

DATA SUFFICIENCY ANALYSIS FOR AUTOMATIC SPEECH RECOGNITION

JACOB ANDREAS CORNELIUS BADENHORST

DATA SUFFICIENCY ANALYSIS FOR AUTOMATIC SPEECH RECOGNITION

By

J.A.C. Badenhorst

Submitted in fulfilment of the requirements for the degree

Master in Engineering

at the

Potchefstroom Campus, North-West University

Supervisor: Prof. E. Barnard

Co-supervisor: Dr. M.H. Davel

May 2009

DATA SUFFICIENCY ANALYSIS FOR AUTOMATIC SPEECH RECOGNITION

The languages spoken in developing countries are diverse and most are currently under-resourced from an automatic speech recognition (ASR) perspective. In South Africa alone, 10 of the 11 official languages belong to this category. Given the potential for future applications of speech-based information systems such as spoken dialog system (SDSs) in these countries, the design of minimal ASR audio corpora is an important research area. Specifically, current ASR systems utilise acoustic models to represent acoustic variability, and effective ASR corpus design aims to optimise the amount of relevant variation within training data while minimising the size of the corpus. Therefore an investigation of the effect that different amounts and types of training data have on these models is needed.

With this dissertation specific consideration is given to the data sufficiency principals that apply to the training of acoustic models. The investigation of this task lead to the following main achievements: 1) We define a new stability measurement protocol that provides the capability to view the variability of ASR training data. 2) This protocol allows for the investigation of the effect that various acoustic model complexities and ASR normalisation techniques have on ASR training data requirements. Specific trends with regard to the data requirements for different phone categories and how these are affected by various modelling strategies are observed. 3) Based on this analysis acoustic distances between phones are estimated across language borders, paving the way for further research in cross-language data sharing.

Finally the knowledge obtained from these experiments is applied to perform a data sufficiency analysis of a new speech recognition corpus of South African languages: The Lwazi ASR corpus. The findings correlate well with initial phone recognition results and yield insight into the sufficient number of speakers required for the development of minimal telephone ASR corpora.

Keywords: speech recognition, acoustic variability, corpus design, resource-scarce languages, acoustic models, model distances, telephone ASR corpora.

DATATOEREIKENDHEIDSANALISE VIR OUTOMATIESE SPRAAKHERKENNING

'n Wye verskeidenheid tale wat slegs oor beperkte hulpbronne vir outomatiese spraakherkenning (OSH) beskik, word dikwels in ontwikkelende lande aangetref. In Suid-Afrika behoort 10 van die 11 amptelike tale steeds tot hierdie kategorie. Weens die moontlike impak wat toekomstige gebruike van spraakgebaseerde inligtingstelsels (soos spraakdialoogstelsels) in hierdie lande kan hê, is die ontwerp van OSH-korpusse 'n belangrike navorsingsveld. Bestaande OSH-stelsels maak van akoestiese modelle gebruik om akoestiese veranderlikheid te beskryf. Effektiewe OSH-korpusontwerp poog om die hoeveelheid relevante leerdata te optimeer en terselfdertyd die grootte van die korpus te minimeer. Om hierdie balans te handhaaf word 'n ondersoek na die effek wat verskillende hoeveelhede en tipes data op hierdie modelle uitoefen, benodig.

In hierdie verhandeling word daar spesifiek aandag aan die beginsels van toereikende data wat geld vir die suksesvolle opleiding van akoestiese modelle gegee. Die volgende hoofuitkomst is deur hierdie ondersoek bereik: 1) Ons het 'n nuwe stabiliteitsmetingsprotokol ontwerp wat die moontlikheid bied om die veranderlikheid van OSH-leerdata te beskryf; 2) Hierdie protokol skep die moontlikheid om die effek van verskeie akoestiese modelkompleksiteite en normaliseringstegnieke op leerdata noukeurig te ondersoek. Spesifieke patrone met betrekking tot toereikende data is opgemerk vir verskillende foonkategorieë, asook wat die uitwerking van die gebruik van die verskeie modelleringstegnieke hierop is; en 3) Bogenoemde analise laat ook toe dat akoestiese afstande tussen die fone van verskillende tale effektief beraam kan word en maak deure oop vir toekomstige navorsingsgeleenthede in kruistaaldeling van data.

Die kennis wat uit bogenoemde eksperimente verkry is, word nuttig aangewend om 'n datatoereikendheidsanalise uit te voer op 'n nuwe OSH-korpus vir Suid-Afrikaanse tale, te wete die Lwazi OSH-korpus. Bevindinge van hierdie analise korreleer goed met die oorspronklike foonherkenningsresultate en verskaf nuwe insig in die hoeveelheid sprekers wat benodig word vir die ontwikkeling van minimale OSH-korpusse van telefoondata.

Sleuteltermes: spraakherkenning, akoestiese veranderlikheid, korpusontwerp, hulpbronne van tale, akoestiese modelle, modelafstande, OSH-korpusse van telefoondata.

TABLE OF CONTENTS

CHAPTER ONE - INTRODUCTION	1
1.1 Context	1
1.2 Problem statement	2
1.3 Literature review	3
1.3.1 Acoustic model analysis	3
1.3.1.1 Acoustic models in ASR	3
1.3.1.2 Model analysis techniques	4
1.3.2 Data requirements for ASR	5
1.3.2.1 Data selection strategies	5
1.3.2.2 The effect of confusability	6
1.3.3 Cross-language sharing of data	6
1.3.3.1 Language-independent modelling	7
1.3.3.2 Data pooling	7
1.3.3.3 Language-adaptive modelling	8
1.3.3.4 Model adaptation	8
1.3.4 Motivation for efficient ASR corpus design techniques	9
1.4 Overview of dissertation	10
CHAPTER TWO - MODEL DISTANCE ESTIMATION	12
2.1 Introduction	12
2.2 Background	12
2.2.1 Distance metrics and similarity measures	12
2.2.2 The Bhattacharyya Bound	13
2.3 Distance estimation for Single Gaussian Models	15
2.4 Distance estimation for Gaussian Mixture Models	16
2.4.1 Bhattacharyya estimator	16
2.4.2 Verifying the accuracy of estimated Bhattacharyya bound	18
2.4.3 Number of samples	18
2.5 Conclusion	21

CHAPTER THREE - MODEL STABILITY ESTIMATION	22
3.1 Introduction	22
3.2 Factors that influence data variance	22
3.3 Modelling data variance	24
3.3.1 Parameterisation	24
3.3.2 Normalisation	25
3.3.3 Model types	25
3.3.3.1 Current modelling structure	25
3.3.3.2 Model complexity	25
3.4 General approach	26
3.4.1 Model saturation	26
3.4.2 A three-dimensional measure of acoustic stability	27
3.4.3 The effect of subset selection	29
3.4.3.1 Experimental setup	29
3.4.3.2 Experiment 1: Mean of stability measurement	30
3.4.3.3 Experiment 2: Cumulative standard deviation of the bound	32
3.4.3.4 Experiment 3: Cumulative standard deviation of the mean of the bound	33
3.5 Conclusion	35
CHAPTER FOUR - DATA SUFFICIENCY ANALYSIS TECHNIQUES	36
4.1 Introduction	36
4.2 Data and experimental setup	37
4.3 Initial analysis	37
4.4 Single Gaussian model	38
4.4.1 Number of phone observations required per speaker	39
4.4.2 Number of speakers required per phone	40
4.5 Gaussian mixture model	41
4.5.1 Number of samples	42
4.5.2 Effect of additional model complexity	42
4.6 Effect of context-dependence	45
4.7 Bootstrapping	46
4.8 Cepstral Mean Normalisation	47
4.9 Conclusion	48
CHAPTER FIVE - DATA SUFFICIENCY ANALYSIS OF THE LWAZI ASR CORPUS	52
5.1 Introduction	52
5.2 Data and experimental design	52
5.2.1 Lwazi ASR corpus	52

5.2.2	Experimental setup	53
5.3	Analysis of phone variability	54
5.4	Distances between languages	55
5.5	ASR results	57
5.6	Impact of data reduction	58
5.7	Conclusion	59
CHAPTER SIX - CONCLUSION		61
6.1	Introduction	61
6.2	Summary of contribution	62
6.2.1	Literature	62
6.2.2	Main contributions	62
6.3	Further application and future work	64
6.4	Conclusion	65
APPENDIX A - ANNEXTURE		66
A.1	Effect of context-dependence	66
A.2	Phone stability	67
A.3	Distances between languages	71
REFERENCES		77

