic Modelling of the Five Accents of SAE

The three acoustic modelling approaches described in Section 4.2 were applied to the combination of the AE, BE, CE, EE and IE training sets described in Section 3.3. In this section we present phone and word recognition results obtained in an oracle recognition configuration.

6.6.1 Language Models and Pronunciation Dictionaries

Separate accent-specific phone backoff bigram LMs were used in phone recognition experiments while accent-independent word backoff bigram LMs trained on the combined set of training transcriptions of all five accents were used in word recognition experiments. These LMs have already been described in Section 5.1, but for ease of reference the phone and word LM perplexities are repeated in Table 6.19. Recognition systems employed a pooled accent-independent PD obtained by combining all five accent-specific PDs described in Section 5.2. The resulting dictionary contains pronunciations for 7256 words and on average 2.10 pronunciations per word. OOV rates are shown in Table 6.19.

Table 6.19: Phone and word bigram language model (LM) perplexities and OOV rates measured on the evaluation sets of all five SAE accents.

Accent	Accent-specific phone LM		Five-accent accent-independent word LM		
	Bigram types	Perplexity	Bigram types	Perplexity	OOV rate (%)
AE	1891	14.40	11 580	24.07	1.82
BE	1761	15.44	9639	27.87	2.84
CE	1834	14.12	10 641	27.45	1.40
EE	1542	12.64	10 451	24.90	1.08
IE	1760	14.24	11 677	25.55	1.73

6.6.2 Phone Recognition Experiments

Figure 6.19 shows the average evaluation set phone recognition accuracy using eight-mixture triphone HMMs. These were obtained by application of the three acoustic modelling approaches. Recognition performance is shown for a range of model sizes with the systems leading to optimal performance on the development set indicated by circular markers. For each acoustic modelling approach a single curve indicating the average accuracy between the five accents is shown. The number of physical states is calculated as described in Section 6.2.2.

The results presented in Figure 6.19 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 6.20 summarises the evaluation set performance and the number of states for the systems optimised on the development set. Multi-accent acoustic modelling is shown to outperform both accent-specific and accent-independent modelling. Table 6.24 indicates that these improvements are statistically significant at the 99.9% and 80% levels, respectively. 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% [5].

Since the triphone clustering process optimises the overall likelihood, improvements for each individual accent are not guaranteed. Table 6.21 shows the per-accent phone accuracies for

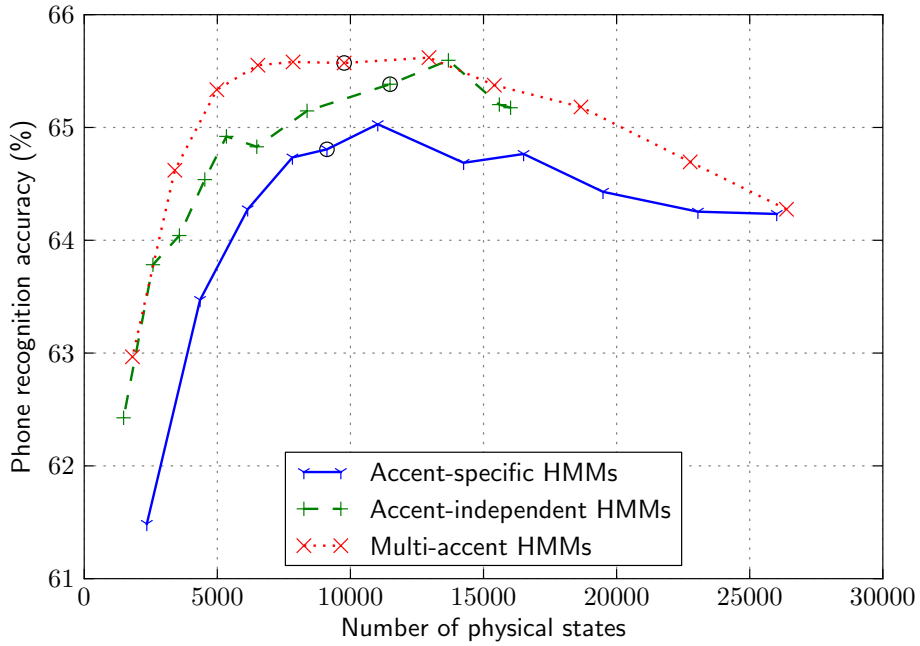


Figure 6.19: Average evaluation set phone accuracies of accent-specific, accent-independent and multi-accent systems as a function of the total number of distinct HMM states. Circular markers indicate the systems delivering optimal performance on the development set.

the systems optimised on the development set. These results show that multi-accent acoustic modelling improves the phone recognition accuracy relative to the remaining two approaches for CE and IE. For BE, accent-specific modelling yields the best performance while for AE and EE, accent-independent modelling leads to the best results. Nevertheless, the average accuracy over all five accents is highest for the multi-accent models. In general the trends in per-accent performance are consistent with the perplexities given in Table 6.19. Best accuracy is achieved for EE (which has the lowest phone perplexity) while phone recognition accuracy is lowest for BE (which has the highest phone perplexity). Similar accuracies are achieved for AE, CE and IE, which have similar phone perplexities.

Table 6.21 also indicates 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 model sets. For BE, on the other hand, it is better to do the latter. 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 as noted in [5] and [48, 50] (see Section 2.4). A further important observation, which will be supported by the word 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.

Table 6.20: The number of states and evaluation set phone accuracies for the five-accent systems delivering optimal performance on the development set and identified by the circular markers in Figure 6.19.

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

Table 6.21: Evaluation set phone accuracies (%) for each accent individually for the five-accent systems delivering optimal performance on the development set and identified by the circular markers in Figure 6.19.

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

6.6.3 Word Recognition Experiments

Figure 6.20 shows the average evaluation set word recognition accuracy using eight-mixture triphone HMMs. These were obtained by application of the three acoustic modelling approaches. Once again, performance is shown for a range of acoustic models with differing numbers of physical states, with the systems leading to optimal performance on the development set identified by circular markers. For each acoustic modelling approach a single curve indicating the average accuracy between the five accents is shown. The number of physical states is calculated as described in Section 6.2.2.

Figure 6.20 indicates that, over the range of models considered, multi-accent modelling consistently outperforms both accent-specific and accent-independent acoustic modelling. The evaluation set performance for the systems optimised on the development set is summarised in Table 6.22. For these systems, multi-accent acoustic modelling outperforms both accent-specific and accent-independent modelling. The performance improvements exhibited by the multi-accent system (82.78% accuracy) relative to both the accent-specific system (81.53%) as well as the accent-independent system (81.52%) were found to be significant at the 99.9% level, as indicated in Table 6.24.

Table 6.23 presents the word accuracies separately for each accent. For all accents, better performance is achieved when the multi-accent models are used even though multi-accent modelling does not guarantee per-accent performance improvements. For CE, EE and IE, accent-independent models obtained by pooling the training data result in better performance than that achieved by the accent-specific models. 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.

As in the phone recognition experiments, the relative per-accent performance is in general consistent with the perplexities given in Table 6.19. The best accuracy is achieved for AE which has the lowest word perplexity while word recognition accuracy is lowest for BE which has the highest perplexity. Although IE has a lower perplexity than CE, its word recognition performance is lower. This peculiar result for IE seems to indicate that IE may be acoustically more difficult to recognise than the other accents. Interestingly, Table 6.23 indicates that the accent-specific and multi-accent modelling approaches yield slightly better performance for the second language AE than for the first language EE, while this was not true for the corresponding phone accuracies in Table 6.21. We believe that this can be attributed to the word perplexities which are lower for AE than for EE while the opposite is true for the phone LMs (Table 6.19). Furthermore, as pointed out in Section 3.2.1, the EE and AE accents of SAE are closely related. This is also supported by the results presented in Sections 6.2, 6.4 and 6.5.

Although the phone recognition results in Figure 6.19 indicate that multi-accent acoustic modelling does not outperform accent-independent modelling at all system sizes, the word recognition

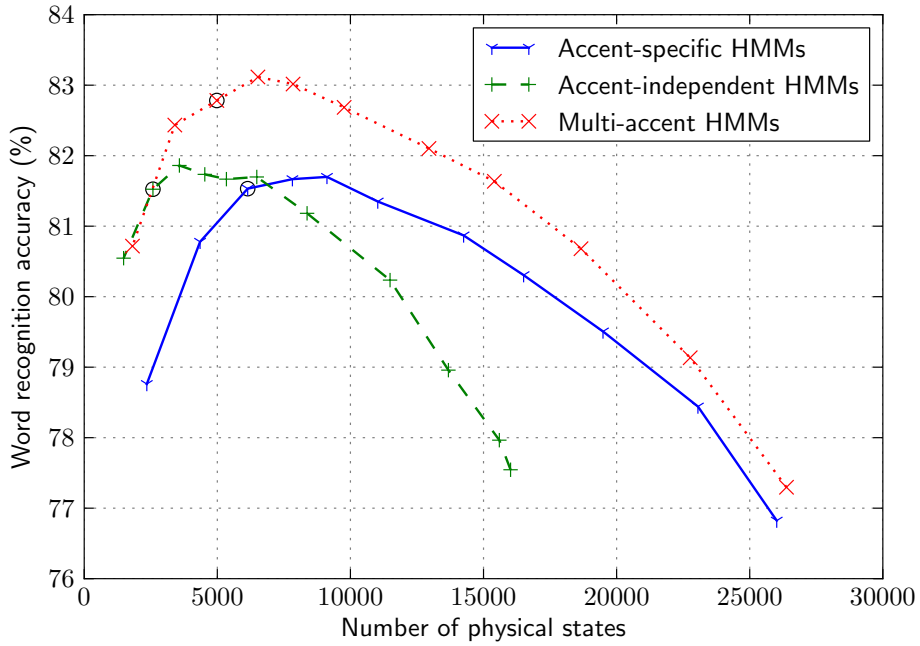


Figure 6.20: Average evaluation set word accuracies of accent-specific, accent-independent and multi-accent systems as a function of the total number of distinct HMM states. Circular markers indicate the systems delivering optimal performance on the development set.

Table 6.22: The number of states and evaluation set word accuracies for the five-accent systems delivering optimal performance on the development set and identified by the circular markers in Figure 6.20.

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

Table 6.23: Evaluation set word accuracies (%) for each accent individually for the five-accent systems delivering optimal performance on the development set and identified by the circular markers in Figure 6.20.

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

Table 6.24: Statistical significance levels (%) of improvements for pair-wise comparisons of the five-accent systems delivering optimal performance on the development set and described in Tables 6.20 and 6.22, calculated by using bootstrap confidence interval estimation [80].

Comparison	Phone recognition	Word recognition
Accent-specific vs. accent-independent	99	50
Accent-specific vs. multi-accent	99.9	99.9
Accent-independent vs. multi-accent	80	99.9

results indicate consistent improvements over both accent-specific and accent-independent modelling with significance at the 99.9% level. As in Section 6.5, these results correspond to some degree to the results presented in [5] in which multilingual acoustic modelling was shown to outperform language-specific and language-independent modelling for four languages spoken in South Africa. The performance improvements obtained here are, however, greater. The improvements achieved by multi-accent modelling are more consistent in the five-accent case than for either AE+EE (Section 6.2) or BE+EE (Section 6.3), and in word recognition experiments the improvements are slightly greater compared to those obtained for the AE+BE+EE combination (Section 6.5).

6.6.4 Analysis of the 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 that they become increasingly less frequent towards the leaves. Figure 6.21 analyses the decision-trees of the multi-accent system delivering optimal word accuracy on the development set (4982 states, Table 6.22). 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, the separation of BE and IE tends to occur more frequently than for the other accents. As discovered in Sections 6.6.2 and 6.6.3, BE was also the only accent for which accent-specific (i.e. separate) modelling led to higher phone and word recognition accuracies compared to accent-independent (i.e. pooled) models. Furthermore, the similarity analysis presented in Section 3.5 indicated that BE and IE are relatively different from the other accents. The proportion of accent-based questions is much larger for the five-accent system than for any of the other accent combinations considered in this chapter. The analysis is also more consistent with the multilingual case presented in [5], although accent-based questions seem to be asked more consistently through the depths of the decision-trees while language-based questions were more focussed at levels close to the root nodes in the multilingual case.

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 6.22 as a function of the 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 6.22, approximately 37% of the total increase at the root nodes is afforded by accent-based questions. This proportion is similar to the proportion calculated for the AE+BE+EE accent combination, but lower than the corresponding figure of 74% for multilingual acoustic modelling [5]. Figures 6.21 and 6.22 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 multi-accent system with the largest number of parameters, revealed similar trends.

6.6.5 Analysis of Cross-Accent Data Sharing

In order to determine to what extent and for which accents data sharing ultimately takes place in 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 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 6.23 illustrates how these proportions change as a function of the total number of clustered states in a system.

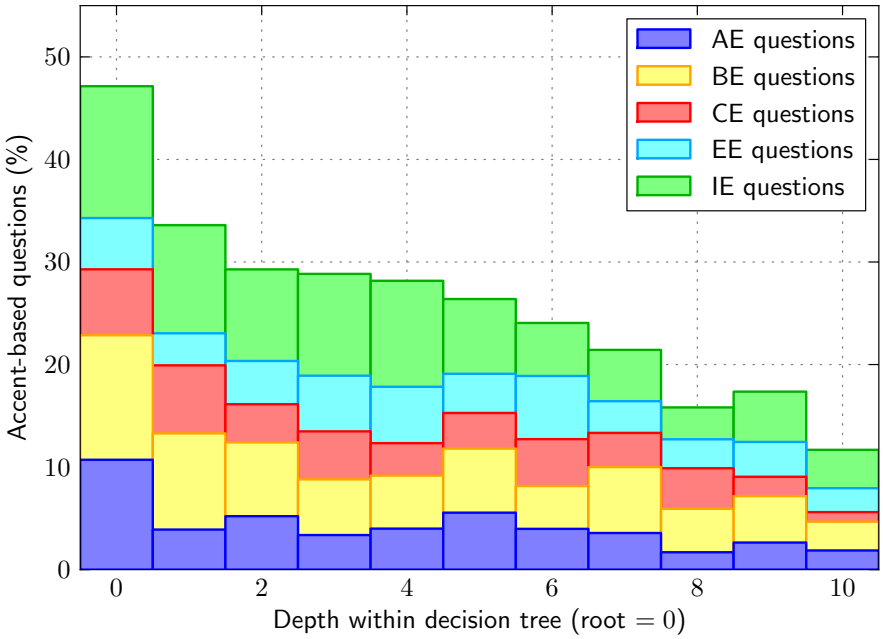


Figure 6.21: Analysis showing the percentage of questions that are accent-based at various depths within the decision-trees for the multi-accent system with optimal development set word accuracy. The questions enquire whether the accent of a basephone is AE, BE, CE, EE or IE. The individual proportions of these questions are also shown.

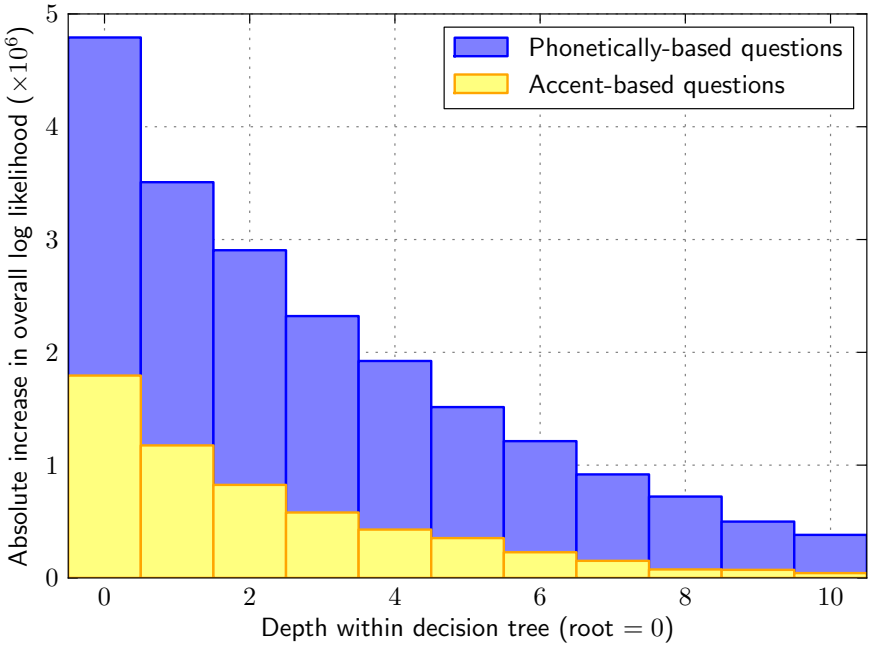


Figure 6.22: Analysis showing the contribution made to the increase in overall log likelihood by the accent-based and phonetically-based questions, respectively, within the decision-trees for the multi-accent system with optimal development set word accuracy.

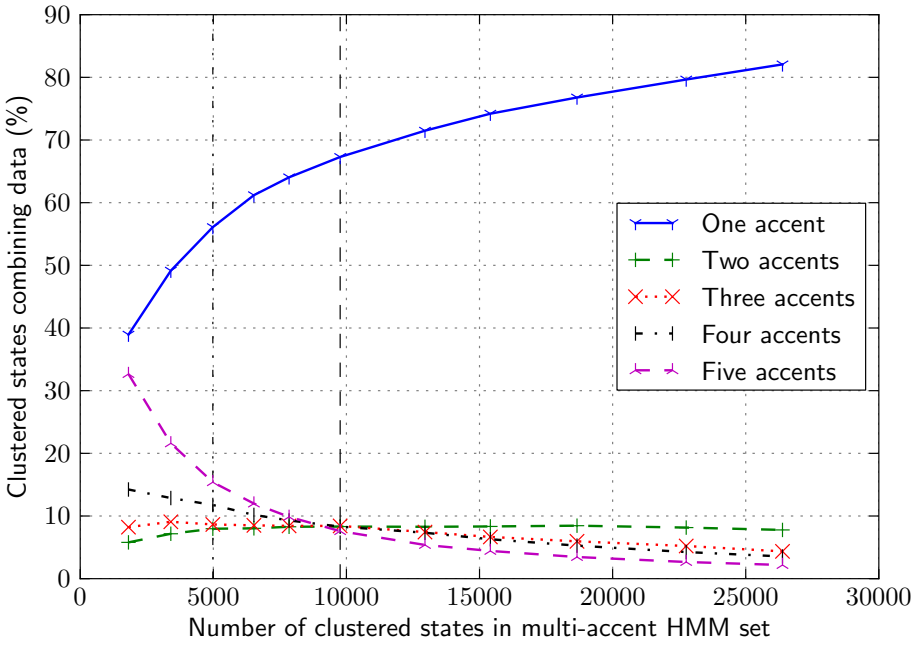


Figure 6.23: Proportion of state clusters combining data from one, two, three, four or all five accents. The dashed and dot-dashed vertical lines indicate the number of states for, respectively, the phone and the word recognition systems with optimal performance on the development set.