LIST OF FIGURES

2.1	<i>Decision rule: Minimum misclassification at the optimal decision boundary x_0 (Reproduced from Bishop, 2006 [1])</i>	13
2.2	<i>Difference between estimated distance measure and analytically calculated values for three different single Gaussian model comparisons and different numbers of samples . .</i>	19
2.3	<i>Estimation accuracy of distance measure on Gaussian mixture models containing 4 mixture components each ($e = 35$)</i>	20
3.1	<i>Mean of stability measurements for the phones /ah/,/ih/,/s/ and /z/</i>	29
3.2	<i>Mean of stability measurements for the phones /l/,/d/,/t/ and /p/</i>	30
3.3	<i>Cumulative standard deviation of the bound for the phones /ah/,/ih/,/s/ and /z/</i>	31
3.4	<i>Cumulative standard deviation of the bound for the phones /l/,/d/,/t/ and /p/</i>	32
3.5	<i>Cumulative standard deviation of stability measurements for the phones /ah/,/ih/,/s/ and /z/</i>	33
3.6	<i>Cumulative standard deviation of stability measurements for the phones /l/,/d/,/t/ and /p/</i>	34
4.1	<i>Speaker-and-utterance three-dimensional plot for the phone /ah/</i>	38
4.2	<i>Effect of number of phone utterances per speaker on mean of Bhattacharyya bound for different phone groups using data from 20 speakers</i>	39
4.3	<i>Effect of number of phone utterances per speaker on mean of Bhattacharyya bound for different phone groups using data from 50 speakers</i>	40
4.4	<i>Effect of number of speakers on mean of Bhattacharyya bound for different phone groups using 100 utterances per speaker</i>	41
4.5	<i>Effect of number of phone utterances per speaker on mean of Bhattacharyya bound for different phone groups using data from 20 speakers and 6 mixtures</i>	42
4.6	<i>Effect of number of speakers on mean of Bhattacharyya bound for different phone groups using 100 utterances per speaker and 6 mixtures</i>	43
4.7	<i>Comparing the effect of model complexity on the relative distance to asymptote for the phones /ah/,/ih/,/z/ and /s/</i>	44
4.8	<i>Comparing the effect of model complexity on the relative distance to asymptote for the phones /n/,/l/,/d/ and /t/</i>	45
4.9	<i>Effect of number of phone utterances per speaker on mean of Bhattacharyya bound for different contexts of the phone /ah/ using data from 20 speakers</i>	46
4.10	<i>Effect of number of phone utterances per speaker on mean of Bhattacharyya bound for different contexts of the phone /n/ using data from 20 speakers</i>	47

4.11	<i>Effect of number of phone utterances per speaker on mean of Bhattacharyya bound for different contexts of the phone /s/ using data from 20 speakers</i>	48
4.12	<i>Effect of number of phone utterances per speaker on mean of Bhattacharyya bound for different contexts of the phone /p/ using data from 20 speakers</i>	49
4.13	<i>Effect of number of phone utterances per speaker on mean of Bhattacharyya bound for different phone groups using data from 20 speakers using bootstrapping approach</i>	50
4.14	<i>Effect of number of speakers on mean of Bhattacharyya bound for different phone groups using 100 utterances per speaker using bootstrapping approach</i>	50
4.15	<i>Effect of number of phone utterances per speaker on mean of Bhattacharyya bound for different phone groups using data from 20 speakers applying bootstrapping and Cepstral Mean Normalisation</i>	51
4.16	<i>Effect of number of speakers on mean of Bhattacharyya bound for different phone groups using 100 utterances per speaker applying bootstrapping and Cepstral Mean Normalisation</i>	51
5.1	<i>Effect of number of phone utterances per speaker on mean of Bhattacharyya bound for different phone groups using data from 30 speakers</i>	54
5.2	<i>Effect of number of speakers on mean of Bhattacharyya bound for different phone groups using 20 utterances per speaker</i>	55
5.3	<i>Effective distances in terms of the mean of the Bhattacharyya bound between a single phone (/n/-nbl top and /a/-nbl bottom) and each of its closest matches within the set of phones investigated</i>	56
5.4	<i>Effective distances in terms of the mean of the Bhattacharyya bound between a single phone (/m/-nbl top and /i/-nbl bottom) and each of its closest matches within the set of phones investigated</i>	57
5.5	<i>Effective distances in terms of the mean of the Bhattacharyya bound between a single phone (/g/-nbl top and /s/-nbl bottom) and each of its closest matches within the set of phones investigated</i>	58
5.6	<i>The influence of a reduction in training corpus size on phone recognition accuracy . . .</i>	59
A.1	<i>Effect of number of phone utterances per speaker on mean of Bhattacharyya bound for the parent phones /ah/, /n/, /s/ and /p/ using data from 20 speakers</i>	68
A.2	<i>Effect of number of phone utterances per speaker on mean of Bhattacharyya bound for different phone groups using data from 30 speakers</i>	68
A.3	<i>Effect of number of phone utterances per speaker on mean of Bhattacharyya bound for different phone groups using data from 30 speakers</i>	69
A.4	<i>Effect of number of phone utterances per speaker on mean of Bhattacharyya bound for different phone groups using data from 30 speakers</i>	69
A.5	<i>Effect of number of speakers on mean of Bhattacharyya bound for different phone groups using 20 utterances per speaker</i>	70

A.6	<i>Effect of number of speakers on mean of Bhattacharyya bound for different phone groups using 20 utterances per speaker</i>	70
A.7	<i>Effect of number of speakers on mean of Bhattacharyya bound for different phone groups using 20 utterances per speaker</i>	71
A.8	<i>Effective distances in terms of the mean of the Bhattacharyya bound between a single phone (/n/-afr top and /a/-afr bottom) and each of its closest matches within the set of phones investigated</i>	72
A.9	<i>Effective distances in terms of the mean of the Bhattacharyya bound between a single phone (/n/-ssw top and /a/-ssw bottom) and each of its closest matches within the set of phones investigated</i>	72
A.10	<i>Effective distances in terms of the mean of the Bhattacharyya bound between a single phone (/n/-ven top and /a/-ven bottom) and each of its closest matches within the set of phones investigated</i>	73
A.11	<i>Effective distances in terms of the mean of the Bhattacharyya bound between a single phone (/n/-zul top and /a/-zul bottom) and each of its closest matches within the set of phones investigated</i>	73
A.12	<i>Effective distances in terms of the mean of the Bhattacharyya bound between a single phone (/m/-ssw top and /i/-ssw bottom) and each of its closest matches within the set of phones investigated</i>	74
A.13	<i>Effective distances in terms of the mean of the Bhattacharyya bound between a single phone (/m/-ven top and /i/-ven bottom) and each of its closest matches within the set of phones investigated</i>	74
A.14	<i>Effective distances in terms of the mean of the Bhattacharyya bound between a single phone (/m/-zul top and /i/-zul bottom) and each of its closest matches within the set of phones investigated</i>	75
A.15	<i>Effective distances in terms of the mean of the Bhattacharyya bound between a single phone (/d/-afr top and /s/-afr bottom) and each of its closest matches within the set of phones investigated</i>	75
A.16	<i>Effective distances in terms of the mean of the Bhattacharyya bound between a single phone (/b/-ssw top and /s/-ssw bottom) and each of its closest matches within the set of phones investigated</i>	76
A.17	<i>Effective distances in terms of the mean of the Bhattacharyya bound between a single phone (/g/-zul top and /s/-zul bottom) and each of its closest matches within the set of phones investigated</i>	76

LIST OF TABLES

3.1	<i>Difference between biased (Equation 3.5) and un-biased estimates of the cumulative standard deviations of the bound at 400 speaker permutations</i>	34
3.2	<i>Standard deviations for phones of stability measurements at 400 speaker permutations</i>	35
4.1	<i>Comparative inter- and intra-speaker variability for different phone types</i>	41
4.2	<i>Number of samples required for accurate estimation of similarity values</i>	43
5.1	<i>A summary of audio data available for different languages</i>	53
5.2	<i>Initial results for South African ASR systems. The column labelled “avg # phones” lists the average number of phone observations for each phone for each speaker</i>	59
A.1	<i>Triphone clusters for different contexts of the phone /ah/</i>	66
A.2	<i>Triphone clusters for different contexts of the phone /n/</i>	67
A.3	<i>Triphone clusters for different contexts of the phone /s/</i>	67
A.4	<i>Triphone clusters for different contexts of the phone /p/</i>	71

CHAPTER ONE

INTRODUCTION

1.1 CONTEXT

When building speech recognition systems for previously unsupported languages, data scarcity is a problem. Even if the required audio can be obtained, just labelling the data are still a cumbersome process. Due to these factors, the decisions that are made during automatic speech recognition (ASR) corpus design are of key importance. Current speech recognition systems represent audio training data in terms of acoustic models. Depending on the application, these models are typically defined on a phone level and are of various complexities. Encoded audio data for each separate phone is accumulated and then associated with an acoustic model, during the training process, as defined for the particular speech recognition system.

Linguistically, phones can be divided into different categories: Vowels, Nasals, Plosives, Fricatives, etc. For the various phone categories, it is easy to grasp that training data representing a specific phone type will have different qualities. For example vowels are voiced sounds, typically requiring much more time to be realised than, for example, a plosive sound that might be voiced or unvoiced. With currently used encoding schemes, the effect of these differences ultimately translates to different data requirements for the acoustic models of the various phones of an ASR system.

Due to data scarcity problems, work has been done on data sharing between languages. Given the phonetic similarities that span across families of related languages, it makes sense to investigate these approaches for speech recognition of resource-scarce languages. When no target language training data exists, an ASR system trained on a different source language can be used directly (cross-language transfer). Not surprisingly cross-language transfer suffers heavily from the acoustic differences that exist between the sound systems of source languages. Also, not necessarily all sounds that are needed by the target language are catered for. Therefore other approaches try to use data from the target language (if such resources exist), no matter how limited these resources may be.

It is important to give careful consideration to the acoustic differences that exist between the recordings that are being used as training data. In the multilingual case these differences include: The phonetic inventories of the different languages, pronunciation differences between similar phones of the source and target languages and channel effects (different speakers and microphones). When attempting to generalise speech recognition training methods for the multilingual case these factors represent significant obstacles. This is also the reason why trying to select the right chunks of data to train on linguistically, quickly results in complex annotations [2].

Data-sharing approaches that include data-driven techniques are more successful because of the acoustic distance which accounts for pronunciation differences, speaker differences and also inherently models channel effects are taken into consideration as well. These methods serve to generate sophisticated groupings that fit the acoustic characteristics of the target language. In literature, the strategies that have been used to accomplish acoustic modelling across languages range from linguistic model selection to acoustic model adaptation, data pooling and combinations of approaches. Although the combination of training data in this fashion is based on good assumptions, substantial experimentation has to be performed with regard to optimisation. During this process the effect of previous assumptions on the particular data sets might be difficult to judge from end results or may even be disproved.

1.2 PROBLEM STATEMENT

We are interested in the data requirements when developing ASR systems in resource-scarce environments. As the question “How much data is enough?” is too broad to answer, we focus on an analysis of the effect of additional data on acoustic model stability. Such an analysis should address the type and amount of training data used to construct models as well as acoustic model complexity. Given this context, defining distance metrics between different acoustic models is a viable option to investigate the acoustic properties of training data. Such metrics will, however, have to take into account data sufficiency concerns. Acoustic models eventually reach a state where they remain stable when additional training data are added. At this point of “data sufficiency” the model optimally represent the training data.

Specifically, the main objectives of this dissertation are:

1. The development of an acoustic stability measure that can be used during data sufficiency analysis.
2. Utilizing the above measure to analyse the effect of different types and amounts of training data on the stability of acoustic models of varying complexity.
3. Performing a data sufficiency analysis on the newly developed Lwazi ASR corpus, and correlating analytical results with ASR accuracies observed.

1.3 LITERATURE REVIEW

This literature review describes prior work related to acoustic modelling with limited training data and focuses on four main areas of interest:

1. Analysis of acoustic models,
2. ASR data requirements,
3. Cross-language sharing of data for ASR systems, and
4. The motivation for efficient ASR corpus design techniques

1.3.1 ACOUSTIC MODEL ANALYSIS

1.3.1.1 ACOUSTIC MODELS IN ASR

A core component of current speech recognition architecture is the acoustic modelling component, designed to classify much of the variability normally found within speech data. Before any acoustic modelling can be performed, training data are converted into a useful representation during a process called encoding. During encoding, the useful information is extracted from the audio recordings as features that undergo changes over time. Sound groupings of the training data feature observations can then be made (many of which can also be linguistically motivated) and associated with specifically designed acoustic models to represent the variability seen in the training data.

A popular type of acoustic model structure currently used is the Hidden Markov Model (HMM). An HMM is a finite state machine which changes state every time instance [3]. To allow for specific state transitions, a state topology is defined. In addition, state transition probabilities also characterise transition behaviour between the states. Some states are associated with a probability density estimator. These states are called emitting states and are capable of classification. The complex structure of the HMM thus enables temporal modelling of variability (commonly found within the speech signal) via its ability to perform classification using different density estimators at different time instances.

For phone recognition, a single HMM model is generally constructed for every acoustically different sound unit defined for the specific recognition problem. Because of the fact that the speech spectra of a single phone is still a time-varying sequence, an HMM topology of three emitting states is normally used. Any parametric density estimator can be used as output probability distributions, but these models may lack the flexibility to obtain added model complexity. Due to easy generalisation, Gaussian mixture models (GMMs) are a very common implementation choice, to achieve good performance with a continuous density HMM speech recognition system. Also, due to co-articulation effects, the definition of context-dependent phone models are found to be useful [4]. Automatic speech recognition toolkits are available that uses the HMM as an acoustic model when building speech recognition systems. One such implementation choice is the Hidden Markov Model Toolkit (HTK) [3].

The above speech recognition architecture already adjusts (to a degree) to the amount of training data available during acoustic model training. In practice there exists a data insufficiency problem due to the resulting large number of model parameters within the HMM framework. Also, the training data are usually unevenly spread so that a method is required to balance model complexity against data availability [4]. This happens both at the triphone level, when context dependency is taken into account and during density estimation, when more or less mixtures can be used to model the training data.

Triphone tying typically utilises tree based clustering techniques [4]. First of all the training data are pooled, combining similar emitting states of all triphone models based on a single monophone model. At this stage, each mono-phone state is typically modelled with a single Gaussian density estimator. The decision tree then tries to subdivide these data pools (directed by a predetermined set of questions), calculating an entropy measure and limiting the minimum amount of data per pool via a threshold on the data occupation counts. The end result is an inventory of distributions that can be used for groups (clusters) of triphones.

It follows that as training data increases and triphone models are better estimated, the amount of acoustically different clusters will increase up to a point where the definition of the sound units (triphones) for the particular task breaks down. If the amount of data allows, increased accuracy can then be obtained by incrementing the Gaussian mixtures of these models.

1.3.1.2 MODEL ANALYSIS TECHNIQUES

For a specific task, the question that arises with regard to acoustic measurements is to what extent these measurements accurately portray the differences that are relevant during classification. For speech recognition the ideal would be to analyse the acoustic models being used within a system directly. This can be a difficult task, given the complexity of the HMM models used and the lack of stability of these models because of data insufficiency.

Several techniques exist that allow the comparison of density estimators. One of these, the Bhattacharyya distance is an important measure of separability between two single mixture Gaussian distributions. It can be derived analytically from the Bhattacharyya bound, which in turn is defined as an upper bound on the Bayes error [5]. Calculation of the bound is tractable between single Gaussians, but intractable for complex distributions such as Gaussian mixture models used in speech recognisers. One can estimate the Bhattacharyya bound via Monte Carlo sampling techniques [6]. If the number of samples drawn from the represented distributions is sufficient, the technique should converge to the actual Bhattacharyya bound. The Bhattacharyya bound has been used successfully to analyse acoustic models on a phone level [7].

Prior work in music analysis also exploits Gaussian mixture models (GMMs) [8] and Hidden Markov Models (HMMs) [9] to represent a piece of music. As in speech recognition, these models are then compared to find similarity between different pieces of music. To do so, Monte Carlo approaches are also used, but here training data feature observations (MFCCs for this specific implementation)

are used as samples and the likelihood of these samples then calculated as similarity measure on the acoustic models of another piece of music. These approaches are then further extended incorporating a Markov chain to measure similarity between HMMs as well.

Approaches to measure the distance between HMMs directly are described in [10], [11] and [12]. Here the authors make use of the Kulback-Liebler divergence between two probabilistic models. This definition is then expanded to a divergence measure between HMMs. An HMM can be seen as a first order Markov chain with a number of states and random variables related to each of these states. Various assumptions are made in the attempts to model this behaviour. Simulations using Monte Carlo methods are also employed.

1.3.2 DATA REQUIREMENTS FOR ASR

Given the current HMM-based modelling of speech sounds (phones) within speech recognition systems, a definite concern is the amount of data that is actually required to achieve sufficient recognition accuracy. A perspective on this work is provided by data selection research that focuses on the optimal selection of a subset of training data from a larger corpus, in order to better understand the effect of amount of training data on ASR accuracy. Armed with the knowledge obtained, corpus design becomes possible, where the size of training corpora can be minimized. This is achieved through various techniques that aim to include as much variability in the data as possible, while ensuring that the corpus matches the intended operating environment as accurately as possible.

1.3.2.1 DATA SELECTION STRATEGIES

With regard to corpus design, there are three data selection strategies that are primarily employed:

1. Explicit specification of phonotactic, speaker and channel variability during corpus development,
2. automated selection of informative subsets of data from large corpora, with the smaller subset yielding comparable results [13],
3. or the use of active learning to improve the accuracy of existing speech recognition systems [14].

All three of the techniques provide a perspective on the sources of variation inherent in a speech corpus, and the effect of this variation on speech recognition accuracy.

In [13], Principal Component Analysis (PCA) is used to cluster data acoustically. These clusters then serve as a starting point for selecting the optimal utterances from a training database. As a consequence of the clustering technique, it is possible to characterise some of the acoustic properties of the data being analysed, and to obtain an understanding of the major sources of variation, such as different speakers and genders. Interestingly, the effect of utterance length has also been analysed as a main source of variation [14].

Active and unsupervised learning methods can be combined to circumvent the need of transcribing massive amounts of data [14]. The most informative untranscribed data is selected for a human to label, based on acoustic evidence of a partially and iteratively trained ASR system. From such work it becomes evident that the optimisation of the amount of variation inherent to training data is needed, since randomly selected additional data does not necessarily improve recognition accuracy. By focusing on the selection (based on existing transcriptions) of a uniform distribution across different speech units such as words and phonemes, improvements are obtained [15].