From Figure 6.23 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 development set phone accuracy (9765 states, Table 6.20; indicated with the dashed vertical line in Figure 6.23), approximately 33% of state clusters contain a mixture of accents, while 44% of state clusters contain a mixture of accents for the multi-accent system delivering optimal development set word accuracy (4982 states, Table 6.22; indicated with the dot-dashed vertical line in Figure 6.23). 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 [5].

In order to determine which accents are shared most often by the clustering process, Figures 6.24, 6.25 and 6.26 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 6.24 that the largest proportion of two-accent clusters results from the combination of AE and EE and of AE and CE. All other combinations are far less common. In Section 3.2.1 the influence of Afrikaans on EE was noted, and this may account for a higher degree of similarity between AE and EE. The similarity of AE and EE is also supported by the results presented in Sections 6.2 and 6.5. The influence of Afrikaans on CE and the use of Afrikaans as a first language by many CE speakers (Section 3.2.4) may in turn explain a phonetic similarity and therefore a higher degree of sharing between AE and CE. Figure 6.25 indicates that AE, CE and EE form the most frequent three-accent combination, followed by the combination of BE, CE and EE. Furthermore, Figure 6.26 shows that the two most frequent four-accent combinations are AE, CE, EE, IE and AE, BE, CE, EE, both of which 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 separated most often from the others during the clustering process, Figure 6.27 presents the proportion of state clusters consisting of just one accent.

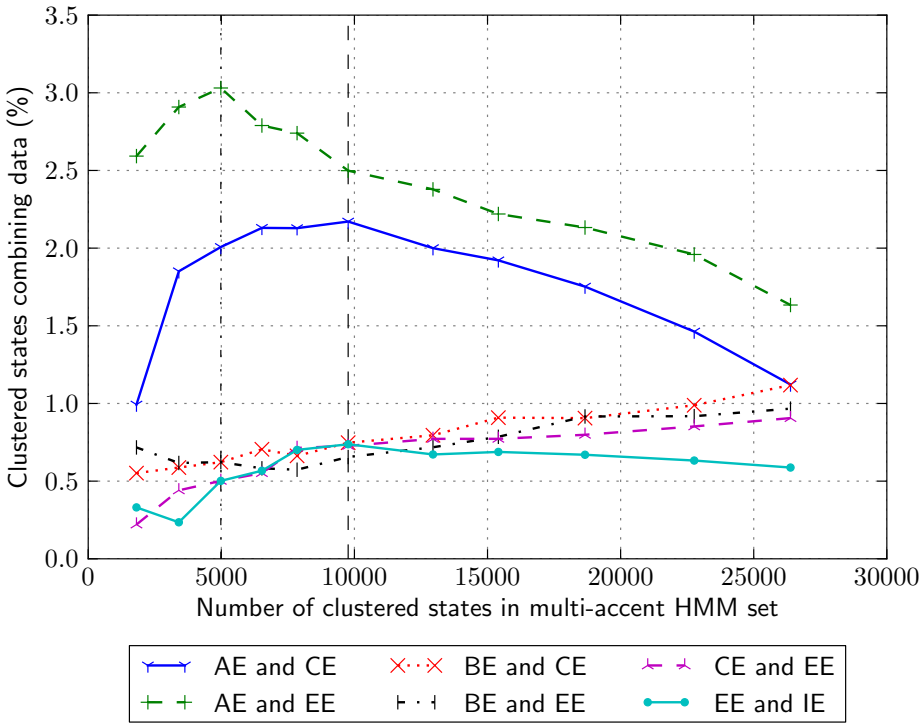


Figure 6.24: Proportion of state clusters combining data from various combinations of two accents. The dashed and dot-dashed vertical lines indicate the number of states for, respectively, the phone and the word recognition systems with optimal performance on the development set.

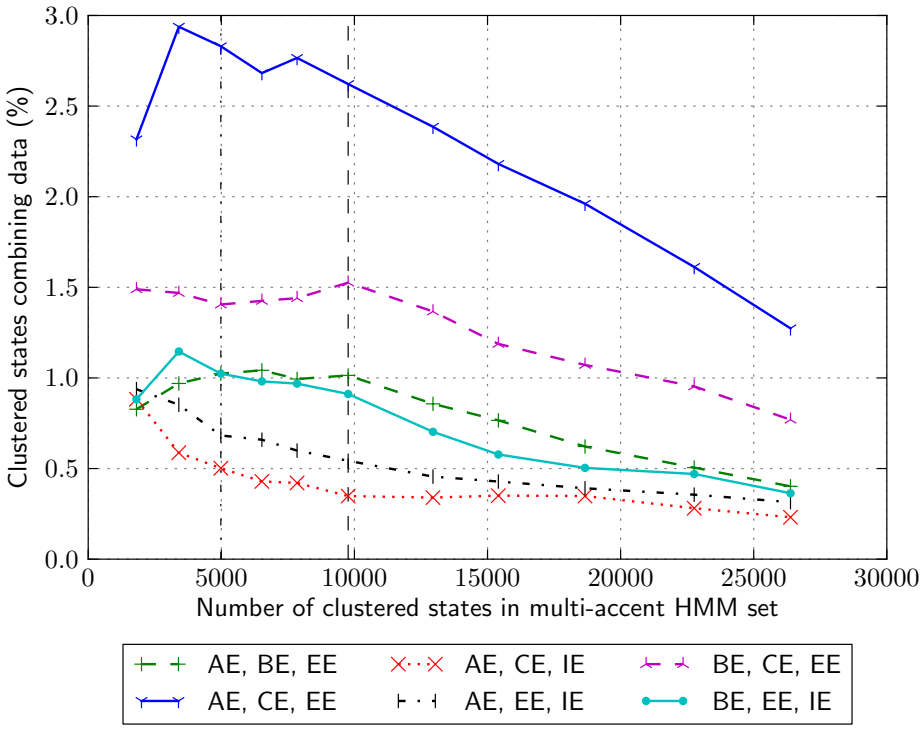


Figure 6.25: Proportion of state clusters combining data from various combinations of three accents. The dashed and dot-dashed vertical lines indicate the number of states for, respectively, the phone and the word recognition systems with optimal performance on the development set.

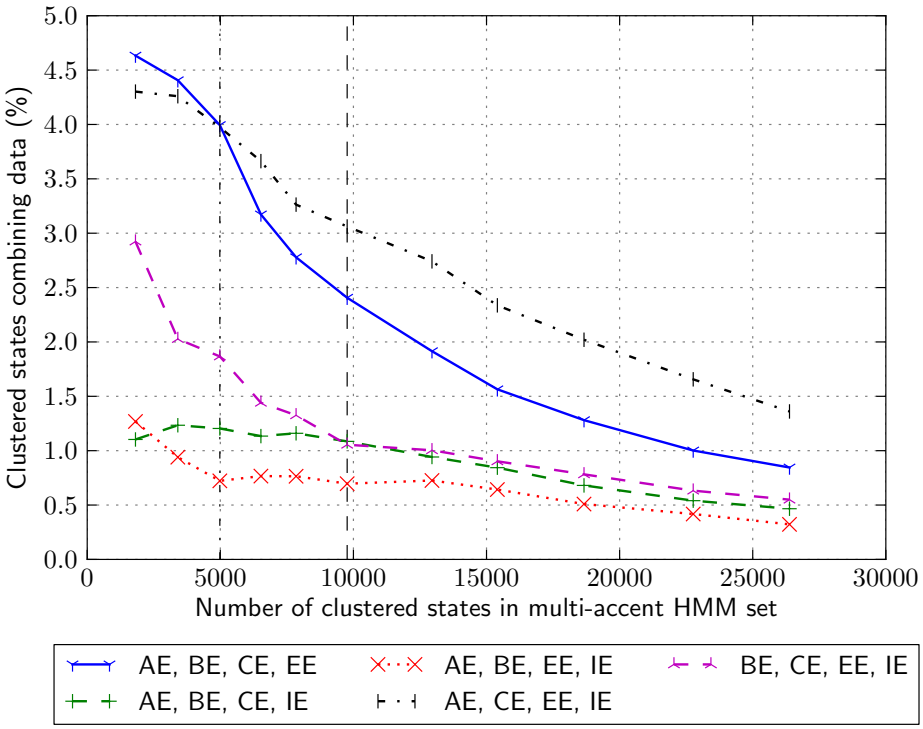


Figure 6.26: Proportion of state clusters combining data from various combinations of four accents. The dashed and dot-dashed vertical lines indicate the number of states for, respectively, the phone and the word recognition systems with optimal performance on the development set.

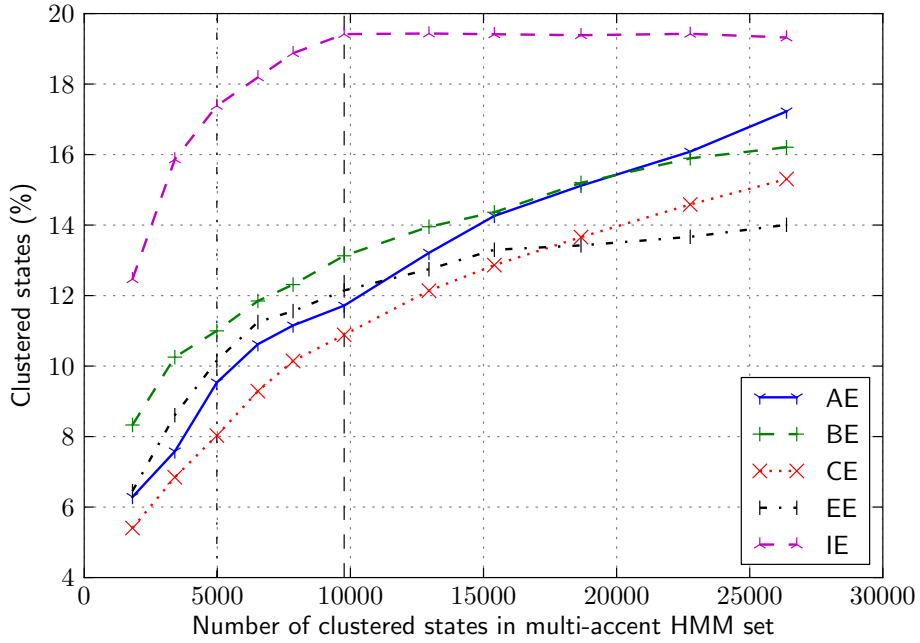


Figure 6.27: Proportion of state clusters containing data from just one accent. The dashed and dot-dashed vertical lines indicate the number of states for, respectively, the phone and the word recognition systems with optimal performance on the development set.

The most striking feature of this figure 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 6.27 lends further support to our conclusion that AE, CE and EE are most similar, since these three accents occur least frequently in single-accent clusters. The analysis presented in this section concurs with the discussion in Section 3.2 in which the influence of the AE, CE and EE accents on one another was noted. This is also consistent with the analysis of accent similarity described in Section 3.5 which indicated the similarity of AE, CE and EE while BE and IE were shown to be more different.

6.6.6 Summary for Five-Accent Acoustic Modelling

As with the AE+BE+EE combination (Section 6.5), multi-accent systems achieved consistent improvements over the accent-specific as well as accent-independent systems in both phone and word recognition experiments when applied to the five accents of SAE. Further analysis showed that the decision-trees cluster a relative high proportion of states together for AE, CE and EE, whereas a relative high proportion of states are modelled separately for BE and IE. The improvements in word recognition accuracy achieved by multi-accent models over accent-specific and accent-independent models is greater for the five-accent combination than for the AE+EE (Section 6.2), the BE+EE (Section 6.3) and the AE+BE+EE (Section 6.5) combinations. Performance improvements are also greater than those achieved in multilingual acoustic modelling [5].

6.7 Summary and Conclusions

In this chapter we evaluated three acoustic modelling techniques: (i) accent-specific modelling, in which multiple accents are modelled separately; (ii) accent-independent modelling, in which acoustic training data is pooled across accents; and (iii) multi-accent modelling, which allows selective data sharing across accents. We applied these modelling approaches to different combinations of SAE accents in order to obtain some intuition into the behaviour of the approaches for accent combinations which differ in their degree of similarity.

In Section 6.2 we considered acoustic modelling of AE and EE, two accents which are quite similar. Both phone and word recognition experiments indicated that accent-independent acoustic models outperform accent-specific models and that multi-accent models achieve slightly poorer performance compared to accent-independent models. Acoustic modelling of BE and EE, two accents which are relatively dissimilar, was considered in Section 6.3. Phone and word recognition results indicated that accent-specific models are superior to accent-independent models and that multi-accent acoustic models yield similar or improved performance relative to accent-specific models for these accents. The recognition results for the AE+EE and BE+EE accent pairs indicated that the decision of whether to separate or to pool data across accents depends greatly on the accents involved and, in particular, on the similarity of the accents.

The experiments presented in Section 6.5, which considered modelling of the AE+BE+EE accent combination, indicated that, while the decision to pool or to separate the training data depends on the particular accents in question, multi-accent modelling allows almost all of the possible gain to be obtained in a data-driven manner. In both phone and word recognition experiments, multi-accent acoustic modelling consistently outperformed the two alternative approaches. This behaviour was also observed in the experiments described in Section 6.6 in which acoustic modelling of all five SAE accents was considered. For the five-accent combination, the superior performance of multi-accent acoustic models was illustrated in both phone and word recognition experiments. Multi-accent models yielded a word recognition accuracy of 82.78%,

which represents an improvement of 1.25% absolute over accent-specific and accent-independent models. These improvements were found to be statistically significant at the 99.9% level.

Throughout this chapter, both recognition performance as well as the analysis of the multi-accent decision-trees highlighted the relative similarity of the AE, CE and EE accents while the BE and IE accents were found to be more dissimilar from the other accents and each other. This corresponds to the analysis of accent similarity presented in Section 3.5 as well as the analysis of pronunciation dictionary similarity described in Section 5.2.

For the experiments presented in this chapter we used phone backoff bigram language models, accent-independent word backoff bigram language models and pooled accent-independent pronunciation dictionaries. These design decisions were made based on preliminary experiments which are described in Appendix C. In that appendix, we evaluate the performance achieved when alternative language and pronunciation modelling approaches are used. Importantly, we show that the relative performance of the three acoustic modelling approaches does not change when the alternative language models and dictionaries are used.

Finally, the experiments described in this chapter were based on an oracle recognition setup in which the accent of each test utterance is assumed to be known. By configuring the setup in this way, acoustic modelling effects are isolated since the effects of misclassifications are not taken into account. In the next chapter we consider the development of systems for which the accent of the input speech is assumed to be unknown.

Chapter 7

Analysis of Misclassifications in Parallel Recognition Experiments

A distinction can be made between oracle recognition and parallel recognition when recognition of accented speech is considered, as discussed in Sections 2.1.2 and 2.6. In Chapter 6 an oracle recognition configuration was used. In this configuration, each test utterance is presented only to the recogniser tailored to the matching accent. This was done in order to evaluate the different acoustic modelling approaches without allowing performance to be influenced by the effects of accent misclassifications. In this chapter we consider parallel recognition in which each test utterance is presented to a bank of accent-specific speech recognisers and the output with the highest associated likelihood is selected. This represents the practical scenario where the accent of the test data is assumed to be unknown. By comparing oracle and parallel system performance, we analyse the effects on recognition performance caused by misclassifications. These occur in the accent identification (AID) process that is performed implicitly during parallel recognition.

As in Chapter 6, we consider different combinations of SAE accents: AE+EE, BE+EE and the five-accent combination. As noted in the previous chapters, the AE+EE pair represents accents which are quite similar, while the BE+EE pair represents accents which are relatively dissimilar. This is the motivation for considering these two accent pairs first. Both phone and word recognition experiments are performed and all three acoustic modelling approaches described in Section 4.2 are evaluated.

7.1 Related Research[1]

When multiple accents have to be simultaneously recognised, one approach is to explicitly precede accent-specific speech recognition with AID [54]. This is illustrated in Figure 2.6. Two alternatives to this approach are described by Chengalvarayan [4]. The first is a system where a bank of accent-specific recognisers is run in parallel and the output with the highest associated likelihood is selected (Figure 2.3). AID is thus performed implicitly during recognition. The second alternative is to train a single model set by pooling data across accents (Figure 2.5). For recognition of American, Australian and British English, the latter exhibited the best performance in [4]. In [53], recognition of non-native English from six European countries was considered. Identification followed by recognition gave comparable performance to an oracle system, but both were outperformed by a pooled system.

[1] This section provides a brief summary of relevant literature which has already been described in Section 2.6.

7.2 System Optimisation, Configuration and Objectives

By running multiple accent-specific recognisers[2] in parallel and then selecting the output with the highest associated likelihood, we performed speech recognition experiments for the AE+EE, BE+EE and five-accent combinations. This system configuration is illustrated in Figure 7.1 for the five-accent case. The selection of the highest scoring result can be performed independently for each utterance, leading to per-utterance AID, or for each speaker, leading to per-speaker AID. We considered both cases. Furthermore, the performance of these parallel systems was compared to those of oracle systems in which each test utterance is presented only to the correct accent-specific recogniser. This configuration is illustrated in Figure 7.2 for the five-accent combination.

[2]'Accent-specific recogniser' used in this context could refer to a system employing accent-specific or multi-accent acoustic models since both are dependent on accent. A recogniser employing accent-independent acoustic models with an accent-specific language model could also be considered an accent-specific recogniser.

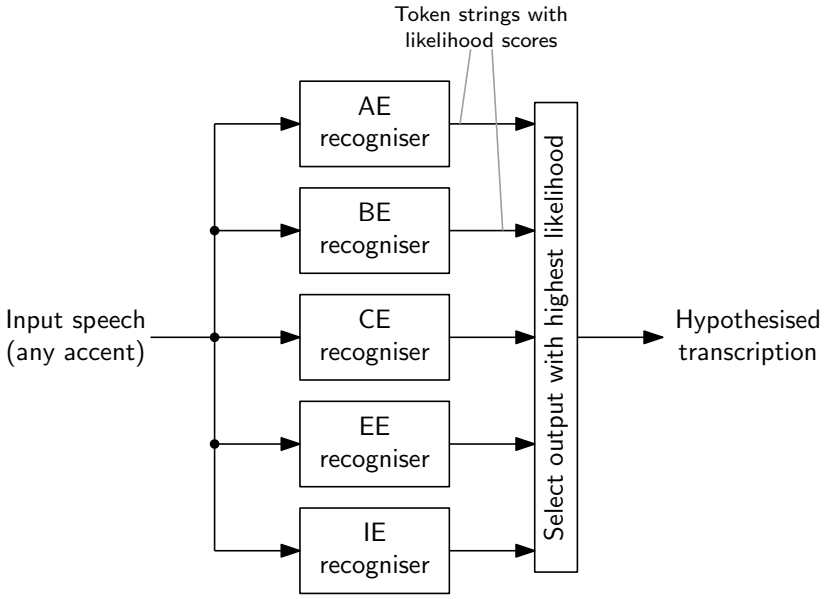


Figure 7.1: A speech recognition system employing multiple accent-specific recognisers in parallel for simultaneous recognition of the five SAE accents.

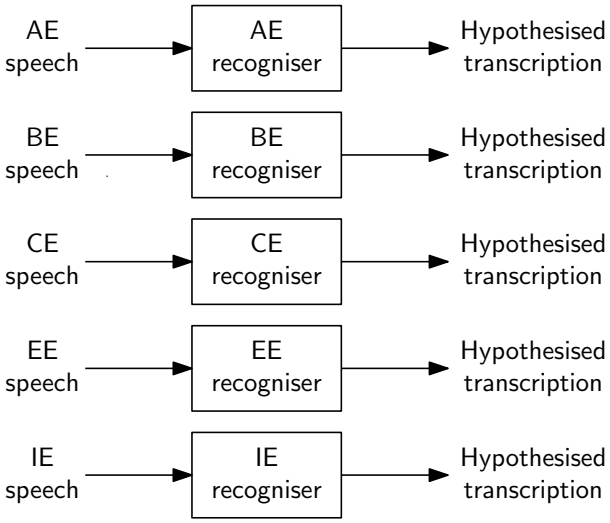


Figure 7.2: Oracle speech recognition of the five SAE accents in which test utterances are presented only to the accent-specific recognition system matching the accent of that utterance.

The experiments described in Appendix C indicate that accent-specific language modelling is advantageous for phone recognition while accent-independent language modelling is superior for word recognition. Despite these results, we evaluated both accent-specific and accent-independent language models (LMs) in the phone and word recognition experiments presented in this chapter. This was done since, although the word accent-specific LMs might for instance be poorly trained, these models could result in better discrimination between accents in a parallel configuration. The three acoustic modelling approaches described in Section 4.2 were therefore used in combination with both accent-specific and accent-independent LMs. A single recogniser was used for the systems employing both accent-independent acoustic and language models and these systems correspond to the pooled systems described in Section 7.1.

The optimisation approach described in Section 5.3.3 was also followed to optimise the systems for the experiments in this chapter. In particular, all phone and word recognition accuracies given in this chapter were measured on the evaluation set for systems optimised exclusively on the development set. Each oracle system used the same acoustic and language models as the parallel system it was compared to. Some of the oracle results given in this chapter match the results presented in Chapter 6 (compare for instance the phone recognition results in Table 6.2 to the results for the oracle systems employing the accent-specific LMs given in Table 7.3). The parallel recognition systems perform AID as a by-product of recognition and these accuracies can also be measured. When we regard our systems as AID systems, phone-based systems correspond to the parallel phone recognition identification approach described in [82], while our word-based systems are comparable to the identification systems used in [83]. One of the chief aims of this chapter is to determine the degree of performance deterioration caused by accent misclassifications that occur when running accent-specific recognition systems in parallel.

7.3 Parallel Recognition of AE+EE and BE+EE

Following the methodology described in Section 7.2, oracle and parallel speech recognition systems were developed using the combination of the AE and EE as well as of the BE and EE training sets described in Section 3.3.

7.3.1 Language Models and Pronunciation Dictionaries

Accent-specific phone LMs trained individually on the transcriptions of each accent (Section 5.1), as well as AE+EE and BE+EE accent-independent phone LMs trained on the combined phone-level training set transcriptions of the two accents involved, were applied in phone recognition experiments. Perplexities for these phone LMs are given in Table 7.1. Accent-specific word LMs trained on the respective transcriptions of each accent, as well as AE+EE and BE+EE accent-independent word LMs trained on the training transcriptions of all five SAE accents (Sections 6.2.1 and 6.3.1), were applied in word recognition experiments. Perplexities for these word LMs are given in Table 7.2. The AE+EE and BE+EE accent-independent LMs differ only in their vocabularies, which were taken from the respective training sets of each accent pair. The accent-specific pronunciation dictionaries (PDs) described in Section 5.2 were used with the accent-specific word LMs while the AE+EE and BE+EE accent-independent PDs described in Sections 6.2.1 and 6.3.1 were employed with the corresponding accent-independent word LMs. Out-of-vocabulary (OOV) rates are shown in Table 7.2.

Table 7.1: Phone bigram language model (LM) perplexities (perp.) measured on the evaluation sets of the AE, BE and EE accents.

Accent	Phone bigram types	Accent-specific LM perp.	AE+EE LM perp.	BE+EE LM perp.
AE	1891	14.40	15.03	-
BE	1761	15.44	-	16.95
EE	1542	12.64	13.34	13.99

Table 7.2: Word bigram language model (LM) perplexities and OOV rates measured on the evaluation sets of the AE, BE and EE accents.

Accent	Bigram types	Accent-specific LMs		AE+EE LM		BE+EE LM	
		Perplexity	OOV (%)	Perplexity	OOV (%)	Perplexity	OOV (%)
AE	11 580	25.81	4.87	23.48	3.06	-	-
BE	9639	30.30	6.90	-	-	26.74	3.87
EE	10 451	28.97	4.74	23.85	2.12	24.04	3.04

7.3.2 Phone Recognition Experiments

Tables 7.3 and 7.4 show the average phone recognition and per-utterance AID accuracies measured on the evaluation sets for the AE+EE and BE+EE systems, respectively, developed according to the methodology described in Section 7.2. Per-utterance AID was performed for the parallel systems. Because a single recogniser was used for the systems employing both accent-independent acoustic and language models, identical results were obtained for the oracle and parallel tests. AID cannot be performed by these fully accent-independent systems.

The results in Tables 7.3 and 7.4 indicate that for both accent pairs and all three acoustic modelling approaches, the systems employing accent-specific LMs outperform those with accent-independent LMs in both recognition and identification. This agrees with the phone LM perplexities given in Table 7.1, which indicates that the perplexities of the accent-independent LMs are higher than those of the accent-specific LMs. It also agrees with the results presented in Section C.1 which indicate that accent-specific language modelling is superior to accent-independent modelling for phone recognition.

The results in Tables 7.3 and 7.4 also indicate that in all cases the oracle systems outperform their parallel counterparts. Using bootstrap confidence interval estimation [80], these improvements were calculated to be statistically significant at at least the 98% level. This indicates that, if improved AID could be performed prior to recognition, phone recognition accuracies would improve.

When the AID accuracies in Tables 7.3 and 7.4 are compared, it is evident that BE and EE can be distinguished with much higher accuracy than AE and EE. This agrees with the analysis in Section 3.5 which indicated that AE and EE are more similar than BE and EE (also see Figure 6.11). Of the three acoustic modelling approaches, accent-independent models yield the best results in both oracle and parallel recognition experiments for the AE+EE accent pair. Multi-accent acoustic modelling yields slightly improved performance over accent-specific modelling for the BE+EE pair, with accent-independent modelling performing worst in both oracle and parallel recognition. In general, phone recognition accuracies are lower for BE+EE compared to AE+EE. This is attributed mainly to the higher phone LM perplexities of BE compared to the other two accents, as indicated in Table 7.1.

Table 7.3: Performance of AE+EE oracle and parallel phone recognition systems when applying per-utterance AID. Average phone recognition accuracies (%) and per-utterance AID accuracies are indicated.

Model set	Accent-specific LMs			Accent-independent LM		
	Oracle	Parallel	AID (%)	Oracle	Parallel	AID (%)
Accent-specific	68.80	67.49	77.57	67.96	66.94	77.07
Accent-independent	70.01	69.38	64.85	68.64	68.64	-
Multi-accent	69.81	69.16	77.93	68.63	68.26	76.49

Table 7.4: Performance of BE+EE oracle and parallel phone recognition systems when applying per-utterance AID. Average phone recognition accuracies (%) and per-utterance AID accuracies are indicated.

Model set	Accent-specific LMs			Accent-independent LM		
	Oracle	Parallel	AID (%)	Oracle	Parallel	AID (%)
Accent-specific	65.04	64.08	90.95	63.76	62.93	90.60
Accent-independent	63.98	63.61	85.35	62.41	62.41	-
Multi-accent	65.08	64.35	91.78	63.86	63.43	90.81

7.3.3 Word Recognition Experiments

Tables 7.5 and 7.6 show the average word recognition and per-utterance AID accuracies measured on the evaluation sets for the AE+EE and BE+EE systems, respectively, developed according to the methodology described in Section 7.2. Per-utterance AID was performed for the parallel systems.

The results in Tables 7.5 and 7.6 show consistently superior performance for the systems employing accent-independent LMs compared to those employing accent-specific LMs. This is not only the case for recognition but also, more surprisingly, for AID. Table 7.2 shows that the perplexities and OOV rates measured on the evaluation sets are also in all cases higher for the accent-specific LMs than for the accent-independent LMs. This is attributed to the very small amount of data available for LM training (Table 3.2) and corresponds to the experimental results discussed in Section C.2. We therefore focus on systems using accent-independent LMs in the following comparison of the oracle and parallel recognition tests, although the accent-specific LM systems show similar trends. Although even the accent-independent LMs may be considered poorly trained, they are common to all recognition systems. Furthermore, the accent-independent PD is also common to all corresponding systems. Hence, the individual recognition systems used in the parallel recognition approaches are distinguished solely by their acoustic models.

For the AE+EE systems using accent-independent LMs, the parallel systems employing accent-specific and multi-accent acoustic models show small improvements over the corresponding oracle systems. These improvements have been calculated to be statistically significant at the 99% and 74% confidence levels for the two approaches, respectively. Although the improvements are small, it is noteworthy that accent misclassifications do not lead to deteriorated system performance. Instead, the misclassifications improve overall recognition performance indicating that some test utterances are better matched to the acoustic models of the other accent. In contrast we observe deteriorated performance for the BE+EE pair when using a parallel recognition approach with the accent-independent LM. The superior performance of the BE+EE oracle systems is statistically significant at the 99% level for both the accent-specific and multi-accent acoustic modelling approaches. The results also indicate that the parallel recognition performance of the multi-accent acoustic models is better than that achieved using accent-specific and accent-

Table 7.5: Performance of AE+EE oracle and parallel word recognition systems when applying per-utterance AID. Average word recognition accuracies (%) and per-utterance AID accuracies are indicated.

Model set	Accent-specific LMs			Accent-independent LM		
	Oracle	Parallel	AID (%)	Oracle	Parallel	AID (%)
Accent-specific	78.70	81.38	78.72	84.01	84.63	80.23
Accent-independent	79.57	82.38	67.72	84.78	84.78	-
Multi-accent	79.50	81.85	78.07	84.78	84.88	78.22

Table 7.6: Performance of BE+EE oracle and parallel word recognition systems when applying per-utterance AID. Average word recognition accuracies (%) and per-utterance AID accuracies are indicated.

Model set	Accent-specific LMs			Accent-independent LM		
	Oracle	Parallel	AID (%)	Oracle	Parallel	AID (%)
Accent-specific	71.64	71.51	92.26	76.69	76.07	93.23
Accent-independent	71.63	72.43	86.87	75.38	75.38	-
Multi-accent	72.40	72.22	92.81	77.35	76.75	93.16

Table 7.7: Performance of AE+EE oracle and parallel word recognition systems when applying per-speaker AID. Average word recognition accuracies (%) and per-utterance AID accuracies are indicated for systems employing the AE+EE accent-independent LM.

Model set	Oracle	Parallel	AID (%)
Accent-specific	84.01	84.09	91.95
Multi-accent	84.78	84.81	94.75

independent acoustic models for both accent pairs. The extent of these improvements depends on the accents involved and confidence levels vary between 60% and 99%.

As for the phone recognition case, AID accuracies indicate that BE and EE can be distinguished with much higher accuracy than AE and EE. Again, this agrees with the similarity analysis presented in Section 3.5. In general, word recognition accuracies are also lower for BE+EE compared to AE+EE. This is attributed mainly to the higher word LM perplexities of BE compared to the other two accents, as indicated in Table 7.2.

7.3.4 Per-Speaker AID

The word recognition performance of parallel AE+EE systems applying per-speaker AID is shown in Table 7.7, where the oracle results are unchanged from Table 7.5 and the accent-independent LM was used. A comparison between Tables 7.7 and 7.5 shows that, although per-speaker AID improves identification accuracy, it leads to deteriorated recognition performance compared to parallel systems performing per-utterance AID. The parallel results given in Table 7.7, however, are still marginally better than those achieved by the oracle systems. For BE+EE systems, on the other hand, per-speaker AID leads to perfect AID for both acoustic modelling approaches. Hence the BE+EE systems employing per-speaker AID achieve the performance of the oracle systems as indicated in Table 7.6, which represents an improvement over the per-utterance AID results.

7.4 Parallel Recognition of the Five Accents of SAE

Following the methodology described in Section 7.2, oracle and parallel speech recognition systems were developed using the combination of the AE, BE, CE, EE and IE training sets described in Section 3.3.

7.4.1 Language Models and Pronunciation Dictionaries

Accent-specific and accent-independent phone backoff bigram LMs, trained in a similar fashion to those described in Section 7.3.1, were used in phone recognition tests. Perplexities for these phone LMs are given in Table 7.8. For word recognition, accent-specific PDs (Section 5.2) were used with accent-specific word backoff bigram LMs (Section C.2), while a pooled accent-independent PD (Section 6.6) and an accent-independent word backoff bigram LM (Section C.3) were used together in contrasting systems. Word LM perplexities and OOV rates are indicated in Table 7.9.

Table 7.8: Phone bigram language model (LM) perplexities measured on the evaluation sets of all five SAE accents.

Accent	Phone bigram types	Accent-specific LM perp.	Accent-independent LM perp.
AE	1891	14.40	15.33
BE	1761	15.44	18.39
CE	1834	14.12	14.60
EE	1542	12.64	13.85
IE	1760	14.24	15.16

Table 7.9: Word bigram language model (LM) perplexities and OOV rates measured on the evaluation sets of all five SAE accents.

Accent	Word bigram types	Accent-specific LMs		Accent-independent LMs	
		Perplexity	OOV rate (%)	Perplexity	OOV rate (%)
AE	11 580	25.81	4.87	24.07	1.82
BE	9639	30.30	6.90	27.87	2.84
CE	10 641	30.87	5.24	27.45	1.40
EE	10 451	28.97	4.74	24.90	1.08
IE	11 677	26.22	5.09	25.55	1.73

7.4.2 Phone Recognition Experiments

Table 7.10 shows the average phone recognition and per-utterance AID accuracies measured on the evaluation sets for the five-accent systems developed according to the methodology described in Section 7.2. Per-utterance AID was performed for the parallel systems. Because a single recogniser was used for the systems employing both accent-independent acoustic and language models, identical results were obtained for the oracle and parallel tests. AID cannot be performed by these fully accent-independent systems.

The results in Table 7.10 indicate that for all three acoustic modelling approaches and for both oracle and parallel recognition, the systems employing accent-specific LMs outperform those

Table 7.10: Performance of five-accent oracle and parallel phone recognition systems when applying per-utterance AID. Average phone recognition accuracies (%) and per-utterance AID accuracies are indicated.

Model set	Accent-specific LMs			Accent-independent LM		
	Oracle	Parallel	AID (%)	Oracle	Parallel	AID (%)
Accent-specific	64.81	63.46	64.04	64.08	62.88	63.26
Accent-independent	65.60	64.55	37.98	63.93	63.93	-
Multi-accent	65.57	64.50	62.96	64.65	63.82	61.64

with accent-independent LMs in both recognition and identification. This corresponds to the phone LM perplexities given in Table 7.8 which indicate that the perplexities of the accent-independent LMs are higher than those of the accent-specific LMs. It is also consistent with the results presented in Sections 7.3.2 and C.1 in which accent-specific LMs were found to be superior to accent-independent LMs.

The results in Table 7.10 also indicate that in all cases the oracle systems outperform their parallel counterparts. Using bootstrap confidence interval estimation, these improvements were found to be statistically significant at at least the 99.9% level in all cases. This indicates that, if improved AID could be performed prior to recognition, phone recognition accuracies would improve. We shall see that the word recognition experiments show different trends, as was the case for the pair-wise AE+EE and BE+EE tests in Section 7.3.

7.4.3 Word Recognition Experiments

Table 7.11 shows the average word recognition and per-utterance AID accuracies measured on the evaluation sets for the five-accent systems developed according to the methodology described in Section 7.2. Per-utterance AID was performed for the parallel systems. The results in Table 7.11 indicate consistently superior performance for the systems employing accent-independent LMs compared to those employing accent-specific LMs. This is not only the case for recognition, but also for AID. Similar trends were observed for the AE+EE and BE+EE combinations in Section 7.3. As for those cases, Table 7.9 shows that the perplexities and OOV rates measured on the evaluation sets are in all cases higher for the accent-specific LMs than for the accent-independent LMs. As in Section 7.3, this is attributed to the very small amount of data available for LM training (Table 3.2) and also corresponds to the results of the additional experiments described in Section C.2. We therefore focus on systems using accent-independent LMs in the following comparison of the oracle and parallel recognition tests, although the accent-specific LM systems show similar trends. Although even the accent-independent LMs may be considered poorly trained, they are common to all recognition systems. Furthermore, the accent-independent PD is also common to all systems. Hence, the individual recognition systems used in the parallel recognition approaches are distinguished solely by their acoustic models.