1.3.2.2 THE EFFECT OF CONFUSABILITY

In [16] the authors present a novel framework for predicting speech recognition errors, with the eventual goal of developing a metric of lexical confusability that takes into account information from acoustic, pronunciation, and language models of the recogniser.

Lexical confusability is related to and compounds acoustic confusability, due to the fact that the goal of speech recognition is ultimately to make sense of an acoustic speech signal. Increased lexical confusability will thus result in more speech recognition errors. Adding new pronunciation variants to a pronunciation dictionary for example, leads indirectly to a higher acoustic confusability, because more pronunciation variance is allowed [16].

Language models suffer from the same lexical confusability effects as pronunciation dictionaries. Similar words are confusable, because of the similarities of their respective pronunciations. The pronunciation complexities of words also differ for various languages. There exists an interplay between the specific phone set for a language, as used with a speech recognition system, and the pronunciations of words using these phone sets.

Phone sets lead to different levels of acoustic confusability. Some languages may even have various highly confusable phones, from an ASR point of view. Therefore the definition of sound units used within an ASR system is important. Ideally the optimal set is those definitions that allow for the discriminate classification between the different variations being modelled. In order to optimise the definition of sound units, a merging and splitting technique is defined in [17]. This optimisation algorithm employs a Monte Carlo based entropy metric (also known as the KL2 metric, previously used in problems such as the segmentation of acoustic data), to compute distances between phone-based HMMs. A Monte Carlo estimation approach is used since there exists no closed form expression for many of the distance measures between more complicated distributions such as Gaussian mixtures or samples of a non-stationary random process [17]. Instead, to simulate HMM behaviour, a generated sequence from the HMM being considered is analysed. Comparing HMMs in this fashion however, has a high computational cost and is slow with regard to convergence [11].

1.3.3 CROSS-LANGUAGE SHARING OF DATA

In this section we discuss some of the main strategies for cross-language data sharing, including language-independent and language-adaptive modelling. The goal of language-independent mod-

elling is to create a combined acoustic model suitable for the simultaneous recognition of various languages. In contrast, the goal of language-adaptive modelling is the adaptation of pre-existing models towards an optimal recognition of a new target language, using only limited adaptation data from this target language [18]. In all approaches, the way in which the various cross-language units are mapped to one another has a significant influence on results achieved.

1.3.3.1 LANGUAGE-INDEPENDENT MODELLING

During language-independent modelling, a combined acoustic model is created suitable for simultaneous recognition of multiple languages. When attempting to generalise existing speech recognition modelling methods for this purpose, we are faced with two main tasks: determining suitable seed models for the initialisation of acoustic models for a target language, and dealing with the potentially large phonetic mismatch between the phonetic inventories of the source languages and the target language [18].

Within the multilingual domain, the definition of acoustic units used for the interpretation of the training data, becomes all the more important. Usually there is a greater number of units that needs to be defined between all of the languages. A more difficult question is how to combine similar acoustic units within this domain, especially at the context-dependent triphone level. Subtle differences exist due to pronunciation differences. This is why just trying to select the right chunks of data to train on linguistically, quickly results in complex annotations [2].

In literature various approaches have been taken to try and facilitate efficient data sharing between languages. Generally more linguistic information is involved and selections are made according to linguistic or acoustic measures between the units.

1.3.3.2 DATA POOLING

Data pooling, a popular data sharing strategy, simply pools data from various sources together during ASR system development. When this is done, the way in which the various cross-language units are mapped to one another influences results significantly. Initial approaches selected similar sounds on a phone level using the International Phonetic Alphabet (IPA) [18–20]. The results for doing so however, leads to deteriorations in performance [18,20]. In [19] some improvement is obtained when using closely related speech databases. It should be noted, however, that the amount of target data available to this experiment has been very limited. An acoustic mismatch delivered poor results for a similar experiment with another speech database.

Linguistically-motivated questions to cluster related contexts during decision tree clustering have the advantage of discriminating between acoustically similar and different phones between languages based on an entropy measure. This is accomplished by tagging the phones of each language to preserve the information about the language they belong to, and then testing the effect that clustering specific groups of phones together has on the entropy measure [18,20]. Results indicate a significant improvement over the pure model selection case and even comparable results to a baseline system.

Experiments were also conducted to cluster related contexts “from the bottom up” or agglomerative clustering. That is, grouping together contexts based on a similarity measure. This is different from the approach followed with decision tree clustering, where data is first pooled and then these pools are broken down into smaller clusters [21]. Comparable results to tree based clustering methods are obtained.

Combining IPA-based phone selection with acoustic distance measures to create effective mappings or clustering between phonetic inventories of source data have also been tried. These methods are most successful when the phone inventory of the source and target languages has maximal overlap [19]. Normally a confusion matrix between the phones of two languages is constructed and the phone mappings between the target and source languages are then derived from this matrix.

Inter-language models that incorporate acoustic information between similar phones of different domains, possibly at the sub-phone level, by sharing states and distributions rather than whole models show promise. The technique is especially valuable if the target language contains phones not found in any of the source languages [22]. The effect of this technique is fairly similar to model adaptation, in the sense that models are partly altered to be more robust in the domain they are intended for.

The amount of source data, amount of target data, the acoustic mismatch between the various languages and the level of confusability inherent in the target language are all factors that combine to affect the success of these various approaches.

1.3.3.3 LANGUAGE-ADAPTIVE MODELLING

For language adaptive-modelling, the goal is to utilise source and target language data in such a way that the optimal recognition of the target language with the available resources becomes possible. According to [18], depending on the amount of training data available there are three approaches: Cross-language transfer, language adaptation and bootstrapping. Cross-language transfer is used when no target training data are available. Research in this area investigates whether source data from the same family of languages can be used to recognise the target language data. Language adaptation is used when a very limited amount of target language training data are available. The main challenge then is to identify suitable acoustic models to start from, after which general model adaptation techniques are used. Lastly the bootstrapping approach can be followed when a sufficiently large amount of target language data do exist. During this approach models are initialized with carefully selected training data of similar acoustics as the intended end models. It has been proved that cross-language seed models (applying bootstrapping) achieve lower word error rates than flat starts or random models [18].

1.3.3.4 MODEL ADAPTATION

It is possible to use model adaptation methods to smooth out the differences between models of source and target data. Care has to be taken not to lose important acoustic characteristics of the target language in the process. Model-based speaker adaptation algorithms modify the parameters

of HMMs and can be divided into the maximum a posteriori (MAP) adaptation family, parameter transformation based adaptation and a family related to speaker clustering techniques [23]. Two candidate adaptation techniques that have been widely used in literature are mean-only MAP adaptation and Maximum Likelihood Linear Regression (MLLR) [23]. These fall into the MAP adaptation and parameter transformation based adaptation families respectively.

Techniques that use variants of MAP adaptation to model and correct for channel factors using feature transformations have also been applied successfully. Since some feature-based transformations, such as feature warping, do not rely on a specific model, they can be used as an additional front-end processing step for any recognition system that takes advantage of this compensation technique [24].

Eigenspace-based methods that have also proved useful form part of the speaker clustering family. These are powerful techniques to analyse variations, typically applied to speaker adaptation of acoustic models. One such adaptation technique, referred to as Eigenvoices [25], uses Principal Component Analysis (PCA) to convert acoustic model parameters to a lower dimensional space. A weighted sum of a cluster of acoustic models is then computed and this interpolated model (eigenvoice) is then used to represent the acoustic characteristics of that particular cluster. Adaptation is performed by estimating a weighted combination of the chosen eigenvoices.

1.3.4 MOTIVATION FOR EFFICIENT ASR CORPUS DESIGN TECHNIQUES

Developing countries, such as the African countries, can potentially benefit from the use of spoken dialog systems (SDSs) [26]. The general idea of SDSs is to provide speech-based access to electronic information through the use of a telephone. Access to traditional computer infrastructure in Africa is low, but telephone networks and especially cellular networks are spreading rapidly. This is the main reason for the belief that SDSs can have a significant impact in these countries and may even empower illiterate or semi-literate people, 98% of whom live in the developing world.

In developing countries a wide range of applications for SDSs exists. The most important of these include: education (using speech-enabled learning software or kiosks), information systems in agriculture [27], health care [28, 29] and government services [30].

In order to realise SDSs in Africa, technology components such as text-to-speech (TTS) systems and automatic speech recognition (ASR) systems are required. ASR systems for these applications require specific resources: electronic pronunciation dictionaries, annotated audio corpora and recognition grammars. For the African languages, speech recognition systems are only available for a very limited number and to our knowledge no service available to the general public currently uses ASR in an indigenous African language.

A major drawback for the development of ASR systems in the indigenous African languages is the lack of linguistic resources and audio corpora needed to build these systems. Modern speech recognition systems require the use of training data relevant to the application (speech that is appropriate for the recognition task in terms of the language used, profile of the speakers, speaking style,

etc.). The audio data has to be carefully transcribed orthographically with the use of markers to indicate important events (such as non-speech sounds) that might occur in the data. Also, for most applications, a large number of speakers is required to obtain good speaker-independent recognition accuracies.