The recognition results in Table 7.11 show that the parallel per-utterance AID system employing the accent-independent LM and the accent-specific acoustic models is outperformed by its corresponding oracle system. In contrast, the parallel recognition system employing multi-accent acoustic models show a small improvement over its oracle counterpart. Although this improvement is small and only significant at the 66% level, it is noteworthy that accent misclassifications do not lead to deteriorated system performance. Instead, the misclassifications improve overall recognition performance indicating that some test utterances are better matched to the multi-accent acoustic models of another accent. The results also indicate that the recognition performance of the multi-accent acoustic models is better than those achieved using accent-

Table 7.11: Performance of five-accent oracle and parallel word recognition systems when applying per-utterance AID. Average word recognition accuracies (%) and per-utterance AID accuracies are indicated.

Model set	Accent-specific LMs			Accent-independent LM		
	Oracle	Parallel	AID (%)	Oracle	Parallel	AID (%)
Accent-specific	75.34	77.89	66.66	81.53	81.31	67.60
Accent-independent	75.82	80.20	44.56	81.67	81.67	-
Multi-accent	76.34	80.15	64.88	82.78	82.85	65.39

Table 7.12: Performance of five-accent oracle and parallel word recognition systems when applying per-speaker AID. Average word recognition accuracies (%) and per-utterance AID accuracies are indicated for systems employing the accent-independent LM.

Model set	Oracle	Parallel	AID (%)
Accent-specific	81.53	81.69	89.84
Multi-accent	82.78	82.85	89.81

specific and accent-independent models. The improvements of the parallel per-utterance AID system employing multi-accent acoustic models over the systems employing accent-specific and accent-independent acoustic models are statistically significant at the 99.9% level in both cases.

7.4.4 Per-Speaker AID

The word recognition performance of parallel systems employing the accent-independent LM and applying per-speaker AID is shown in Table 7.12, where the oracle results are identical to those given in Table 7.11. A comparison between Tables 7.12 and 7.11 indicate that per-speaker AID improves recognition accuracy for the systems using accent-specific acoustic models (outperforming the accent-specific oracle system) while recognition accuracy is unchanged for the multi-accent systems. In both cases identification accuracies are substantially improved. Again, multi-accent acoustic modelling performs best. The improvements of the per-speaker AID multi-accent system over the per-speaker AID system employing accent-specific acoustic models and over the accent-independent system were calculated to be statistically significant at the 99.9% level.

In summary, we are led to the surprising conclusion that superior AID prior to accent-specific speech recognition does not necessarily lead to superior speech recognition accuracy. This is supported by both the per-utterance and per-speaker AID word recognition experiments. Furthermore, from our comparison of acoustic modelling approaches it is apparent that it is better to employ parallel speech recognisers with multi-accent acoustic models than to pool accent-specific acoustic training data and use the resulting accent-independent acoustic models. Our experiments considering per-speaker AID indicate that, when employing multi-accent acoustic models, recognition accuracy is neither improved nor impaired compared to per-utterance AID despite a significant increase in AID accuracy. Small improvements are observed for the parallel system employing accent-specific acoustic models and per-speaker AID over its oracle and parallel per-utterance AID counterparts.

7.4.5 Analysis of Accent Misclassifications

In order to investigate the implicit AID taking place in the parallel recognition systems and the resulting improvement or deterioration compared to oracle recognition, we present an analysis

of the misclassifications taking place in the per-utterance AID accent-independent LM word recognition systems employing accent-specific and multi-accent acoustic models.

A misclassification occurs when the accent of the recogniser selected for a particular utterance during parallel recognition is different from the accent with which that utterance is labelled. The procedure for our analysis is as follows. For each misclassified utterance, the transcription from the corresponding oracle system as well as the transcription from the recogniser selected during parallel recognition is obtained. Both these transcriptions are aligned to the reference transcription. By comparing the two alignments, we can determine whether the misclassification results in an improvement, in a deterioration, or in no change in performance for the parallel recognition system relative to the performance of the oracle system. The resulting effect on recognition performance can also be calculated as follows.

A speech recogniser is evaluated by aligning output transcriptions to reference transcriptions using dynamic programming and then calculating the recognition accuracy as [75, p. 205]:

$$\text{Percent Accuracy} = \frac{N - D - S - I}{N} \times 100\% \tag{7.1}$$

where N is the total number of tokens in the reference transcriptions, D the number of deletion errors, S the number of substitution errors and I the number of insertion errors. By distinguishing tokens, deletions, substitutions and insertions associated with utterances which are misclassified from those associated with utterances which are correctly identified, one can calculate the "contribution" of the misclassified and correctly identified utterances to the overall accuracy of a parallel recognition system:

$$\begin{aligned}
\text{Percent Accuracy} &= \frac{N - D - S - I}{N} \times 100\% \\
&= \frac{N_{\text{misclassified}} - D_{\text{misclassified}} - S_{\text{misclassified}} - I_{\text{misclassified}}}{N} \times 100\% \\
&\quad + \frac{N_{\text{correct}} - D_{\text{correct}} - S_{\text{correct}} - I_{\text{correct}}}{N} \times 100\% \\
&= \text{Contrib}_{\text{misclassified}} + \text{Contrib}_{\text{correct}}
\end{aligned} \tag{7.2}$$

By further distinguishing between misclassifications which result in no change, in improvements and in deterioration, the first term in the final line of equation (7.2) can be written as[3]

$$\text{Contrib}_{\text{misclassified}} = \text{Contrib}_{\text{no-effect}} + \text{Contrib}_{\text{improve}} + \text{Contrib}_{\text{worse}} \tag{7.3}$$

For each of the consequences of misclassification, these contributions were calculated as part of the analyses presented in Tables 7.13 and 7.14 for the parallel accent-specific and multi-accent systems, respectively. We also calculated the contributions to the overall accuracy which would have resulted if the oracle system output was used instead of the parallel system output for the misclassified utterances. These contributions as well as the resulting effect of using the parallel systems instead of the oracle systems are indicated in the last two columns of Tables 7.13 and 7.14.

Tables 7.13 and 7.14 indicate that for both acoustic modelling approaches a considerable number of misclassified utterances have no effect on performance, with the minority of misclassifications leading to improvements or deterioration in accuracy for the parallel systems relative to the performance of the oracle systems. Although the overall average length of misclassifications is shorter than the overall average length of the evaluation set utterances (which is approximately 2

[3] Although recognition accuracy is not a probability but rather a rate, these contributions could analogously be seen as probability mass carried by the different types of utterances.

Table 7.13: Analysis for the parallel per-utterance AID word recognition system employing the accent-independent LM and accent-specific acoustic models, indicating the contribution (contrib.) and effect on recognition accuracy (acc.) due to misclassifications.

Misclassification consequence	No. of utterances	No. of tokens	Average length (s)	Contrib. to acc. (%)	Contrib. if oracle (%)	Δcontrib. (%)
No effect	905	2721	1.42	14.18	14.18	0
Improved acc.	135	802	2.66	4.28	2.99	+1.29
Deteriorated acc.	162	773	2.22	2.54	4.05	−1.52
Total/Average†	1202	4296	1.67†	20.99	21.22	−0.23

Table 7.14: Analysis for the parallel per-utterance AID word recognition system employing the accent-independent LM and multi-accent acoustic models, indicating the contribution (contrib.) effect on recognition accuracy (acc.) due to misclassifications.

Misclassification consequence	No. of utterances	No. of tokens	Average length (s)	Contrib. to acc. (%)	Contrib. if oracle (%)	Δcontrib. (%)
No effect	1040	3213	1.46	17.14	17.14	0
Improved acc.	120	705	2.66	3.60	2.41	+1.19
Deteriorated acc.	124	559	2.05	1.67	2.79	−1.12
Total/Average†	1284	4477	1.63†	22.41	22.34	+0.07

seconds) we observe that the average length of utterances leading to improved accuracy (approximately 2.7 seconds) is longer. Compared to the evaluation set, utterances leading to deteriorated accuracy (approximately 2.1 seconds) are similar in length and utterances resulting in no effect are shorter (approximately 1.4 seconds) for both acoustic modelling approaches.

The results in Table 7.11 indicate that the parallel accent-specific system (with an accuracy of 81.31%) is slightly outperformed by the corresponding oracle system (81.53% accuracy). The analysis in Table 7.13 for the accent-specific system indicates that there are significantly more misclassifications leading to deterioration in accuracy than misclassifications leading to improvement. Although this is the case, the misclassified utterances resulting in improvements are longer and the analysis shows that the number of tokens involved in improved and deteriorated performance is approximately equal. The analysis indicates that the 1.29% absolute improvement due to misclassifications is outweighed by the 1.52% decrease in accuracy due to misclassifications leading to a deterioration. This ultimately results in the 0.23% absolute drop in performance when performing parallel recognition.

For the multi-accent acoustic modelling approach, the results in Table 7.11 indicate that parallel recognition (82.85% accuracy) leads to improved recognition performance compared to oracle recognition (82.78% accuracy). The analysis in Table 7.14 indicate that, although the number of misclassifications leading to deterioration and those leading to improvements are approximately equal, the number of tokens involved in improved performance is greater. This corresponds to the longer average length of the utterances yielding improvements. The accuracy contribution analysis indicates that the effect of misclassifications leading to improvements (+1.19%) is slightly greater than for misclassifications leading to deterioration (−1.12%) and the overall result is a very slight improvement of 0.07% absolute in word recognition accuracy.

Table 7.15 shows the AID confusion matrix for the parallel multi-accent system employing per-utterance AID and the accent-independent LM. The general trends in the table indicate that AE and CE as well as AE and EE utterances are confused most often. IE and CE are also often confused. The diagonal of Table 7.15 indicates that AE, CE and EE utterances are

misclassified most often, IE are misclassified less often, and BE are identified correctly most often. This analysis corresponds to the general trends observed in the similarity analysis presented in Section 3.5 as well as the experimental results and conclusions in Chapter 6 which highlighted the similarity of AE, CE and EE while BE and IE were found to be more different. The AID confusion matrix for the parallel accent-specific system employing per-utterance AID and the accent-independent LM indicates similar trends.

Table 7.15: AID confusion matrix for the parallel per-utterance AID word recognition system employing the accent-independent LM and multi-accent acoustic models, with confusions indicated in percentages.

		Hypothesised accent				
		AE	BE	CE	EE	IE
Actual accent	AE	62.41	3.48	17.42	11.32	5.37
	BE	4.16	78.26	7.38	3.36	6.85
	CE	16.50	6.77	53.88	8.60	14.25
	EE	21.94	3.70	7.27	58.83	8.26
	IE	2.43	7.40	10.98	7.75	71.45

7.5 Summary and Conclusions

In this chapter we evaluated the speech recognition performance of systems employing parallel accent-specific recognisers for three combinations of SAE accents: AE+EE, BE+EE and all five SAE accents. In order to determine the effect of misclassifications in the accent identification (AID) process that occurs implicitly during parallel recognition, the performance of these systems was compared to the performance of oracle systems in which test utterances were presented to matching accent-specific recognisers. The performance of parallel recognition systems employing accent-specific and multi-accent acoustic models was also compared to accent-independent speech recognition achieved by pooling acoustic and language model training data.

In phone recognition experiments, oracle systems outperformed parallel recognition systems for all of the accent combinations and acoustic and language modelling approaches considered. Different trends were observed in word recognition experiments. The results in Section 7.3 for the AE+EE and BE+EE accent pairs indicated that the observed improvement or deterioration in word recognition performance of multiple accent-specific systems running in parallel relative to an oracle system depends on the similarity of the accents involved. Word recognition experiments demonstrated that, despite AID errors, parallel systems performing implicit per-utterance AID slightly outperformed oracle systems for the AE+EE configuration. This was not the case for the BE+EE accent pair. However, parallel systems based on per-speaker AID showed oracle or better word recognition performance for both accent pairs.

In the word recognition experiments of the five-accent combination described in Section 7.4, the parallel system performing implicit per-utterance AID and employing accent-specific acoustic models were outperformed by the corresponding oracle recognition system. In contrast, the parallel system employing multi-accent acoustic models resulted in slightly superior performance compared to the corresponding oracle system. This was the case for the multi-accent word recognition system applying per-utterance AID as well as for the multi-accent system applying per-speaker AID. The former achieved an AID accuracy of 65.39% while the latter achieved an AID accuracy of 89.81%. Both these systems achieved a word recognition accuracy of 82.85% which represents a very slight improvement relative to the accuracy of 82.78% achieved by the oracle system. The parallel accent-specific system performing per-speaker AID also outperformed its oracle counterpart. Whether or not per-speaker AID can be employed will be determined

by the practical speech recognition setup. We conclude that AID errors made during parallel recognition do not necessarily lead to deteriorated speech recognition accuracy and may in fact lead to slight improvements. Furthermore, we speculate that such improvements are possible for similar accents but are less likely for accents that differ widely from each other. Analysis of the type of misclassifications which occur during parallel recognition indicated that the number of misclassifications leading to improvements is very similar to the number leading to a deterioration, and that it is only with a small margin that the overall performance of the five-accent multi-accent word recognition system outperforms its oracle counterpart.

Of the three acoustic modelling approaches considered, multi-accent modelling, which supports selective cross-accent sharing of acoustic training data, yielded superior or comparable performance to the other two approaches in both phone and word recognition. In word recognition experiments, parallel systems employing multi-accent acoustic models (achieving a word recognition accuracy of 82.85% for the five-accent combination) outperformed systems employing accent-independent acoustic models obtained by pooling acoustic and language model training data across accents (which achieved a word recognition accuracy of 81.67% for the five-accent combination). This was the case for all three accent combinations (AE+EE, BE+EE and the five-accent combination) and was achieved by parallel systems performing either implicit per-utterance or per-speaker AID. For the five-accent case, these improvements, as well as the improvements obtained over parallel systems employing accent-specific acoustic models (word recognition accuracy of 81.31%), were calculated to be statistically significant at the 99.9% level.

Chapter 8

Accent Reclassification Experiments

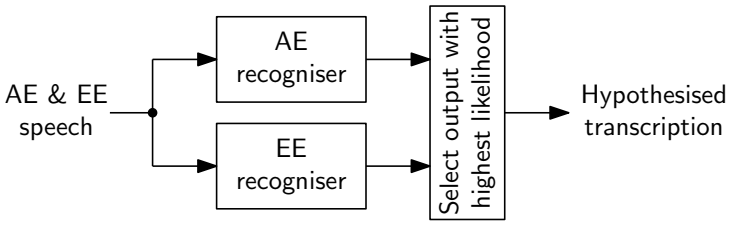
In order to train multi-accent speech recognition systems, accent labels must be assigned to training set utterances. For the AST databases, these accent labels were assigned according to the speaker's mother-tongue and ethnicity (Section 3.1). However, the experiments presented in Chapter 7 indicated that the accent labels assigned to some utterances in the AST databases might be inappropriate. In this chapter we consider the iterative accent reclassification of the training set utterances in an attempt to improve the labelling consistency of the training set. The speech recognition performance of models trained on the reclassified data is compared with that of models trained on the original data, as well as with systems in which training data are pooled across accents. Two acoustic modelling approaches are considered for the reclassification configuration: accent-specific and multi-accent acoustic modelling. We investigate accent reclassification for two pairs of SAE accents: AE+EE and BE+EE. As observed in the previous chapters (e.g. Sections 3.5 and 6.4, Figure 6.11), the former pair represents two relatively similar accents while the latter pair represents two accents which are more dissimilar.

8.1 Accent Reclassification

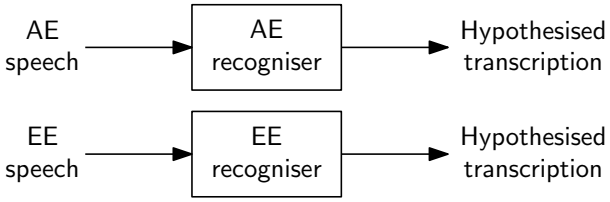
As part of the experiments described in Section 7.3.3 we considered word recognition of AE and EE using a system of two accent-specific recognisers operating in parallel, as illustrated in Figure 8.1(a). It was shown that this configuration slightly outperformed an oracle setup in which accented speech was presented to the matching accent-specific recogniser, illustrated in Figure 8.1(b). The finding that configuration (a) outperforms configuration (b) indicates that accent misclassifications do not always lead to deteriorated speech recognition accuracies. Instead, in some cases a different accent's recogniser produces a better accuracy than the recogniser of the correct accent. It appears then that the accent to which an utterance has been consigned in the training/test data is not always the most appropriate. In the light of these results we proceed to reclassify the accent of each utterance in the training set using a set of first-pass acoustic models obtained using the original databases, and then to retrain the acoustic models using these newly assigned accent classifications.

This approach is illustrated in Figure 8.2. Using the unmodified training data, initial accent-specific[1] HMMs are obtained. These models are then used to reclassify the training data. Transcriptions reflecting the new accent classifications are subsequently used to train new accent-

[1] As mentioned in Section 7.2, 'accent-specific' used in this context could refer to a system employing either accent-specific or multi-accent acoustic models since both are dependent on accent.



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



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

Figure 8.1: The two recognition configurations considered in Section 7.3 for recognition of AE and EE.

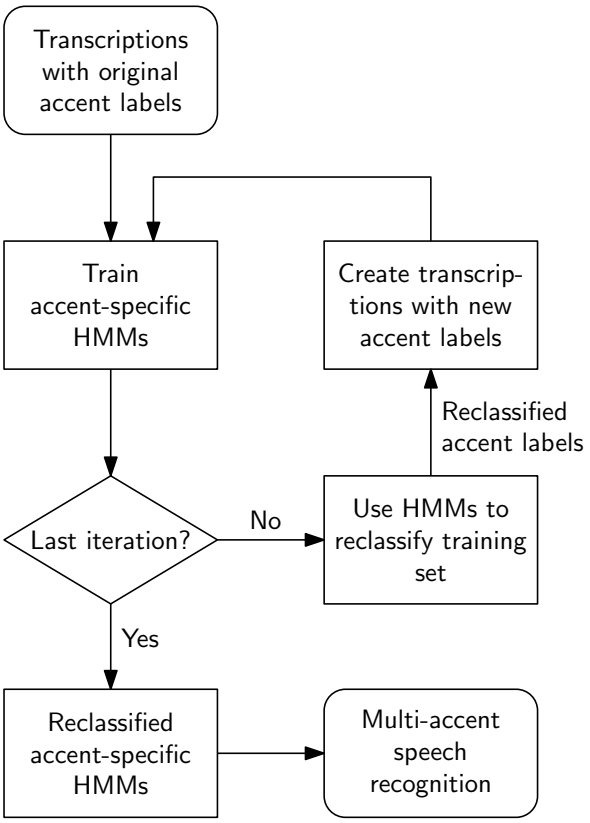


Figure 8.2: Reclassification of training data and subsequent retraining of acoustic models.

specific HMMs. Multiple iterations of reclassification can be performed in this manner, although we only performed a single iteration. The proposed reclassification approach aims to compensate, during training, for the inexact assignment of accent labels to some utterances in the original training data. Note that this approach is computationally expensive, because it requires the entire training set to be recognised in a parallel configuration.