Given the above context, it is no surprise that the complexity related to audio corpora design is highly correlated with the amount of data being required, because of the difficulties experienced with recruitment of multiple speakers and the curative transcription of their respective speech utterances. These factors are major stumbling blocks for speech technology development.

Consequently, the development of tools and guidelines to minimise the complexity related to ASR corpus design is a valuable contribution. The ability to design the smallest corpora that will be sufficient for typical ASR applications is of great importance. For this specific domain, techniques that share data across languages are in demand as minimal corpora can be extended and are expected to benefit by sharing data across languages. As mentioned in the previous sections, subtle acoustic differences exist between similar phones of different languages and consequently these tools need to indicate when data sharing will be beneficial or detrimental.

1.4 OVERVIEW OF DISSERTATION

The structure of the dissertation is as follows:

- Chapter 2 describes the measurement of model differences. The specific strategies to calculate discrete values that serve to describe model separability are given. Importantly a distinction is also made between distance metrics and similarity measures. The distance metric defined forms the basis for the model stability measure, defined in the next chapter.
- In Chapter 3 different types of ASR data variability are introduced. Current approaches to effectively model ASR data are discussed. Finally, in the context of these two factors (data variability and model complexity) a new model stability measurement protocol is defined and evaluated.
- In Chapter 4 the effect of different amounts and types of training data on the stability of acoustic models is investigated. The defined stability measurement protocol is applied to analyse the effectiveness of different types of acoustic models in modelling various different sets of training data.
- The insights gained from the data sufficiency experiments performed in previous chapters are finally applied in Chapter 5 to perform a data sufficiency analysis of a new telephone-based ASR corpus (The Lwazi ASR corpus). Based on the findings, experiments are extended to include distances between phones of different languages. The results of initial ASR systems for these different languages are also presented to show the correlation between phone recognition results and the amount of ASR training data.

- Finally, we conclude in Chapter 6 and summarise the most relevant literature, our contributions, as well as possible future work that may be conducted.

CHAPTER TWO

MODEL DISTANCE ESTIMATION

2.1 INTRODUCTION

Speech recognition systems represent the acoustic information obtained from training data as acoustic models. Within current ASR systems, the specific definitions of these models are generally based on sound assumptions about acoustically dissimilar properties found within the speech data. Model complexities vary between implementations and applications, but ultimately probability densities are used to represent the acoustic information. It follows that if distances can be estimated between such densities, it should be possible to investigate the estimation (learning behaviour) of these densities on various selections of data and eventually obtain a better understanding of the data requirements of models. This chapter describes the distance metrics that are applied between various acoustic models throughout the remaining chapters to follow.

The chapter is structured as follows:

- Section 2.2 provides the theoretical background associated with the Bhattacharyya Bound;
- Section 2.3 deals with calculation of the bound for single Gaussian Models on speech data; and
- In Section 2.4 estimation of a similarity measure for Gaussian Mixture Models is discussed.

2.2 BACKGROUND

2.2.1 DISTANCE METRICS AND SIMILARITY MEASURES

As mentioned in Section 1.3.1.2 various techniques have been described to accurately classify the differences between acoustic models similar to those used in ASR systems. Any acoustic model is, however, just a representation of a set of training data which captures the desired acoustic information.

Depending on the amount of training data and the acoustic model complexity, the success of the intended representation may vary.

Model complexity adds to the problem, because of the tractability of approaches to analyse model differences for various acoustic models. Analytical bounds that describe the separability between probability densities are available for simple parametric density estimators. When more complex density estimators are employed, Monte Carlo sampling methods can be employed to estimate a similar separability. Sampling introduces added variability to the measurements that need to be dealt with. As described in Section 1.3.1.2 distance measures have even been defined for complex HMM models. As with sampling, these strategies differ with regard to distance variability.

Given the above, distance measurements between different acoustic models are at best an estimation. Various factors come into play, from the actual data used for model construction up to effects introduced by the very different approaches to distinguish between models. When measurements on data are considered (in chapters to follow), we refer to similarity measurements. The main reason for this is that we work with the Bhattacharyya bound values directly, which are greater for higher similarity (contrary to distance, which is a measure of dissimilarity). The measurements also compound the introduced effects of model construction and measurement strategies.

2.2.2 THE BHATTACHARYYA BOUND

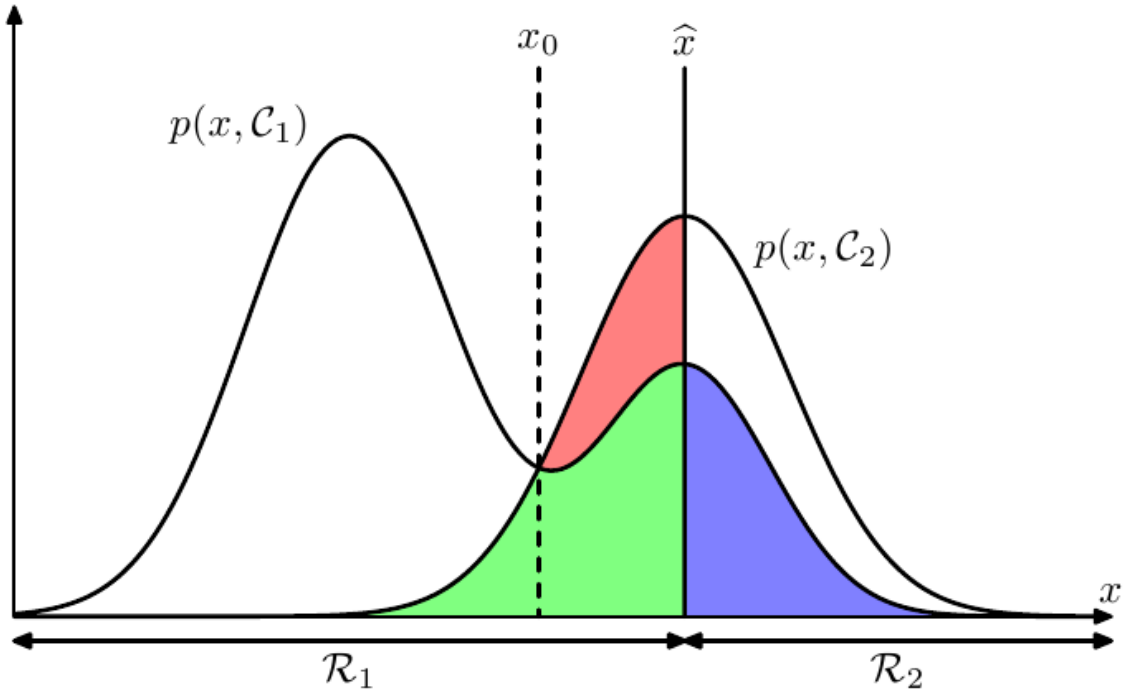


Figure 2.1: *Decision rule: Minimum misclassification at the optimal decision boundary x_0 (Reproduced from Bishop, 2006 [1])*

The separability between two probability density functions can be measured by a widely-used

upper bound of the Bayes error, namely the Bhattacharyya bound [7]. If we view two probability distributions in the light of a classification problem where the first probability distribution P_1 represents the probability of class C_1 and the second probability distribution P_2 represents the probability of class C_2 , then the Bayes error is the minimum misclassification rate¹.

In Figure 2.1, this concept is being illustrated: If we take $\hat{x}$ to be the decision boundary and consider the decision regions R_1 and R_2 , then all values of x within R_1 is classified as belonging to C_1 and similarly all x in R_2 as C_2 . For this particular case, the coloured surfaces all represent regions where x can be misclassified. For $x < \hat{x}$ points of C_2 are classified as C_1 and for $x > \hat{x}$ points of C_1 as C_2 . It is possible to optimise this classification algorithm by simply moving the decision boundary. Optimal classification is possible at the indicated decision boundary $\hat{x} = x_0$, because of the fact that the red region disappears and the combined area of the green and blue regions remains constant when the decision boundary is moved for the particular example. The probability of a misclassification for a particular decision boundary as represented in the figure is therefore given by [1]:

$$\begin{aligned} p(\text{error}) &= p(x \in R_1, C_2) + p(x \in R_2, C_1) \\ &= \int_{R_1} p(x, C_2)dx + \int_{R_2} p(x, C_1)dx \end{aligned} \quad (2.1)$$

with $p(x, C_1)$ and $p(x, C_2)$ denoting the class conditional density functions for the two possible classes respectively. Consequently the Bayes error ϵ is then this probability with the decision boundary at the optimal location using prior probabilities P_1 and P_2 respectively and is represented by:

$$\epsilon = \int \min[P_1 p(x, C_1), P_2 p(x, C_2)]dx \quad (2.2)$$

Following the derivation in [5], an upper bound on the Bayes error ϵ can be derived due to the fact that the geometric mean of two positive numbers is larger than the smaller number. Equation 2.3 states this relationship for two numbers a and b respectively:

$$\min[a, b] \leq a^s b^{1-s} \quad 0 \leq s \leq 1 \quad (2.3)$$

When the above relationship between two positive numbers is applied to the Bayes error we obtain the Chernoff bound:

¹Note that this means that two very similar models will approach the maximum bound of 0.5 (an expected 50% misclassification rate) while two very different models will have lower bounds, possibly even approaching 0 (an expected 0% Bayes error)

$$\epsilon = P_1^s P_2^{1-s} \int p^s(x, C_1) p^{1-s}(x, C_2) dx \quad 0 \leq s \leq 1 \quad (2.4)$$

The Chernoff bound still requires selection of the optimal parameter s . Further approximation can be obtained by the relaxation of this condition and choosing $s = 0.5$. This simplified bound is then referred to as the Bhattacharyya bound:

$$\epsilon_\mu = \sqrt{P_1 P_2} \int \sqrt{p(x, C_1) p(x, C_2)} dx \quad (2.5)$$

We utilise this bound directly as a similarity measure between two probability density functions. For simple parametric density functions, analytical equations can be used to evaluate the bound. It is also possible to estimate the bound for more complex functions such as Gaussian mixture models. In Section 2.4 such a similarity measure is derived.