8.2 System Optimisation, Configuration and Objectives

For the experiments described in this chapter, only word recognition was considered since the labelling inconsistency described above was observed only in the relative performance of oracle and parallel word recognition systems. The AE+EE and BE+EE acoustic model sets used in Sections 6.2.3, 6.3.3 and 7.3.3 were used as the first-pass HMMs for reclassification. Sections 5.3.3 and 7.2 describe the optimisation of these models in more detail. The same decision-tree likelihood improvement thresholds used to cluster the first-pass models were used for the reclassified models. As in Sections 6.2.3, 6.3.3 and 7.3.3, the AE+EE and BE+EE accent-independent word language models (LMs) trained on the combined training sets of all five SAE accents were used in combination with the accent-independent pronunciation dictionaries (PDs) obtained by pooling the pronunciations in the available accent-specific PDs of the two corresponding accents. LM perplexities and out-of-vocabulary (OOV) rates are given in Table 7.1.

We performed word recognition experiments for the AE+EE and BE+EE accent pairs. Using the accent-specific and multi-accent acoustic modelling approaches described in Section 4.2, we trained reclassified models using the approach described in Section 8.1 and employed these models in parallel during recognition. Reclassification and parallel recognition was performed on a per-utterance basis. As a baseline, we considered models trained on the unmodified training data using the same two acoustic modelling approaches – these are the accent-specific and multi-accent acoustic models used in Sections 6.2.3, 6.3.3 and 7.3.3 and the first-pass models used for the experiments described here. Systems employing these models were used to perform both oracle and parallel recognition. These two configurations are illustrated in Figure 8.1 for the AE+EE pair. As a further benchmark we considered the performance of the AE+EE and BE+EE accent-independent recognition systems also evaluated in Sections 6.2.3, 6.3.3 and 7.3.3. Since these systems employ accent-independent acoustic models, LMs and PDs, accent identification (AID) is not possible and reclassification cannot be performed. However, these systems represent an important baseline. Parallel recognition systems perform AID implicitly and these accuracies can also be measured. These accuracies are calculated relative to the originally assigned accent labels and are therefore not relevant to the evaluation of the reclassified systems.

The chief aim of this chapter is to determine whether the reclassified systems can improve on parallel recognition performance compared to systems trained on the originally labelled data. By performing these experiments in pairs, we are considering one scenario where accents are quite similar (AE+EE) and a second scenario where accents are relatively different (BE+EE).

8.3 Experimental Results

Using the combination of the AE and EE as well as the BE and EE training sets described in Section 3.3, we performed word recognition experiments using the systems described in Section 8.2. Tables 8.1 and 8.2 respectively show the AE+EE and BE+EE average word recognition and AID accuracies measured on the corresponding evaluation sets. The performance of the systems trained on the original data are identical to those given in Tables 7.5 and Tables 7.6. Oracle performance is indicated for these systems, but is not relevant to the reclassified systems. Because a

Table 8.1: Performance of AE+EE systems employing HMMs trained on the original databases, as well as systems employing reclassified HMMs. Word recognition accuracies (%) and AID accuracies are given.

Model set	Original HMMs			Reclassified HMMs	
	Oracle	Parallel	AID (%)	Parallel	AID (%)
Accent-specific	84.01	84.63	80.23	84.58	78.07
Accent-independent	84.78	84.78	-	-	-
Multi-accent	84.78	84.88	78.22	84.61	76.64

Table 8.2: Performance of BE+EE systems employing HMMs trained on the original databases, as well as systems employing reclassified HMMs. Word recognition accuracies (%) and AID accuracies are given.

Model set	Original HMMs			Reclassified HMMs	
	Oracle	Parallel	AID (%)	Parallel	AID (%)
Accent-specific	76.69	76.07	93.23	75.86	93.37
Accent-independent	75.38	75.38	-	-	-
Multi-accent	77.35	76.75	93.16	76.60	92.40

Table 8.3: Accuracy difference (%) and the corresponding significance (sig.) levels (%), calculated by using bootstrap confidence interval estimation [80], of the superior performance of the original systems over the corresponding reclassified parallel systems.

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

single recogniser is used for the systems employing accent-independent models, identical results are obtained for the oracle and parallel tests. AID and accent reclassification is not possible with these fully accent-independent systems.

The results in Table 8.1 indicate that, as observed in Section 7.3.3, the AE+EE parallel systems employing accent-specific and multi-accent acoustic models show small improvements over the corresponding oracle systems. In contrast the results in Table 8.2 indicate deteriorated performance for the BE+EE pair when using the original parallel systems compared to oracle recognition. The results also indicate, as noted in Section 7.3.3, that the recognition performance of the original systems employing multi-accent acoustic models is better than that achieved by the original systems employing accent-specific and accent-independent acoustic models for both accent pairs.

When comparing the performance of the original and reclassified parallel recognition systems given in Tables 8.1 and 8.2, deterioration in system performance is observed for both accent-specific and multi-accent acoustic modelling approaches and for both accent pairs. Except for the BE+EE accent-specific systems, the AID accuracy of the reclassified systems is lower than that of the corresponding original systems in all other cases, as one might expect. Using bootstrap confidence interval estimation [80], the statistical significance levels of the improved performance of the original parallel systems over the reclassified systems were calculated and are shown in Table 8.3. It is evident that the differences are significant only at low levels varying between 56% and 80%. Nevertheless, the deterioration in performance after a single iteration of reclassification is consistent across all the considered accent pairs and acoustic modelling approaches. The AE+EE reclassified systems also show deteriorated performance in comparison to the accent-

independent system, while the BE+EE reclassified systems still show superior performance.

For the experiments presented above, per-utterance AID was applied for both the accent reclassification and parallel recognition procedures. An alternative to per-utterance AID is per-speaker AID in which the selection of the highest recogniser score is done independently for each speaker. We also considered experiments where reclassification was attempted using per-speaker AID. In this case, however, the reclassified training set accent assignments matched the original assignments for all but one speaker (an AE speaker was reclassified as EE by the first-pass multi-accent system). Reclassification using per-speaker AID was therefore not considered further.

Except for the AE+EE and BE+EE accent pairs, we considered additional accent combinations in the preceding chapters. However, the reclassification procedure did not yield any improvements, even for the AE+EE accent pair which represent relatively similar accents. Due to this and the very high computational complexity of the reclassification process, it was decided not to consider additional accent combinations.

8.4 Analysis and Discussion

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

In order to obtain some insight into the somewhat surprising results, we analysed utterances in the training set for which the original and the reclassified accent labels differ. We performed this analysis for the AE+EE multi-accent system and the results are presented in Table 8.4. The analysis indicates that the utterances for which the original accent labels had been changed are generally shorter (1.07 seconds) than both the overall average (2.20 seconds) and the average length of utterances for which accent labels were unchanged (2.28 seconds). Furthermore, the number of original AE utterances reclassified as EE utterances is approximately double the number of EE utterances reclassified as AE. In general, the English proficiency of AE speakers is high [81], which might suggest that some of the AE speakers are simply better matched to the models trained on the EE data and this may explain why AE to EE relabelling is performed more often than the opposite.

Table 8.5 shows an analysis of evaluation set utterances for which the accent classification according to the original AE+EE parallel multi-accent system (84.88% accuracy, Table 8.1) and the corresponding reclassified system (84.61%, Table 8.1) differs, i.e. utterances for which the accent of the accent-specific recogniser selected during recognition is different for the original and reclassified systems. Table 8.5 indicates that, again, the utterances for which classification has changed tend to be shorter with an average length of 1.53 seconds compared to both the overall average of 2.08 seconds and the average length of 2.14 seconds of test utterances for which accent classification was unchanged. The breakup of system accuracy indicates that the overall drop of 0.27% absolute in accuracy is mainly due to deteriorated performance on the utterances for which accent classification was unchanged (85.54% compared to 85.08% word recognition accuracy). Performance on the utterances for which classification had changed indicates a 1.91% absolute improvement in word recognition performance. Rows three and four in Table 8.5 show that this

Table 8.4: Analysis of training set utterances for which the original and the reclassified accent labels differ according to the AE+EE multi-accent system. The 'labels changed' row is the combination of the two rows that follow.

Reclassification effect	Number of utterances	Number of tokens	Average length (s)
Labels unchanged	19 775	96 488	2.28
Labels changed:	1447	3331	1.07
AE → EE	942	2251	1.11
EE → AE	505	1080	1.00
Total/average†	21 222	99 819	2.20†

Table 8.5: Analysis of the change in recogniser selection between the original and reclassified AE+EE multi-accent systems. The 'changed' row is the combination of the two rows that follow.

Recogniser selection	Number of utterances	Average length (s)	Original accuracy (%)	Reclassified accuracy (%)
Unchanged	1241	2.14	85.54	85.08
Changed:	150	1.53	77.19	79.10
AE → EE	63	1.39	74.21	80.00
EE → AE	87	1.63	79.21	78.50
Overall	1391	2.08	84.88	84.61

is the result of superior performance on utterances which previously had been classified as AE but that were identified as EE by the reclassified system.

While Tables 8.4 and 8.5 analyse only the accent classifications and performance of the AE+EE multi-accent systems, the same analysis on the AE+EE accent-specific systems indicates similar trends. Analysis of the BE+EE systems also indicates that the training set utterances for which the manual and automatically derived accent labels differ, as well as the test utterances for which the original and reclassified systems' accent classification are inconsistent, tend to be shorter. However, for the BE+EE case fewer training utterances are relabelled (only approximately 450 out of the total 17 657) and the number of training utterance label changes from BE to EE and vice versa are approximately equal. The original and reclassified systems are also more consistent in test utterance accent assignment and fewer classification changes occur compared to the AE+EE case.

8.5 Summary and Conclusions

In this chapter we evaluated the effect of unsupervised accent reclassification of training set utterances on speech recognition performance. Experiments were performed for two pairs of SAE accents: AE+EE and BE+EE. By classifying the accent of each utterance in the training set using first-pass acoustic models trained on the original databases and then retraining the models based on the new accent labels, reclassified acoustic models were obtained (Section 8.1). Two acoustic modelling approaches were considered for this procedure: accent-specific acoustic modelling and multi-accent acoustic modelling. Systems employing reclassified models were compared with systems employing the original models and with accent-independent systems for which training data were pooled. The reclassified models showed consistently deteriorated performance compared to the original models in parallel speech recognition experiments for both accent pairs and all acoustic modelling approaches considered (Section 8.3). Further analysis

(Section 8.4) indicated that the training utterances for which manual and automatically derived labels differ tend to be shorter. Fewer utterances were relabelled for BE+EE than for AE+EE. For the latter accent pair, accent label changes from AE to EE occurred much more often than from EE to AE. Analysis of evaluation set utterance for which the accent classification according to the original and the reclassified systems differed, indicated that test utterances for which classification had changed also tend to be shorter. Furthermore, the analysis indicated that the AE+EE reclassified systems improved accuracy for test utterances which were classified as EE by the reclassified systems, but were previously classified as AE in the original parallel systems. However, recognition performance was poorer for the utterances for which classification was unchanged which ultimately resulted in deteriorated performance. We conclude that the proposed relabelling procedure does not lead to performance improvements and that the best strategy remains the use of the originally labelled training data.

Chapter 9

Summary and Conclusions

In this thesis we investigated different acoustic modelling approaches and system configurations for speech recognition of 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). Our experiments were based on the African Speech Technology (AST) databases which include five databases corresponding to these five accents. In this chapter we give a summary of the presented study, highlight our contributions, and discuss further avenues of research for future work.

9.1 Acoustic Modelling Approaches

Three acoustic modelling approaches formed the foundation of this research: (i) accent-specific acoustic modelling, in which separate acoustic model sets are trained for each accent; (ii) accent-independent acoustic modelling, which simply involves pooling accented data across accents; and (iii) multi-accent acoustic modelling, which enables selective cross-accent data sharing. For the latter, selective sharing is enabled by extending the decision-tree state clustering process normally used to construct tied-state HMMs by allowing accent-based questions. These three modelling approaches were evaluated and compared in different recognition configurations.

In the first stage of our research we set out to determine which of the modelling approaches yields superior performance when applied to accents of SAE. In this case, each test utterance was presented only to the recogniser tailored to the matching accent. This allowed us to focus on acoustic modelling and, in particular, to avoid the effects that accent misclassifications might have on recognition performance.

For the relatively similar AE+EE accent pair, multi-accent and accent-independent acoustic modelling resulted in similar improvements over accent-specific modelling. For the more dissimilar BE+EE accent pair, however, multi-accent and accent-specific acoustic modelling resulted in similar improvements over accent-independent modelling. From these results we conclude that the decision of whether to pool or to separate training data depends on the particular accents in question. Subsequent experiments with the three-accent AE+BE+EE combination demonstrated that multi-accent acoustic modelling allows this decision to be made in a data-driven manner. For this combination of accents, multi-accent modelling outperformed both accent-specific and accent-independent modelling. The same behaviour was observed in experiments that included all five SAE accents. For the five-accent combination, multi-accent models yielded a word recognition accuracy of 82.78%, which represents an improvement of 1.25% absolute in terms of word recognition accuracy compared to accent-specific and accent-independent models. These improvements were found to be statistically significant at the 99.9% level. Overall, we

conclude that multi-accent modelling is an effective strategy for the acoustic modelling of multiple accents, and leads to statistically significant speech recognition improvements when applied to the five accents of SAE.

9.2 Parallel Recognition of Multiple SAE Accents

In the second stage of our research, we considered the impact which accent misclassifications have on speech recognition accuracy in the scenario where the accent of each test utterance is unknown. In a configuration which we referred to as 'parallel recognition', each test utterance was presented simultaneously to a bank of recognisers, one for each accent. The recognition hypothesis with the highest associated likelihood was then chosen as the final result. We compared this configuration to 'oracle recognition', in which the accent of each test utterance is assumed to be known and each utterance is presented only to the matching model set. This last approach was the approach followed in the first stage of our research.

In word recognition experiments for the AE+EE accent pair, parallel systems slightly outperformed oracle systems while for the BE+EE pair, the opposite was observed. In analogous experiments performed for all five SAE accents, a parallel system employing multi-accent models was shown to very slightly outperform its oracle counterpart. The former exhibited a word recognition accuracy of 82.85% and an implicit accent identification accuracy of 65.39%. Furthermore, the parallel system employing multi-accent acoustic models was shown to outperform a corresponding systems employing accent-independent models obtained by cross-accent pooling of acoustic and language model training data, as well as a parallel system employing accent-specific acoustic models.

We conclude that accent misclassifications occurring during parallel recognition do not necessarily lead to deteriorated speech recognition performance and that multi-accent models are particularly effective in this regard. This is a very important and encouraging conclusion from the perspective of practical system implementation, since the accent of the incoming speech will usually not be known. A secondary conclusion is that, once again, multi-accent acoustic modelling leads to the best performing systems in comparison with the other two approaches. In particular, the improvements of parallel multi-accent systems over accent-independent systems are important since, for the latter, a single model set is used irrespective of the input speech accent.

9.3 Accent Reclassification

The result of the parallel recognition experiments raised the suspicion that the accent labels assigned to some training/test utterances might be inappropriate. This led to the final stage of our reseach, in which we considered the unsupervised reclassification of the training set accent labels. By classifying the accent of each utterance in the training set using first-pass acoustic models trained on the original databases and then retraining the models, reclassified acoustic models were obtained. Reclassification of the AE+EE and BE+EE accent pairs was considered using the accent-specific and multi-accent acoustic modelling approaches.

In parallel word recognition experiments, the reclassified models failed to improve on the performance of models trained on the original data. Further analysis indicated that the AE+EE reclassified systems improved accuracy for test utterances which were classified as EE by the reclassified systems, but were previously classified as AE in the original parallel systems. However, recognition performance was poorer for the utterances for which classification was unchanged,

and this ultimately resulted in deteriorated overall performance. We conclude that the proposed relabelling procedure does not lead to any improvements and that training acoustic models on the originally labelled data remains the best approach.

9.4 Contributions

We would like to highlight the following contributions of this work:

- To our knowledge, we are the first to consider multi-accent acoustic modelling of multiple accents in a tied-state HMM system topology.
- This work is the first thorough comparison of accent-specific, accent-independent and multi-accent acoustic modelling for phone and word recognition of accents of SAE.
- Although some authors have compared oracle and parallel recognition (Section 2.6), many authors fail to distinguish between these recognition configurations. A thorough analysis of the effects of accent misclassifications on recognition performance could not be found in the literature. We present such an analysis in Chapter 7.
- Although the proposed accent reclassification approach is somewhat specific to the AST data, this procedure has not been considered elsewhere.

9.5 Further Work

Further avenues of research that can be explored include:

- The performance of accent-dependent or speaker-dependent MAP [31, 32] or MLLR [33] adaptation (Section 2.3.2) or speaker adaptive training [84] can be compared with multi-accent acoustic modelling. One can investigate whether and how multi-accent acoustic modelling can be used in combination with these adaptation techniques.
- The behaviour of the different acoustic modelling approaches can be investigated as the amount of training data is reduced in order to determine the dependency of our results on training set size.
- The three acoustic modelling approaches can be applied to a corpus of South African broadcast news recently collected at Stellenbosch University in order to determine if the findings for the AST databases also hold for this corpus.
- As mentioned in Section 2.5, Caballero et al. [51, 52] proposed the use of one-root decision-trees in contrast to the traditional approach of growing separate decision-trees for each basephone. By applying the former, data can be shared across context-dependent phones with different basephones. This approach can also be evaluated in our SAE systems.
- The reclassification procedure proposed in Chapter 8 can be compared to more traditional speaker clustering algorithms.
- In the proposed reclassification procedure, very few training set utterances are actually relabelled. This is to be expected since the data used for training models were also presented to these models for classification. Accent identification accuracy is therefore very high. An alternative would be to split the training set into non-overlapping parts and then classify one part using models trained on the other parts.