2.3 DISTANCE ESTIMATION FOR SINGLE GAUSSIAN MODELS

For single Gaussian models a closed-form expression of the Bhattacharyya bound exists that can be evaluated analytically. If both the probability density functions $p(x, C_1)$ and $p(x, C_2)$ are Gaussians with mean and covariance matrices of the form [1]:

$$p(x) = N(x|\mu, \Sigma) = \frac{1}{2\pi^{D/2}} \frac{1}{|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right) \quad (2.6)$$

integration of Equation 2.5 leads to a closed-form expression for ϵ_μ . This can be written as [5]:

$$\epsilon_\mu = \sqrt{P_1 P_2} e^{-\mu(1/2)} \quad (2.7)$$

The function $\mu(1/2)$ of Equation 2.7 is referred to as the Bhattacharyya distance [5]:

$$\begin{aligned} \mu(1/2) = & \frac{1}{8}(M_2 - M_1)^T \left[\frac{\Sigma_1 + \Sigma_2}{2} \right]^{-1} (M_2 - M_1) \\ & + \frac{1}{2} \ln \frac{|\frac{\Sigma_1 + \Sigma_2}{2}|}{\sqrt{|\Sigma_1| |\Sigma_2|}} \end{aligned} \quad (2.8)$$

Equation 2.8 is evaluated easily, only requiring the mean vectors (M_1 and M_2) and covariance matrices (Σ_1 and Σ_2) of the respective probability densities. Taking the negative exponent and multiplication by the prior probabilities P_1 and P_2 as defined in Equation 2.7 yields an estimated value for

the Bhattacharyya bound.

2.4 DISTANCE ESTIMATION FOR GAUSSIAN MIXTURE MODELS

The Bhattacharyya bound between two complex Gaussian mixture models cannot be evaluated directly. This is because integration of Equation 2.5 no longer has a closed form when [1]:

$$p(x) = \sum_{k=1}^K \pi_k N(x|\mu_k, \Sigma_k) \quad (2.9)$$

It is possible to estimate the bound via Monte Carlo simulation. Using the sample value of expectation of the integral we derive a Bhattacharyya estimator².

2.4.1 BHATTACHARYYA ESTIMATOR

Given the Bhattacharyya bound and prior probabilities P_1 and P_2 respectively, we assume $P_1 = P_2 = 0.5$ (and assume this explicitly in the remainder of our work). Equation 2.5 then becomes:

$$\epsilon_\mu = 0.5 \int \sqrt{p(x, C_1)p(x, C_2)} dx \quad (2.10)$$

We use the sample value expectation as:

$$\frac{1}{n} \sum_{i=1}^n g(x_i) \approx \int g(x)p(x)dx \quad (2.11)$$

where the expectation of the sample values x_i approaches the expectation of the arbitrary function $g(x)$ with regard to the probability mass function $p(x)$. Substituting $g(x) = \frac{f(x)}{p(x)}$, Equation 2.11 can be written in the form of:

$$\int f(x)dx \approx \frac{1}{n} \sum_{i=1}^n \frac{f(x_i)}{p(x_i)} \quad (2.12)$$

where x_i is drawn from $p(x)$.

If we let: $f(x) = \sqrt{p(x, C_1)p(x, C_2)}$, Equation 2.5 can be rewritten as:

²Note that x can be multi-dimensional

$$\begin{aligned}
\epsilon_\mu &= \frac{0.5}{n} \sum_{i=1}^n \frac{f(x_i)}{p(x_i)} \\
&= \frac{0.5}{n} \sum_{i=1}^n \frac{\sqrt{p(x_i, C_1)p(x_i, C_2)}}{p(x_i)} \\
&= \frac{0.5}{n_1 + n_2} \left[\sum_{i=1}^{n_1} \frac{\sqrt{p(x_i, C_1)p(x_i, C_2)}}{p(x_i, C_1)} + \sum_{j=1}^{n_2} \frac{\sqrt{p(x_j, C_1)p(x_j, C_2)}}{p(x_j, C_2)} \right] \\
&= \frac{0.5}{n_1 + n_2} \left[\sum_{i=1}^{n_1} \sqrt{\frac{p(x_i, C_2)}{p(x_i, C_1)}} + \sum_{j=1}^{n_2} \sqrt{\frac{p(x_j, C_1)}{p(x_j, C_2)}} \right] \tag{2.13}
\end{aligned}$$

where x_i and x_j are drawn from $p(x, C_1)$ and $p(x, C_2)$ respectively. In practice we calculate Equation 2.13 where x_i and x_j are the actual samples and both n_1 and n_2 are the number of samples with regard to each of the two probability density functions ($p(x, C_1)$ and $p(x, C_2)$) respectively.

Samples are generated on a per component basis, taking into account the mixture weight π_k as defined in Equation 2.9. For a total number of samples per model n , the number of samples m_k generated for any particular mixture component can be represented as:

$$m_k = n\pi_k \quad 0 \leq \pi_k \leq 1 \tag{2.14}$$

Equation 2.13 can then be rewritten in terms of Equation 2.9 as well as Equation 2.14 to yield the full similarity measure for the implementation of mixtures:

$$\begin{aligned}
\epsilon_\mu &= \frac{0.5}{n_1 + n_2} \left[\sum_{i=1}^{n_1} \sqrt{\frac{\sum_{k1=1}^{K_{C2}} \pi(k1, C_2) N(x_i | \mu_{k1}, \Sigma_{k1}, C_2)}{\sum_{k2=1}^{K_{C1}} \pi(k2, C_1) N(x_i | \mu_{k2}, \Sigma_{k2}, C_1)}} \right. \\
&\quad \left. + \sum_{j=1}^{n_2} \sqrt{\frac{\sum_{k3=1}^{K_{C1}} \pi(k3, C_1) N(x_j | \mu_{k3}, \Sigma_{k3}, C_1)}{\sum_{k4=1}^{K_{C2}} \pi(k4, C_2) N(x_j | \mu_{k4}, \Sigma_{k4}, C_2)}} \right] \tag{2.15}
\end{aligned}$$

For a typical implementation the number of samples for each of the two densities being compared are generally the same $n = n_1 = n_2$ as well as the total number of mixture components per model $K = K_{C1} = K_{C2}$, in which case:

$$\begin{aligned}
\epsilon_\mu = \frac{0.5}{2n} & \left[\sum_{i=1}^n \sqrt{\frac{\sum_{k1=1}^K \frac{m(k1, C_2)}{n} N(x_i | \mu_{k1}, \Sigma_{k1}, C_2)}{\sum_{k2=1}^K \frac{m(k2, C_1)}{n} N(x_i | \mu_{k2}, \Sigma_{k2}, C_1)}} \right. \\
& \left. + \sum_{j=1}^n \sqrt{\frac{\sum_{k3=1}^K \frac{m(k3, C_1)}{n} N(x_i | \mu_{k3}, \Sigma_{k3}, C_1)}{\sum_{k4=1}^K \frac{m(k4, C_2)}{n} N(x_i | \mu_{k4}, \Sigma_{k4}, C_2)}} \right]
\end{aligned} \tag{2.16}$$

2.4.2 VERIFYING THE ACCURACY OF ESTIMATED BHATTACHARYYA BOUND

In order to show that the derived Bhattacharyya estimator indeed converges to the actual Bhattacharyya bound values, we compare the estimated similarity values with analytically calculated values for a few specifically chosen density estimators. The density estimators are trained on ASR training data (39-dimensional MFCC-based speech vectors) similar to the training data used in Section 4.2. Figure 2.2 shows the difference between analytically evaluated and estimated Bhattacharyya bound values for single Gaussian model comparisons. Three very different model pairs are selected to cover a range of similarity values of the Bhattacharyya bound. The first pair of models is very similar with values close to 0.5, while the second pair is more distinct with values of approximately 0.3 and lastly the third pair is very different at values of below 0.1.