9.6 Overall Summary and Conclusions

In this thesis we demonstrated that multi-accent acoustic modelling, which enables selective cross-accent data sharing by allowing accent-based questions in the decision-tree state clustering process, outperforms accent-specific (i.e. separate) and accent-independent (i.e pooled) acoustic modelling for the five accents of South African English. This conclusion holds both for systems in which each accented test utterance is presented only to the recogniser tailored to that accent, and for systems in which several recognisers tailored to different accents are employed in parallel. For the latter scenario, we found that misclassifications made during the implicit accent identification process did not lead to deteriorated performance. We conclude that, for the speech databases and accents considered, inclusion of accent-based questions in the multi-accent acoustic modelling approach yields statistically significant improvements over the accent-specific and accent-independent modelling approaches. Furthermore, we conclude that when the multi-accent acoustic models are applied in parallel to allow speech recognition without prior knowledge of the incoming accent, the implicit accent identification errors made by the system do not lead to deteriorated performance.

References

[1] Statistics South Africa, "Census 2001: primary tables South Africa: census 1996 and 2001 compared," 2004.

[2] D. van Compernolle, J. Smolders, P. Jaspers, and T. Hellemans, "Speaker clustering for dialectic robustness in speaker independent recognition," in *Proc. Eurospeech*, Genova, Italy, 1991, pp. 723–726.

[3] C. Teixeira, I. Trancoso, and A. Serralheiro, "Recognition of non-native accents," in *Proc. Eurospeech*, Rhodes, Greece, 1997, pp. 2375–2378.

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

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

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

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

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

[9] ——, "Accent reclassification and speech recognition of Afrikaans, Black and White South African English," in *Proc. PRASA*, Johannesburg, South Africa, 2011, pp. 80–84.

[10] D. van Compernolle, "Recognizing speech of goats, wolves, sheep and ... non-natives," *Speech Commun.*, vol. 35, no. 1–2, pp. 71–79, 2001.

[11] S. Steidl, G. Stemmer, C. Hacker, and E. Nöth, "Adaptation in the pronunciation space for non-native speech recognition," in *Proc. Interspeech – ICSLP*, Jeju Island, Korea, 2004, pp. 2901–2904.

[12] D. Crystal, *A Dictionary of Linguistics and Phonetics*, 3rd ed. Oxford, UK: Blackwell Publishers, 1991.

[13] J. C. Wells, *Accents of English*. Cambridge, UK: Cambridge University Press, 1982, vol. 1.

[14] U. Uebler and M. Boros, "Recognition of non-native German speech with multilingual recognizers," in *Proc. Eurospeech*, Budapest, Hungary, 1999, pp. 911–914.

[15] J. J. Humphries, P. C. Woodland, and D. Pearce, "Using accent-specific pronunciation modelling for robust speech recognition," in *Proc. ICSLP*, Philadelphia, PA, 1996, pp. 2324–2327.

[16] L. M. Tomokiyo, "Recognizing non-native speech: characterizing and adapting to non-native usage in LVCSR," Ph.D. dissertation, Carnegie Mellon University, Pittsburgh, PA, 2001.

[17] A. Waibel, P. Geutner, L. M. Tomokiyo, T. Schultz, and M. Woszczyna, "Multilinguality in speech and spoken language systems," in *Proc. IEEE*, vol. 88, no. 8, 2000, pp. 1297–1313.

[18] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens, "Automatic speech recognition and speech variability: a review," *Speech Commun.*, vol. 49, no. 10–11, pp. 763–786, 2007.

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

[20] ——, "The use of accent-specific pronunciation dictionaries in acoustic model training," in *Proc. ICASSP*, Seattle, WA, 1998, pp. 317–320.

[21] N. Beringer, F. Schiel, and P. Regel-Brietzmann, "German regional variants – a problem for automatic speech recognition?" in *Proc. ICSLP*, Sydney, Australia, 1998, pp. 85–88.

[22] P. Fung and W. K. Liu, "Fast accent identification and accented speech recognition," in *Proc. ICASSP*, 1999, pp. 221–224.

[23] L. M. Tomokiyo, "Lexical and acoustic modeling of non-native speech in LVCSR," in *Proc. ICSLP*, Beijing, China, 2000, pp. 346–349.

[24] C. Huang, E. Chang, J. Zhou, and K. F. Lee, "Accent modeling based on pronunciation dictionary adaptation for large vocabulary Mandarin speech recognition," in *Proc. ICSLP*, Beijing, China, 2000, pp. 818–821.

[25] K. Livescu, "Analysis and modeling of non-native speech for automatic speech recognition," Master's thesis, Massachusetts Institute of Technology, 1999.

[26] K. Livescu and J. Glass, "Lexical modeling of non-native speech for automatic speech recognition," in *Proc. ICASSP*, Istanbul, Turkey, 2000, pp. 1683–1686.

[27] 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.

[28] 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.

[29] J. Brousseau and S. A. Fox, "Dialect-dependent speech recognizers for Canadian and European French," in *Proc. ICSLP*, Alberta, Canada, 1992, pp. 1003–1006.

[30] I. Kudo, T. Nakama, T. Watanabe, and R. Kameyama, "Data collection of Japanese dialects and its influence into speech recognition," in *Proc. ICSLP*, Philadelphia, PA, 1996, pp. 2021–2024.

[31] C. H. Lee, C. H. Lin, and B. H. Juang, "A study on speaker adaptation of the parameters of continuous density hidden Markov models," *IEEE Trans. Signal Process.*, vol. 39, no. 4, pp. 806–814, 1991.

[32] J. L. Gauvain and C. H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 291–298, 1994.

[33] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Comput. Speech Lang.*, vol. 9, no. 2, pp. 171–185, 1995.

[34] G. Zavaliagkos, R. Schwartz, and J. Makhoul, "Batch, incremental and instantaneous adaptation techniques for speech recognition," in *Proc. ICASSP*, Detroit, MI, 1995, pp. 676–679.

[35] K. Kirchhoff and D. Vergyri, "Cross-dialectal acoustic data sharing for Arabic speech recognition," in *Proc. ICASSP*, Montreal, Quebec, Canada, 2004, pp. 765–768.

[36] ——, "Cross-dialectal data sharing for acoustic modeling in Arabic speech recognition," *Speech Commun.*, vol. 46, no. 1, pp. 37–51, 2005.

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

[38] Q. Zhang, T. Li, J. Pan, and Y. Yan, "Nonnative speech recognition based on state-level bilingual model modification," in *Proc. ICCIT*, Busan, Korea, 2008, pp. 1220–1225.

[39] Q. Zhang, J. Pan, S. Chan, and Y. Yan, "Nonnative speech recognition based on bilingual model modification," in *Proc. FUZZ-IEEE*, Jeju Island, Korea, 2009, pp. 110–114.

[40] 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.

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

[42] V. Fischer, E. Janke, and S. Kunzmann, "Likelihood combination and recognition output voting for the decoding of non-native speech with multilingual HMMs," in *Proc. ICSLP*, Denver, CO, 2002, pp. 489–492.

[43] V. Fischer, E. Janke, S. Kunzmann, and T. Roß, "Multilingual acoustic models for the recognition of non-native speech," in *Proc. ASRU*, Madonna di Campiglio, Italy, 2001, pp. 331–334.

[44] V. Fischer, E. Janke, and S. Kunzmann, "Recent progress in the decoding of non-native speech with multilingual acoustic models," in *Proc. Eurospeech*, Geneva, Switzerland, 2003, pp. 3105–3108.

[45] Y. Liu and P. Fung, "Partial change accent models for accented Mandarin speech recognition," in *Proc. ASRU*, Saint Thomas, USVI, 2003, pp. 111–116.

[46] ——, "Multi-accent Chinese speech recognition," in *Proc. Interspeech – ICSLP*, Pittsburgh, PA, 2006, pp. 133–136.

[47] T. Schultz and A. Waibel, "Fast bootstrapping of LVCSR systems with multilingual phoneme sets," in *Proc. Eurospeech*, Rhodes, Greece, 1997, pp. 371–374.

[48] ——, "Language independent and language adaptive large vocabulary speech recognition," in *Proc. ICSLP*, Sydney, Australia, 1998, pp. 1819–1822.

[49] ——, "Polyphone decision tree specialisation for language adaptation," in *Proc. ICASSP*, Istanbul, Turkey, 2000, pp. 1707–1710.

[50] ——, "Language-independent and language-adaptive acoustic modeling for speech recognition," *Speech Commun.*, vol. 35, pp. 31–51, 2001.

[51] M. Caballero, A. Moreno, and A. Nogueiras, "Data driven multidialectal phone set for Spanish dialects," in *Proc. Interspeech – ICSLP*, Jeju Island, Korea, 2004, pp. 837–840.

[52] ——, "Multidialectal Spanish acoustic modeling for speech recognition," *Speech Commun.*, vol. 51, pp. 217–229, 2009.

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

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

[55] 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.

[56] S. Bowerman, "White South African English: phonology," in *A Handbook of Varieties of English*, E. W. Schneider, K. Burridge, B. Kortmann, R. Mesthrie, and C. Upton, Eds., vol. 1. Berlin, Germany: Mouton de Gruyter, 2004, pp. 931–942.

[57] ——, "White South African English: morphology and syntax," in *A Handbook of Varieties of English*, B. Kortmann, K. Burridge, R. Mesthrie, E. W. Schneider, and C. Upton, Eds., vol. 2. Berlin, Germany: Mouton de Gruyter, 2004, pp. 949–961.

[58] R. Mesthrie, "Black South African English: morphology and syntax," in *A Handbook of Varieties of English*, B. Kortmann, K. Burridge, R. Mesthrie, E. W. Schneider, and C. Upton, Eds., vol. 2. Berlin, Germany: Mouton de Gruyter, 2004, pp. 962–973.

[59] B. van Rooy, "Black South African English: phonology," in *A Handbook of Varieties of English*, E. W. Schneider, K. Burridge, B. Kortmann, R. Mesthrie, and C. Upton, Eds., vol. 1. Berlin, Germany: Mouton de Gruyter, 2004, pp. 943–952.

[60] P. Finn, "Cape Flats English: phonology," in *A Handbook of Varieties of English*, E. W. Schneider, K. Burridge, B. Kortmann, R. Mesthrie, and C. Upton, Eds., vol. 1. Berlin, Germany: Mouton de Gruyter, 2004, pp. 964–984.

[61] K. McCormick, "Cape Flats English: morphology and syntax," in *A Handbook of Varieties of English*, B. Kortmann, K. Burridge, R. Mesthrie, E. W. Schneider, and C. Upton, Eds., vol. 2. Berlin, Germany: Mouton de Gruyter, 2004, pp. 993–1005.

[62] R. Mesthrie, "Indian South African English: phonology," in *A Handbook of Varieties of English*, E. W. Schneider, K. Burridge, B. Kortmann, R. Mesthrie, and C. Upton, Eds., vol. 1. Berlin, Germany: Mouton de Gruyter, 2004, pp. 953–963.

[63] ——, "Indian South African English: morphology and syntax," in *A Handbook of Varieties of English*, B. Kortmann, K. Burridge, R. Mesthrie, E. W. Schneider, and C. Upton, Eds., vol. 2. Berlin, Germany: Mouton de Gruyter, 2004, pp. 974–992.

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

[65] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY: Springer, 2006.

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

[67] B. Mak and E. Barnard, "Phone clustering using the Bhattacharyya distance," in *Proc. ICSLP*, Philadelphia, PA, 1996, pp. 2005–2008.

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

[69] J. A. C. Badenhorst and M. H. Davel, "Data requirements for speaker independent acoustic models," in *Proc. PRASA*, Cape Town, South Africa, 2008, pp. 147–152.

[70] J. R. Hershey and P. A. Olsen, "Variational Bhattacharyya divergence for hidden Markov models," in *Proc. ICASSP*, Las Vegas, NV, 2008, pp. 4557–4560.

[71] J. A. C. Badenhorst, "Data sufficiency analysis for automatic speech recognition," Master's thesis, Potchefstroom Campus, North-West University, 2009.

[72] J. J. Odell, "The use of context in large vocabulary speech recognition," Ph.D. dissertation, University of Cambridge, 1995.

[73] H. A. Engelbrecht, "Automatic phoneme recognition of South African English," Master's thesis, Stellenbosch University, 2004.

[74] 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.

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

[76] 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.

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

[78] H. Ney, U. Essen, and R. Kneser, "On structuring probabilistic dependencies in stochastic language modelling," *Comput. Speech Lang.*, vol. 8, no. 1, pp. 1–38, 1994.

[79] L. ten Bosch, "ASR, dialects, and acoustic/phonological distances," in *Proc. ICSLP*, Beijing, China, 2000, pp. 1009–1012.

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

[81] 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.

[82] M. A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 1, pp. 31–44, 1996.

[83] T. Schultz, I. Rogina, and A. Waibel, "LVCSR-based language identification," in *Proc. ICASSP*, Atlanta, GA, 1996, pp. 781–784.

[84] T. Anastasakos, J. McDonough, and J. Makhoul, "Speaker adaptive training: a maximum likelihood approach to speaker normalization," in *Proc. ICASSP*, Munich , Germany, 1997, pp. 1043–1046.

[85] L. Loots and T. R. Niesler, "Automatic conversion between pronunciations of different English accents," *Speech Commun.*, vol. 53, no. 1, pp. 75–84, 2010.

Appendix A

Derivations

A.1 Covariance Matrix of a Cluster of Two States

Suppose we have a cluster consisting of two states $\mathbf{S} = \{s_1, s_2\}$ and that the observation vectors for training frames $\mathbf{F}_1$ and $\mathbf{F}_2$ are assigned to states s_1 and s_2, respectively. The mean and covariance matrix for the observation probability density function (PDF) of state s_1 can then be calculated:

$$\mu_{s_1} = \frac{1}{N_1} \sum_{f \in \mathbf{F_1}} \mathbf{o}_f \tag{A.1}$$

$$\Sigma_{s_1} = \frac{1}{N_1} \sum_{f \in \mathbf{F}_1} (\mathbf{o}_f - \mu_{s_1})(\mathbf{o}_f - \mu_{s_1})^{\mathrm{T}} \tag{A.2}$$

where

$$N_1 = \sum_{f \in \mathbf{F}} \gamma_{s_1}(\mathbf{o}_f) \tag{A.3}$$

with similar equations for the observation PDF of state s_2. The common covariance matrix of the cluster is then given by

$$\begin{aligned} \Sigma(\mathbf{S}) &= \frac{1}{N} \sum_{f \in \mathbf{F}} (\mathbf{o}_f - \mu(\mathbf{S}))(\mathbf{o}_f - \mu(\mathbf{S}))^{\mathrm{T}} \\ &= \frac{1}{N_1 + N_2} \left[\sum_{f \in \mathbf{F}_1} (\mathbf{o}_f - \mu(\mathbf{S}))(\mathbf{o}_f - \mu(\mathbf{S}))^{\mathrm{T}} + \sum_{f \in \mathbf{F}_2} (\mathbf{o}_f - \mu(\mathbf{S}))(\mathbf{o}_f - \mu(\mathbf{S}))^{\mathrm{T}} \right] \end{aligned} \tag{A.4}$$

Using (A.1), (A.2) and (A.3) the first term in the brackets in (A.4) can be expanded and simplified as follows:

$$
\begin{aligned}
&\sum_{f\in\mathbf{F}_1}(\mathbf{o}_f-\mu(\mathbf{S}))(\mathbf{o}_f-\mu(\mathbf{S}))^{\mathrm{T}}\\
&=\sum_{f\in\mathbf{F}_1}(\mathbf{o}_f-\mu_{s_1}+\mu_{s_1}-\mu(\mathbf{S}))(\mathbf{o}_f-\mu_{s_1}+\mu_{s_1}-\mu(\mathbf{S}))^{\mathrm{T}}\\
&=\sum_{f\in\mathbf{F}_1}((\mathbf{o}_f-\mu_{s_1})+(\mu_{s_1}-\mu(\mathbf{S})))((\mathbf{o}_f-\mu_{s_1})+(\mu_{s_1}-\mu(\mathbf{S})))^{\mathrm{T}}\\
&=\sum_{f\in\mathbf{F}_1}\big[(\mathbf{o}_f-\mu_{s_1})(\mathbf{o}_f-\mu_{s_1})^{\mathrm{T}}+(\mathbf{o}_f-\mu_{s_1})(\mu_{s_1}-\mu(\mathbf{S}))^{\mathrm{T}}\\
&\qquad+(\mu_{s_1}-\mu(\mathbf{S}))(\mathbf{o}_f-\mu_{s_1})^{\mathrm{T}}+(\mu_{s_1}-\mu(\mathbf{S}))(\mu_{s_1}-\mu(\mathbf{S}))^{\mathrm{T}}\big]\\
&=N_1\cdot\Sigma_{s_1}+\sum_{f\in\mathbf{F}_1}\big[\mathbf{o}_f\cdot\mu_{s_1}^{\mathrm{T}}-\mathbf{o}_f\cdot\mu^{\mathrm{T}}(\mathbf{S})-\mu_{s_1}\cdot\mu_{s_1}^{\mathrm{T}}+\mu_{s_1}\cdot\mu^{\mathrm{T}}(\mathbf{S})+\mu_{s_1}\cdot\mathbf{o}_f^{\mathrm{T}}\\
&\qquad-\mu_{s_1}\cdot\mu_{s_1}^{\mathrm{T}}-\mu(\mathbf{S})\cdot\mathbf{o}_f^{\mathrm{T}}+\mu(\mathbf{S})\cdot\mu_{s_1}^{\mathrm{T}}\big]+\sum_{f\in\mathbf{F}_1}(\mu_{s_1}-\mu(\mathbf{S}))(\mu_{s_1}-\mu(\mathbf{S}))^{\mathrm{T}}\\
&=N_1\cdot\Sigma_{s_1}+N_1\cdot\mu_{s_1}\cdot\mu_{s_1}^{\mathrm{T}}-N_1\cdot\mu_{s_1}\cdot\mu^{\mathrm{T}}(\mathbf{S})-N_1\cdot\mu_{s_1}\cdot\mu_{s_1}^{\mathrm{T}}+N_1\cdot\mu_{s_1}\cdot\mu^{\mathrm{T}}(\mathbf{S})\\
&\qquad+N_1\cdot\mu_{s_1}\cdot\mu_{s_1}^{\mathrm{T}}-N_1\cdot\mu_{s_1}\cdot\mu_{s_1}^{\mathrm{T}}-N_1\cdot\mu(\mathbf{S})\cdot\mu_{s_1}^{\mathrm{T}}+N_1\cdot\mu(\mathbf{S})\cdot\mu_{s_1}^{\mathrm{T}}\\
&\qquad+\sum_{f\in\mathbf{F}_1}(\mu_{s_1}-\mu(\mathbf{S}))(\mu_{s_1}-\mu(\mathbf{S}))^{\mathrm{T}}\\
&=N_1\cdot\Sigma_{s_1}+\sum_{f\in\mathbf{F}_1}(\mu_{s_1}-\mu(\mathbf{S}))(\mu_{s_1}-\mu(\mathbf{S}))^{\mathrm{T}}\\
&=N_1\cdot\big[\Sigma_{s_1}+(\mu_{s_1}-\mu(\mathbf{S}))(\mu_{s_1}-\mu(\mathbf{S}))^{\mathrm{T}}\big]
\end{aligned}
\tag{A.5}
$$

A similar expression can be determined for the second term in the brackets in (A.4). By using this result, (A.4) can be written as

$$
\Sigma(\mathbf{S})=\frac{N_1\cdot\big[\Sigma_{s_1}+(\mu_{s_1}-\mu(\mathbf{S}))(\mu_{s_1}-\mu(\mathbf{S}))^{\mathrm{T}}\big]+N_2\cdot\big[\Sigma_{s_2}+(\mu_{s_2}-\mu(\mathbf{S}))(\mu_{s_2}-\mu(\mathbf{S}))^{\mathrm{T}}\big]}{N_1+N_2}
\tag{A.6}
$$

Appendix B

Phone Set

Table B.1: The resulting phone set used in this research after performing phone mappings.

Description	IPA	ASTbet	Example[1]	Occurrences[2]
Stops				
Voiceless bilabial plosive	p	`p`	s<u>p</u>it	12983
Voiced bilabial plosive	b	`b`	<u>b</u>aby	12574
Voiceless alveolar plosive	t	`t`	<u>t</u>otal	67772
Voiced alveolar plosive	d	`d`	<u>d</u>eath	24996
Voiceless velar plosive	k	`k`	<u>k</u>ick	25847
Voiced velar plosive	ɡ	`g`	<u>g</u>un	5628
Glottal stop	ʔ	`gl`	co_operative	6145
Fricatives				
Voiceless labiodental fricative	f	`f`	<u>f</u>our	28512
Voiced labiodental fricative	v	`v`	<u>v</u>at	17686
Voiceless dental fricative	θ	`th`	<u>th</u>ing	12343
Voiced dental fricative	ð	`dh`	<u>th</u>is	8150
Voiceless alveolar fricative	s	`s`	<u>s</u>ome	54947
Voiced alveolar fricative	z	`z`	<u>z</u>ero	11628
Voiceless post-alveolar fricative	ʃ	`sh`	<u>sh</u>ine	4757
Voiced post-alveolar fricative	ʒ	`zh`	<u>g</u>enre	318
Voiceless velar fricative	x	`x`	lo<u>ch</u>	498
Voiceless glottal fricative	h	`h`	<u>h</u>and	6081
Voiced glottal fricative	ɦ	`hht`	*<u>h</u>and* (Afr.)	1493
Affricates				
Post-alveolar affricate	t͡ʃ	`t_lnksh`	<u>ch</u>ocolate	3772
Voiced post-alveolar affricate	d͡ʒ	`d_lnkzh`	<u>j</u>ug	4035
Trills and flaps				
Alveolar trill	r	`r`	*<u>r</u>oer* (Afr.)	12613
Alveolar flap	ɾ	`fh`	fo<u>r</u>ty (lazy)[3]	881

[1] Mainly English examples except for some words from Afrikaans (Afr.), Xhosa and Zulu.
[2] The number of phone tokens in the combined training set from all five SAE accents as indicated in Table 3.2.
[3] The lazy *vat <u>d</u>it* in Afrikaans is another example.

Table (*continued*)

Description	IPA	ASTbet	Example	Occurrences
Approximants				
Alveolar approximant	ɹ	rt	red	18477
Alveolar lateral approximant	l	l	legs	26996
Palatal approximant	j	j	yes	11097
Voiced labio-velar approximant	w	w	west	19315
Nasals				
Bilabial nasal	m	m	man	19435
Alveolar nasal	n	n	not	76892
Velar nasal	ŋ	nj	thing	7067
Vowels				
High front vowel with duration	iː	i_long	keep	29914
Lax front vowel	ɪ	ic	him	91028
Rounded high front vowel	y	y	*u* (Afrikaans)	146
High back vowel	u	u	*vulani* (Xhosa)	10462
High back vowel with duration	uː	u_long	blue	12823
Lax back vowel	ʊ	hs	push	21039
Mid-high front vowel	e	e	*eweredig* (Afr.)	10751
Mid-high front vowel with duration	eː	e_long	*been* (Afr.)	171
Rounded mid-high back vowel with duration	oː	o_long	*groot* (Afr.)	109
Mid-low front vowel	ɛ	ep	nest	37687
Mid-low front vowel with duration	ɛː	ep_long	fairy	1592
Central vowel with duration	ɜː	epr_long	turn	6256
Rounded mid-low back vowel with duration	ɔː	ct_long	bore	19393
Low back vowel	ɒ	ab	hot	10260
Lax mid-low vowel	ʌ	vt	hut	27941
Low central vowel	a	a	*ukusala* (Zulu)	19098
Low central vowel with duration	aː	a_long	*saak* (Afr.)	2189
Low back vowel with duration	ɑː	as_long	harp	5978
Central vowel (schwa)	ə	sw	the	86174
Mid-low front vowel	æ	ae	average	14270
Mid-low front vowel with duration	æː	ae_long	dad	842

Table B.2: Phone mappings applied to the original English AST databases in order to obtain the smaller set of 50 phones common to all five SAE accents and listed in Table B.1.

Substitution	Occurrences
Stops	
[pʰ] → [p]	20
[p’] → [p]	6
[b̥] → [b]	6
[ɓ] → [b]	38
[tʰ] → [t]	32
[t’] → [t]	12
[d̥] → [d]	1
[j̥] → [j]	1
[ʝ] → [j]	1
[kʰ] → [k]	14
[k’] → [k]	18
Fricatives	
[s’] → [s]	67
[ɬ] → [ʃl]	21
[ɬ’] → [l]	6
[ɮ] → [dl]	5
Clicks	
[ǀ] → [k]	5
[ǀ̃] → [k]	1
[ǀ̰] → [k]	4
[ǀ̤] → [k]	1
[ǁ] → [k]	36
[ǁ̃] → [k]	1
[ǁʰ] → [k]	6
[ǃ] → [k]	2
[ǃ̃] → [k]	1
[ǃ̰] → [k]	2

Substitution	Occurrences
Affricates	
[t͡ʃʰ] → [t͡ʃ]	3
[d͡ʒ] → [z]	21
[k͡xˀ] → [k]	1
Trills and flaps	
[ʀ] → [r]	687
Approximants	
[ɹ̥] → [ɹ]	2
[l̥] → [l]	1
[w̥] → [w]	2
Nasals	
[n̥] → [n]	6
[ɲ] → [ŋ]	8
Vowels	
[i] → [ɪ]	52788[a]
[ɪ̥] → [ɪ]	77
[yː] → [uː]	31
[ʊ̥] → [ʊ]	41
[ø] → [eː]	19
[øː] → [eː]	6
[ɛ̥] → [ɛ]	5
[œ] → [ɜː]	552[b]
[œː] → [ɜː]	484[b]
[o] → [ɔː]	99
[ɔ] → [ɔː]	9543[c]
[ɔ̥] → [ɔː]	5
[ɒ̥] → [ɒ]	3
[ḁ] → [a]	11
[ə̥] → [ə]	33
[æ̥] → [æ]	2

[a] Some transcribers were mistakenly told to only use [i] or [ɪ] for particular databases while other transcribers did not distinguish between these two phones.

[b] Early on in the AST project, a distinction was made between [œ], [œː] and [ɜː], but later in the project only [ɜː] was used.

[c] A distinction between [ɔ] and [ɔː] was only made later on in the AST project, initially only [ɔː] was used.

Appendix C

Language, Pronunciation and Acoustic Modelling Alternatives

In Chapter 6 we employed accent-specific phone backoff bigram language models (LMs) in phone recognition experiments and accent-independent word backoff bigram LMs in word recognition experiments. For word recognition experiments we used a pooled accent-independent pronunciation dictionary (PD) obtained by pooling the pronunciations from the accent-specific dictionaries (described in Section 5.2) for the accents considered in the particular experiments. These design decisions were made based on preliminary experiments which are described in this appendix. We show that the relative performance of the three acoustic modelling approaches (Section 4.2) when applied to the five SAE accents remains unchanged when using alternative LMs and PDs. We also briefly describe some alternative acoustic modelling strategies which we considered but which did not lead to any considerable differences. All experiments in this appendix use an oracle recognition setup, i.e. the accent of each test utterance is assumed to be known.

C.1 Accent-Specific vs. Accent-Independent Phone Language Models

Accent-independent LMs can be employed as an alternative to the accent-specific phone LMs that we used in our main experiments. Following the modelling approach and tools described in Section 5.1, an accent-independent phone backoff bigram LM was trained on the combined set of phone-level training transcriptions of all five accents in the AST databases (approximately 911k phone tokens, Table 3.2). The perplexities of the accent-independent and accent-specific LMs are compared in Table C.1. The table indicates that, for all five accents, the perplexities for the accent-specific LMs are lower than those achieved by the accent-independent LM.

Table C.1: Phone bigram language model (LM) perplexities (perp.) measured on the evaluation sets of all five SAE accents. The accent-specific LMs match the accent of the set under evaluation.

Accent	Phone bigram types	Accent-specific LM perp.	Accent-independent LM perp.
AE	1891	14.40	15.33
BE	1761	15.44	18.39
CE	1834	14.12	14.60
EE	1542	12.64	13.85
IE	1760	14.24	15.16

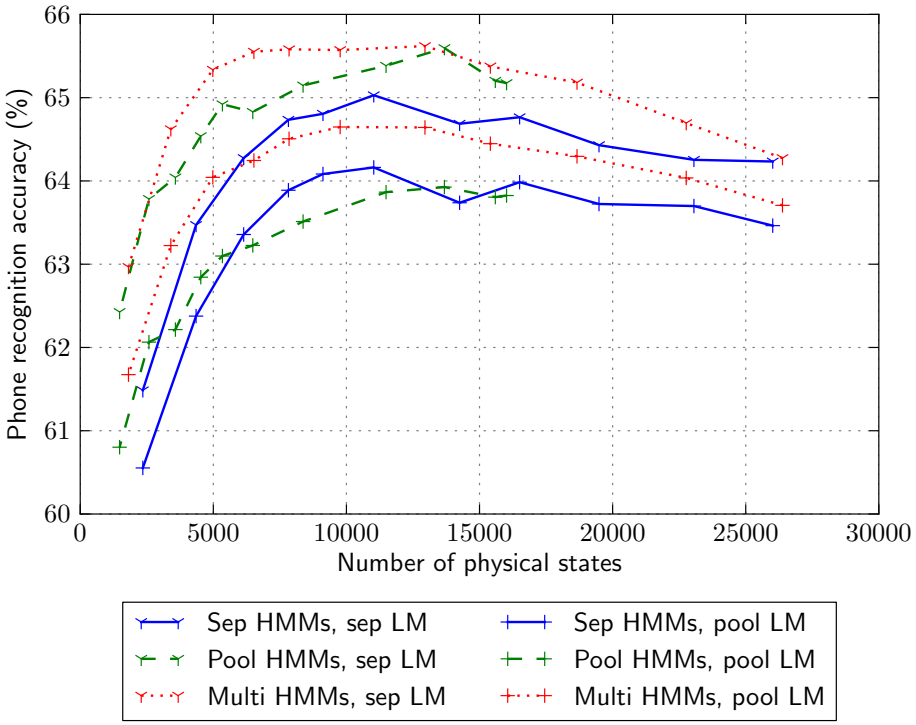


Figure C.1: Average evaluation set phone accuracies of accent-specific (sep HMMs), accent-independent (pool HMMs) and multi-accent (multi HMMs) systems employing accent-specific (sep LM) and accent-independent (pool LM) language models.

Figure C.1 shows the average phone recognition accuracies measured on the evaluation sets of all five accents for accent-specific, accent-independent and multi-accent acoustic models with the accent-specific and accent-independent phone LMs. By comparing systems employing the same acoustic modelling approach but using different LMs, it is evident that the accent-specific LMs consistently outperform the accent-independent LM for all three acoustic modelling approaches.

Since 'accent' refers to pronunciation differences (Section 2.1.1), one might expect that phone sequences differ across accents. We believe that this is the reason for the better performance of the accent-specific LMs, despite the approximately five times larger training set available for the accent-independent LM.

Figure C.1 also indicates that the multi-accent modelling approach yields similar or improved performance compared to the other two acoustic modelling approaches for systems employing the same language modelling strategy. It is interesting, however, that for the systems employing accent-independent phone LMs, accent-specific acoustic models appear to yield superior performance over accent-independent models for several system sizes. This does not happen for the systems employing accent-specific phone LMs.

C.2 Accent-Specific vs. Accent-Independent Word Language Models

As for phone recognition, accent-specific and accent-independent LMs were also compared for the word recognition case. Following the modelling approach and tools described in Section 5.1, accent-specific word backoff bigram LMs were trained for each of the five SAE accents individually on the corresponding training set transcriptions. For each of the five accent-specific LMs, the vocabulary was taken from the training transcriptions of that particular accent. This

Table C.2: Word bigram language model (LM) perplexities and OOV rates measured on the evaluation sets of all five SAE accents. The accent-specific LMs match the accent of the set under evaluation.

Accent	Word bigram types	Accent-specific LMs		Accent-independent LMs	
		Perplexity	OOV rate (%)	Perplexity	OOV rate (%)
AE	11 580	25.81	4.87	22.21	4.87
BE	9639	30.30	6.90	25.77	6.90
CE	10 641	30.87	5.24	25.31	5.24
EE	10 451	28.97	4.74	23.51	4.74
IE	11 677	26.22	5.09	23.36	5.09

allowed the accent-specific PDs described in Section 5.2 to be employed during recognition. Accent-independent word LMs were trained in a similar fashion on the training transcriptions of all five accents in the AST databases (approximately 240k words as indicated in Table 3.2). In order to compare the LMs and to be able to employ the same PDs used in the accent-specific case, five different accent-independent LMs were trained according to the distinct vocabularies of the respective accents. LM perplexities and OOV rates are given in Table C.2. Since the same vocabularies were used, OOV rates are identical for the accent-specific and accent-independent LMs, making a direct comparison of recognition performance possible. In both cases the perplexity of a particular LM was calculated on the evaluation set of the matching accent. The accent-independent LMs described here differ from the LM used in Section 6.6 since the vocabulary of the latter was taken from the combined set of training transcriptions of all five accents.

Table C.2 indicates that, for all five accents, the accent-independent LMs yield lower perplexities. This agrees with the results in Figure C.2 which shows the average word recognition accuracy measured on the evaluation sets where oracle five-accent systems employing accent-specific, accent-independent and multi-accent acoustic models were used with the accent-specific and accent-independent word LMs. Very similar relative trends are observed for the three acoustic modelling approaches for both language modelling strategies: the multi-accent models consistently outperform both the accent-specific and accent-independent models. However, it is evident that systems employing the accent-independent LMs consistently outperform systems employing the accent-specific word LMs. This is in contrast to the phone recognition experiments, in which accent-specific phone LMs were found to outperform accent-independent phone LMs. We ascribe these differences to the observation that, unlike the word sequences, the phone sequences are accent-specific. In addition, the phone LMs are trained on a larger training set.

C.3 Accent-Specific vs. Accent-Independent Pronunciation Dictionaries

For the experiments presented in Section 6.6 an accent-independent PD was obtained by combining the five accent-specific PDs described in Section 5.2. An alternative would be to employ the original accent-specific PDs. We compare these two alternatives in this section. In the previous section we showed that accent-independent language modelling outperforms accent-specific modelling for word recognition and we therefore employ the former in the following experiments.

Since the vocabulary of the accent-specific PDs is consistent with the words in the individual accented training sets, while the vocabulary of the accent-independent PD is consistent with the combination of the words in all five training sets, different LMs are required for the evaluation of the two PDs. In Section C.2 we described accent-independent LMs which were trained on

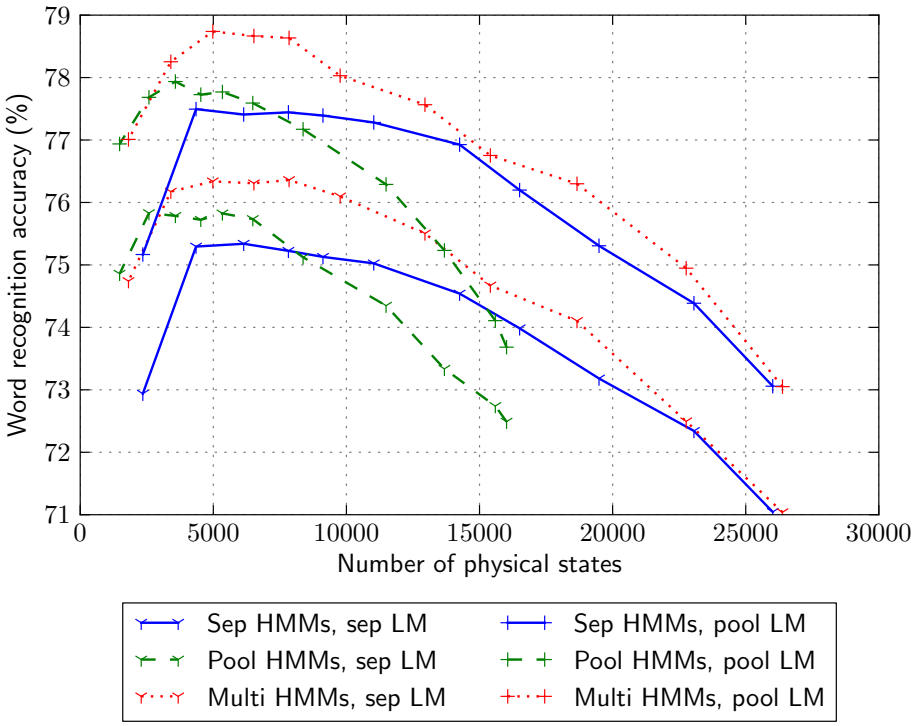


Figure C.2: Average evaluation set word accuracies of accent-specific (sep HMMs), accent-independent (pool HMMs) and multi-accent (multi HMMs) systems employing accent-specific (sep LM) and accent-independent (pool LM) language models.