We evaluate Equation 2.13 for increasing numbers of samples. As the number of samples increases it can be seen that the similarity measure approaches the analytical values for all three of the comparisons. The very similar model pair (close to 0.5) requires fewer samples than the more dissimilar models (at 0.3 or 0.1 respectively).

2.4.3 NUMBER OF SAMPLES

For Gaussian mixture models, no analytical method is available to verify the estimated Bhattacharyya bound values as in Section 2.4.2. Now we rely on the stability of the measurement as an indicator of the number of samples that are needed for accurate estimation. Again a set of model pairs is selected to cover different similarity values of the Bhattacharyya bound and a series of bounds are estimated for every pair with increasing numbers of samples. The samples are generated using different sampling seed values. To estimate the stability of the measurement, it is necessary to repeat the experiment a number of times until the measurement can be estimated within an acceptable level of confidence.

We therefore estimate the mean of the bound (at a particular number of samples), over e experiment runs. If $\bar{x}_e = \frac{1}{e} \sum_{i=1}^e x_i$ is the mean of the e experiment runs and these means are evaluated for experiments of different model pairs sampled at a number of different sample sets J , then the variance between the estimated values of the runs are:

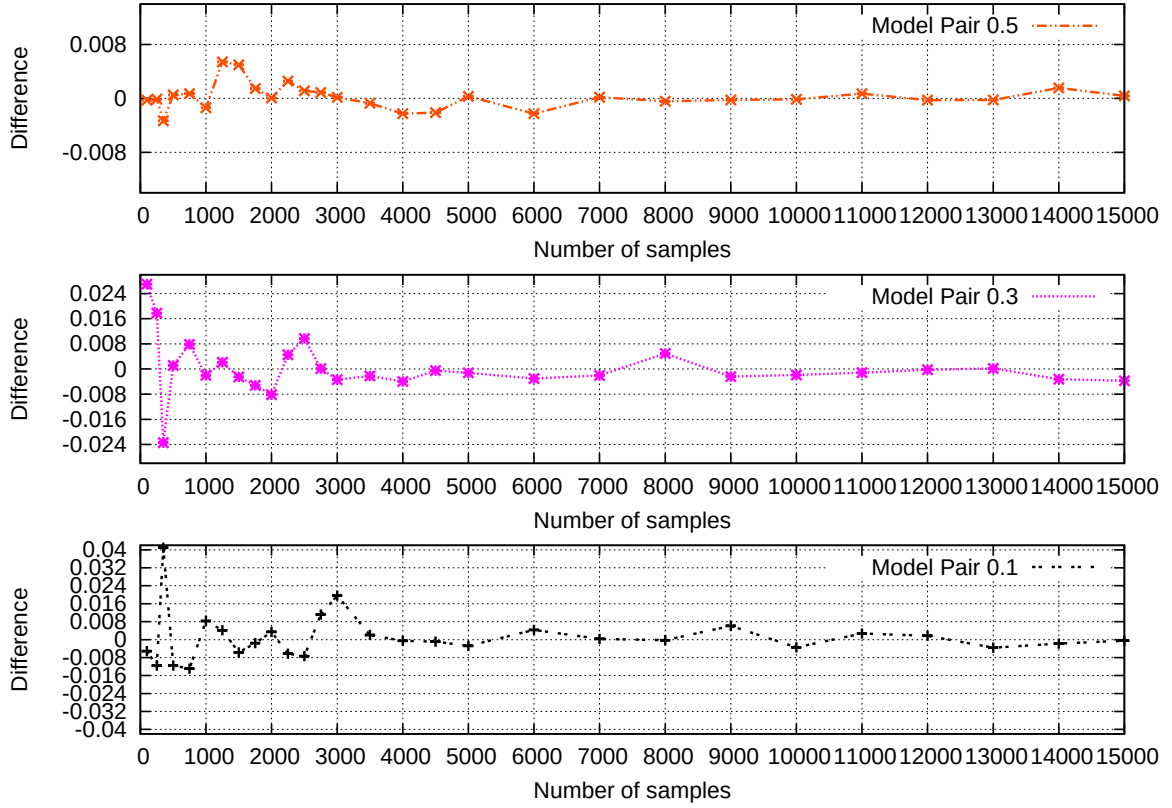


Figure 2.2: Difference between estimated distance measure and analytically calculated values for three different single Gaussian model comparisons and different numbers of samples

$$\text{Var}(\bar{x}_e) = \frac{1}{J} \sum_{j=1}^J (\bar{x}_{ej} - \bar{\bar{x}}_{eJ})^2 \quad (2.17)$$

where

$$\bar{\bar{x}}_e = \frac{1}{J} \sum_{j=1}^J \bar{x}_{ej} \quad (2.18)$$

Finally, in order to determine the sufficient number of samples that are required for the comparison of a particular model pair, the standard deviation of the similarity values can be calculated from the variance at any number of samples. It is then possible to select a number of samples where this standard deviation values fall below an acceptable threshold and consequently the sufficient estimation to allow for accurate model comparison is achieved. The complete process can be thus be summarised as follows:

1. Model pairs that cover a range of similarity values are selected.

2. For each model pair and number of samples s , we run e experiments that each estimates the similarity measure using different sampling seed values.
3. For each number of samples s we calculate the mean of the similarity measures over the e experiments.
4. The number of experiments e is increased until the deviation of the J different means (each associated with a different number of samples) falls below a certain threshold.
5. With e fixed, an estimate of the standard deviation $\hat{\sigma}_{e,s}$ over the e experiments for each s is determined separately.
6. The value of s where the value of $\hat{\sigma}_s$ from (5) falls below an acceptable threshold is chosen.

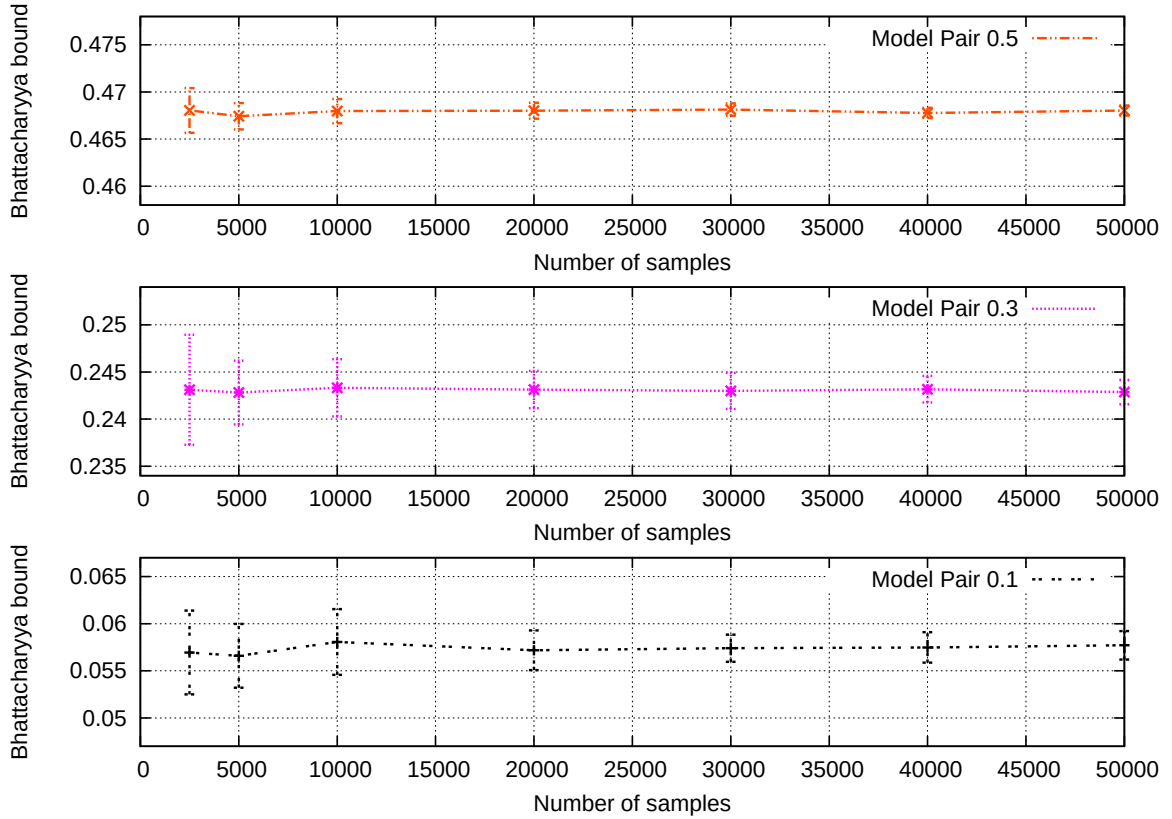


Figure 2.3: Estimation accuracy of distance measure on Gaussian mixture models containing 4 mixture components each ($e = 35$)

An experimental run on Gaussian Mixture models with four mixture components each is shown in Figure 2.3. The three model pairs are chosen to have three different similarity values on the bound of approximately 0.1, 0.3 and 0.5 respectively. At each testing point a specific number of samples are generated for every model. These differ from values of 2,500 to 50,000 samples each. With the number of experiments fixed at $e = 35$, the means are stable enough to accurately measure the standard deviations. It can be seen that many more samples are required for more dissimilar models

(model pairs 0.3 and 0.1) than for similar models (model pair 0.5). At 30,000 samples even these models have standard deviations of less than 0.005.