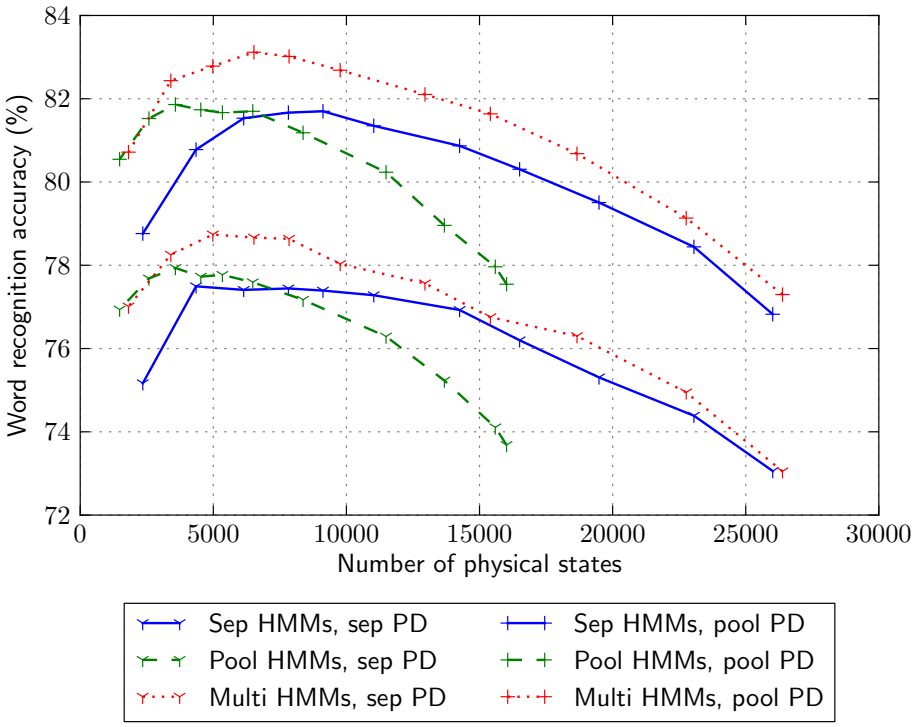


Figure C.3: Average evaluation set word accuracies of accent-specific (sep HMMs), accent-independent (pool HMMs) and multi-accent (multi HMMs) systems employing accent-specific (sep PD) and accent-independent (pool PD) pronunciation dictionaries.

the combined training transcriptions of all five accents in the AST databases, but with the vocabularies taken from the individual accented training sets. The vocabularies of these accent-independent LMs are consistent with the original accent-specific PDs and were therefore used here for the evaluation of the accent-specific PDs. The perplexities of these LMs are given as part of Table C.2. The vocabulary of the accent-independent LM used for the experiments described in Section 6.6 matches the pooled accent-independent PD and this combination was therefore used for evaluation of the accent-independent PD. Perplexities of this LM are given as part of Table 6.19. Since the vocabularies of the two accent-independent LMs are different, the perplexities in Tables 6.19 and C.2 are not directly comparable. The average number of pronunciations per word for both PDs are shown in Table C.3.

Figure C.3 shows the average word recognition accuracy measured on the evaluation sets where oracle five-accent systems employing accent-specific, accent-independent and multi-accent acoustic models were used with the accent-specific and accent-independent PDs and the corresponding LMs. It is evident that systems employing the pooled accent-independent PD significantly outperform systems using the accent-specific PDs. By comparing Table 6.19 with Table C.2 it is apparent that much of this gain can be attributed to the lower OOV rates associated with the accent-independent PD. However, Table C.3 indicates that the average number of pronunciations for the accent-independent PD is very high (2.10 pronunciations per word) compared to those of the accent-specific PDs (ranging between 1.33 and 1.09).

C.4 Other Pronunciation Dictionaries

Since many of the words in the separate accent-specific PDs described in Section 5.2 overlap, the average number of pronunciations per word is significantly increased when pooling the accent-specific PDs. A larger number of pronunciation variants could increase confusability and result in deteriorated speech recognition performance. However, we illustrated in the previous section that the pooled accent-independent PD improves performance compared to the accent-specific PDs since more evaluation set words are covered by the accent-independent PD. In this section we consider two alternatives for obtaining accent-specific PDs with vocabularies that match those of the pooled accent-independent PD. In other words, we obtain accent-specific PDs with pronunciations for all the words in the combined training set transcriptions of all five English AST databases.

The first approach was to augment the pronunciations in each accent-specific PD with the missing pronunciations taken from one of the other accent-specific PDs, without modification. We created such augmented accent-specific PDs for all five accents. When a pronunciation for a word was missing for a particular accent, the most similar accent's PD was consulted and the corresponding pronunciation (or pronunciations) added. If a pronunciation was not found, the next-closest accent's PD was consulted, and so on. We repeated this process for all five

Table C.3: The average number of pronunciations per word for the different pronunciation dictionaries.

Accent	Original accent-specific PDs	Augmented accent-specific PDs	G2P generated accent-specific PDs	Pooled accent-independent PD
AE	1.26	1.19	1.13	2.10
BE	1.33	1.22	1.13	2.10
CE	1.28	1.18	1.12	2.10
EE	1.09	1.12	1.04	2.10
IE	1.16	1.15	1.08	2.10

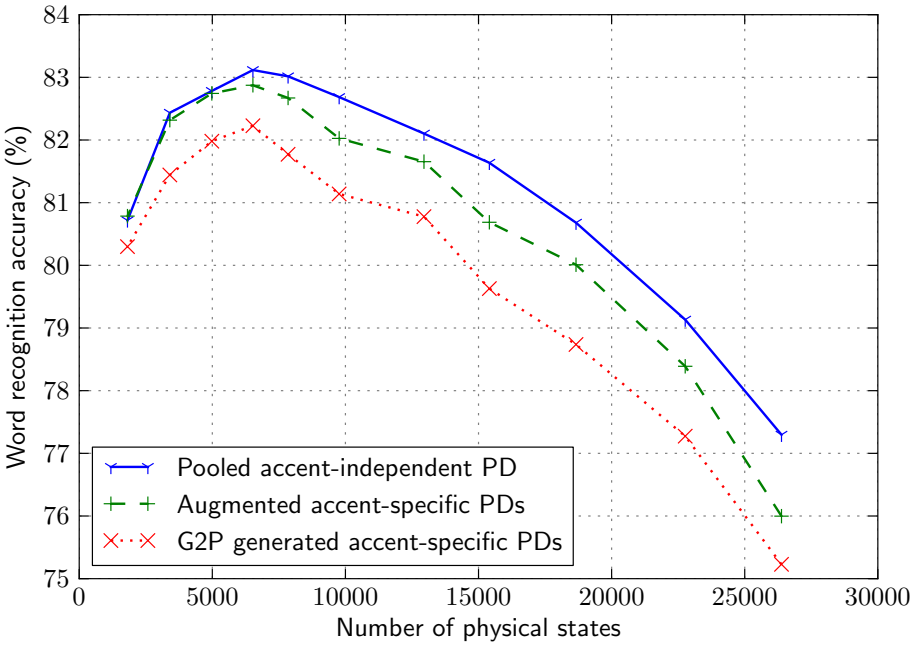


Figure C.4: Average evaluation set word accuracies of multi-accent systems employing pooled accent-independent, augmented accent-specific and G2P generated accent-specific pronunciation dictionaries.

accents. The result was five accent-specific PDs which each contain pronunciations for all the words in the combined training set transcriptions. The "closeness" of accents was based on the PD similarity analysis presented in Section 5.2 and summarised in Table 5.3.

The second approach was to use the tools described in [85] to perform grapheme-to-phoneme (G2P) conversion for all the words missing from each original accent-specific PD. The G2P rules were trained individually for each accent on the available pronunciations in the accent-specific PDs. Each of the accent-specific PDs was then supplemented with G2P generated pronunciations for all the missing words.

The vocabularies of the PDs obtained for both the above approaches, as well as that of the pooled accent-independent PD, are identical. For all three cases the average number of pronunciations per word is indicated in Table C.3. The PDs were evaluated in recognition experiments using the accent-independent LM described in Section 6.6.1, which was trained on the combined training transcriptions of all five English AST databases and was also used in Section C.3.

Figure C.4 shows the average evaluation set word recognition accuracy achieved by oracle five-accent multi-accent systems, respectively employing the pooled accent-independent, the augmented accent-specific and the G2P generated accent-specific PDs. Despite the much higher average number of pronunciations per word for the pooled accent-independent PD, consistently superior performance is achieved by the multi-accent systems employing this PD compared with the two alternatives. Although the results in Figure C.4 are for systems employing multi-accent acoustic models, similar relative trends were observed for systems employing accent-specific and accent-independent acoustic models. We conclude that, of the different approaches considered, the pooled accent-independent PD is the best dictionary to employ.

C.5 Other Experiments

In this section we briefly describe some alternative approaches which we considered but which did not lead to any significant performance improvements. The accent-independent LM and the

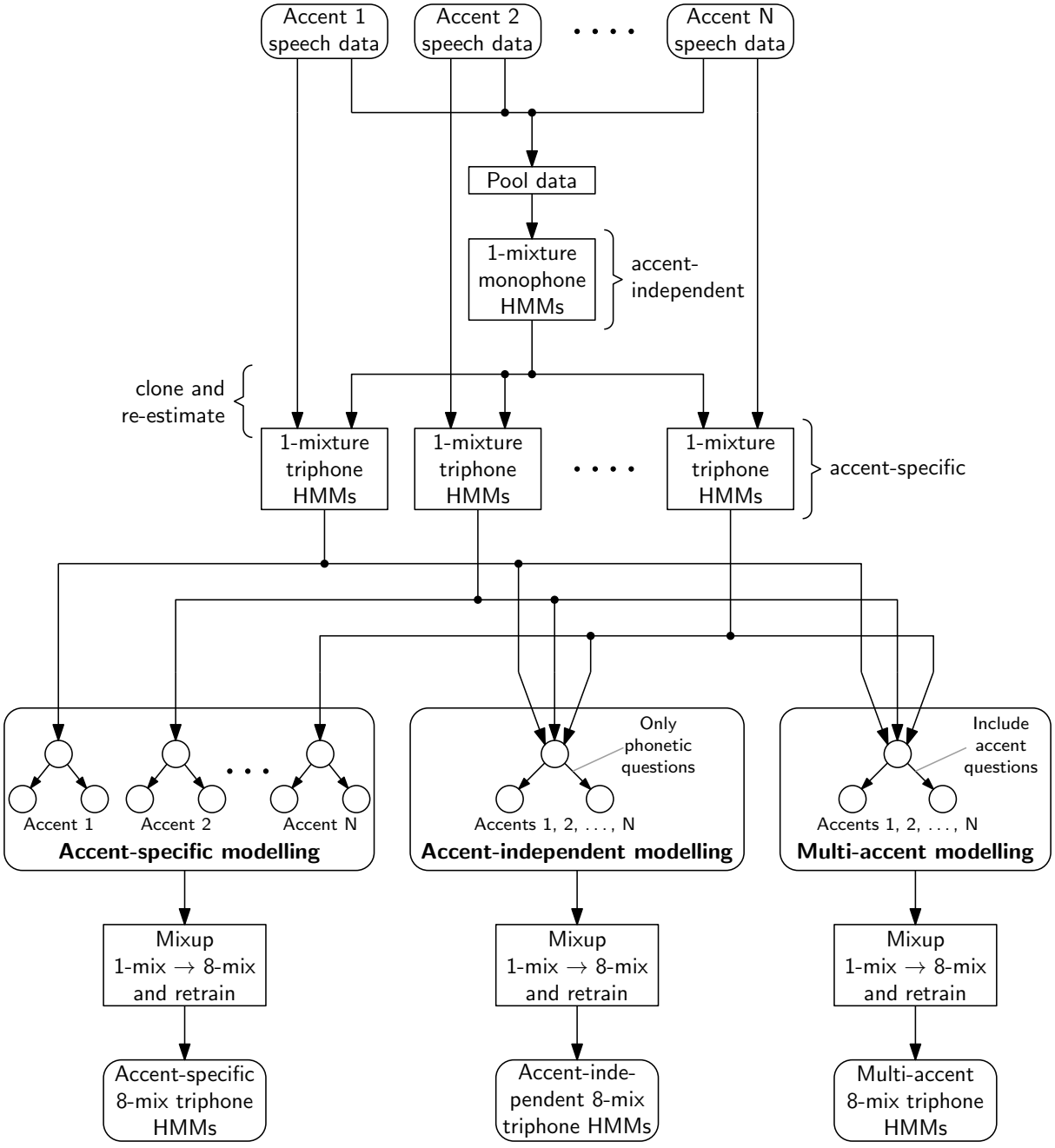


Figure C.5: The acoustic model training process for the accent-specific, accent-independent and multi-accent acoustic modelling approaches applied to an arbitrary number of accents when first training accent-independent monophone HMMs and then cloning and re-estimating these to obtain accent-specific triphone HMMs which are subsequently clustered and then mixed up.

pooled accent-independent PD used for the experiments described in the previous section were employed for the experiments described here.

The accent-specific, accent-independent and multi-accent acoustic models evaluated throughout this research were trained according to the procedure described in Section 4.2.4 and summarised in Figure 4.6. In this procedure a set of accent-specific single-mixture monophone HMMs are trained and subsequently cloned and re-estimated to obtain triphone HMMs. These are then clustered using the three decision-tree strategies corresponding to the three acoustic modelling approaches. After clustering, the number of Gaussian mixtures is increased to yield the final models. An alternative to this would be to first train accent-independent monophone HMMs and then clone these and re-estimate with the accent-specific data in order to obtain initial accent-specific triphone models. Clustering and mixture incrementing can then be performed as in the original approach. This alternative training procedure is illustrated in Figure C.5.

We evaluated and compared this alternative to the original procedure followed. Figure C.6 shows the average evaluation set word recognition accuracy achieved by oracle five-accent multi-accent systems respectively trained using the two approaches illustrated in Figures 4.6 and C.5. The results indicate that there is little to distinguish between the two approaches. Only multi-accent system performance is shown, but the results were similar for both the accent-specific and accent-independent acoustic modelling approaches.

Another design decision described in Section 4.2 was to tie transition probabilities across accents for the accent-independent acoustic modelling approach (see Figure 4.3) while accent-specific transition probabilities were used for the accent-specific (Figure 4.2) and multi-accent (Figure 4.5) acoustic modelling approaches. We also evaluated accent-specific acoustic models for which the transition probabilities were tied across accents, as well as accent-independent models for which the transition probabilities were modelled separately for each accent, and multi-accent acoustic models for which the transition probabilities were tied across accents. In each case only very small changes in recognition performance were observed. This corresponds to the general

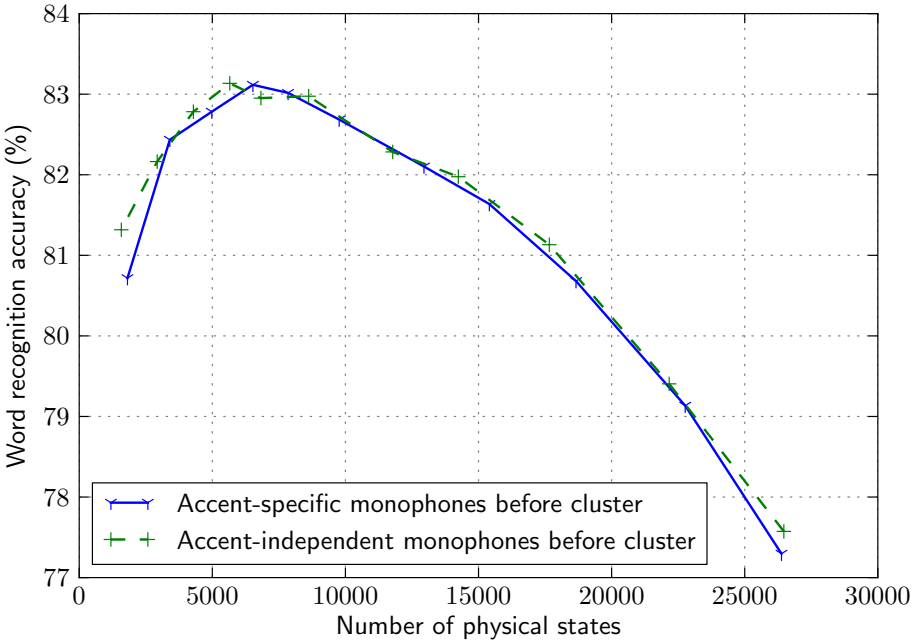


Figure C.6: Average evaluation set word accuracies using multi-accent acoustic models which were obtained by either first training accent-specific monophone HMMs, cloning, clustering and re-estimating as in Figure 4.6; or was obtained by first training accent-independent monophone HMMs, cloning, clustering and re-estimating as in Figure C.5. In both cases the result is eight-mixture triphone HMMs.

perception that the performance of HMM-based ASR systems tends to be robust towards the values of the HMM transition probabilities (e.g. [75, pp. 18, 39]).

C.6 Summary and Conclusions

This appendix described and evaluated several language and pronunciation modelling strategies. We showed that accent-specific language models (LMs) are superior for phone recognition (Section C.1), while an accent-independent LM is superior for word recognition (Section C.2). We ascribe these differences to the larger size of the phone LM training set and the observation that, unlike the word sequences, the phone sequences are accent-specific. An accent-independent pronunciation dictionary (PD) obtained by pooling accent-specific pronunciations was shown to yield better results than the original accent-specific PDs (Section C.3). The former leads to substantially lower OOV rates, which we believe are the reason for the performance differences. Alternative pronunciation (Section C.4) and acoustic (Section C.5) modelling approaches were also described. The most important conclusion from this appendix is that, although absolute performance might differ between configurations, the relative performance of the three acoustic modelling approaches remains unaffected by the language or pronunciation modelling strategy.