2.5 CONCLUSION

In this Chapter, strategies to evaluate model separability in terms of the Bhattacharyya bound are presented. We define and implement a Bhattacharyya bound estimator and evaluate its accuracy. We find that our estimator is highly accurate, both for single Gaussian models and for GMMs, if sufficient sampling is used. Modelling predetermined sets of data, combined with very specific similarity measurements can yield insight into data sufficiency requirements of acoustic models. In the next chapter we define a model stability estimator based on the similarity measure defined here.

CHAPTER THREE

MODEL STABILITY ESTIMATION

3.1 INTRODUCTION

Much is to be gained from an understanding of the role that various effects found within speech data play on the speech recognition parameters that are being used in ASR systems. Speech recognition parameters are generated to effectively represent speech data within speech recognition systems. For the current approaches, these parameters typically include features (the end product of speech encoding) and acoustic models (description and memorisation of the information obtained from features).

Naturally, if the correct encoding schemes that effectively capture the speech information to a particular task are employed, the feature values should change according to the changing speech characteristics over time. These changes can then in turn be described as variance within parameters. Speech recognition systems need good acoustic discriminative capabilities. In order to achieve this capability, successful representation of variance is mandatory.

In this chapter:

- The various effects that influence data variability are discussed in Section 3.2;
- Section 3.3 deals with the important aspects involved with regard to modelling strategies;
- In Section 3.4 a new model stability estimator is defined and evaluated.

3.2 FACTORS THAT INFLUENCE DATA VARIANCE

The complexities of a speech signal present a challenge when effective representation of acoustic information has to be achieved. Human cognition of speech includes many types of information. It

is well known that speech does not only convey linguistic information (the message) but also information about the speaker. For example, a human perceiver of speech can easily tell variations such as gender, age, social and regional origin, health and emotional state and, with a rather high reliability, the identity of a specific speaker [31].

The characterisation and the consequent effect that some of these specific variations have on speech recognition systems are a major research topic. The most prominently researched in literature are [31]:

1. Foreign and regional accents
2. Speaker physiology
3. Speaking style and spontaneous speech
4. Rate of speech
5. Child speech, and
6. Emotional state

These main topics serve to describe a number of unique sources of variation in speech. Note that acoustic channel conditions also play a major role. Using a high quality microphone produces much clearer sounds than recordings made from a telephone. The recording conditions such as sampling frequency, bandwidth and volume are important considerations. Protocols followed to digitally encode the audio data (compression) can alter the acoustic characteristics of speech data. Apart from this sound pollution such as background noise or non-speech signals also affect recording quality (consider the conditions of a public area versus studio audio conditions).

Although the sources (causes) of variations differ, some of the effects they introduce can be grouped together as different realisations of the same thing. Foreign accents, speaking style, rate of speech and child speech all cause pronunciation alterations. The actual alterations however, may differ for the various sources.

On the other hand, speaking style may also include long-term habits such as smoking or singing. Physical activity introduces effects like breathlessness and fatigue. There are also unrelated categories to these such as shouting, whispering, stuttering etc.

An interesting property that is unique to each individual speaker is the effect that the shape of the vocal apparatus has on the speech that a person can produce. This shape place limits on the range within which a particular speaker's voice may vary. Consequently much research has been done in this area and attempts have been made to model the vocal tract response mathematically [31].

With regard to the emotional state of a speaker, it has been noted that a mood change can have a significant impact on features used with speech recognition systems [31]. This is still an emerging field of study today and in most literature, attempts are made to identify variations such as "stressed" or "frustrated". These categories can then further be divided into: fast, angry or scared [31].

From the above it is clear that the sources that introduce variability to ASR training data are numerous and diverse. An equally important realisation is that these different types of variability convey unique types of information. To correctly utilise this information for practical purposes, accurate modelling strategies are crucial.

3.3 MODELLING DATA VARIANCE

In order to use speech information in a speech recognition system an effective representation of speech data are needed. This representation can be described as parameterisation (encoding) and modelling (an interplay between memorisation and generalisation on a given set of speech data). The following sections describe the typical considerations related to speech recognition parameters.

3.3.1 PARAMETERISATION

The extraction of useful information from the complex speech signal for computational purposes is an important process and can be referred to as encoding. Encoding for acoustic modelling purposes generally entails the parameterisation of very specific signal characteristics. Today, various encoding schemes exist that allow for the extraction of different types of information from the speech signal.

Information captured in this way is typically represented as a defined set of features. As the speech of a particular utterance progresses, the speech signal changes over time. It follows that these changes are then also reflected within the features of an applicable encoding scheme. Currently most systems utilise the time-varying spectra of the speech signal to model speech information. These types of parameterisation schemes have proven quite able to represent many different effects that can be distinguished within a speech signal, as described in Section 3.2.

Acoustic differences that influence speech recognition performance are ultimately dependent on the changes (variability) of the features for a chosen encoding scheme, due to the effects found within the actual speech data. This is the reason why Eigenspace-based methods such as Principal Component Analysis (PCA) have proven so useful to cluster data acoustically. As mentioned in Section 1.3.2 and 1.3.3.4, the technique enables characterisation of the acoustic properties of the data being analysed.

Characterisation basically implies the effective identification of all the sources of variation in order of acoustic significance. So far many of the major sources of variation have successfully been classified using general speech data parameterisation schemes (such as MFCCs). Effective identification of these variation sources within a data set is important, because it enables the optimisation of the representation (modelling strategy) of the speech parameters used within a speech recognition system.

3.3.2 NORMALISATION

Techniques also exist that serve to alter the variability of training parameters in a particular manner. Normalisation strategies primarily influence the speech signal such that important information is kept and unwanted distortions are cancelled out [32]. One such option, Cepstral Mean Normalisation (CMN), is an effective normalisation strategy to compensate for channel and speaker variations. Passing a speech signal through a transmission channel, effectively multiplies the speech spectrum with the channel transfer function. In the log-spectral domain, this multiplication is transformed into an addition. The addition can then easily be removed, by subtraction of the cepstral mean from all utterances. In practice the mean has to be estimated, because of the fact that one only has a limited amount of training data available [3].

3.3.3 MODEL TYPES

As mentioned in Section 1.3.1.1, the current acoustic modelling component as used in speech recognition systems is designed to effectively classify much of the variability normally found in speech data. Section 3.2 described the major sources of variation that affect speech recognition performance. Understanding how this variation is modelled relates directly to acoustic model design.

3.3.3.1 CURRENT MODELLING STRUCTURE

Current speech recognition architecture typically allows for the construction of acoustic models based on individual words or phones. This is a useful arrangement, because of the differences in variability that is found for different phones. These definitions have also been successfully extended to the realm of context-dependent triphones as described in Section 1.3.1.1.

Since the speech spectrum of a single phone is a time-varying signal, still further categorisation is possible. Within the framework of HTK, multi-state models can be constructed to allow for the modelling of this time variation. Every state of such a model then corresponds to only a part of the phone that is being modelled and training parameters can also be allocated dynamically to these states during the training process.

The information obtained from the speech parameters are ultimately captured using a probability density function. Current speech recognition architecture typically employs the use of Gaussian mixture probability densities. Depending on the amount of data and variability present within training parameters, increasing the number of mixtures allows for a more accurate representation of the training data.

3.3.3.2 MODEL COMPLEXITY

Depending on the particular parameters being used to represent training data, the variance (richness) of the data itself and the application designed for, representation (modelling) of the acoustics resident within the data may be done at different levels of complexity. To decide on the correct modelling

strategy for a particular task, it is important to give careful consideration to factors such as data sufficiency and also techniques that can serve to alter data variance mentioned in Section 3.3.2.

Given any modelling strategy, data sufficiency places limits on the effectiveness of acoustic models. The three main factors that determine data sufficiency are ultimately the amount of data available, the amount of variance within that data and the modelling strategy selected. Modelling strategies as explained above in 3.3.3.1, are typically designed to capture information of a known artefact within the speech signal (such as different phones). Ideally one seeks to isolate the variance introduced by the event of observing a specific artefact at the highest possible resolution. In practice however, complete isolation is never possible. If we consider the phone example, other distortions such as changing channel characteristics, different contexts (co-articulation effects) and even inconsistent pronunciations all affect the observed variance of a phone. To find the best modelling scheme for a particular task can therefore be a cumbersome process. Assumptions have to be made about the major sources of variation and then verified through experimentation. Typically as the amount of variance introduced by an artefact becomes less it is more difficult to isolate, because of noise introduced by the challenging artefacts. To increase the acoustic resolution, model complexity can only be increased up to a certain point, where we ultimately do not have enough data to justify the effects that we are trying to capture or the influence of other more dominant effects takes over.

For the remainder of this dissertation, the main issues related to model complexity that are addressed can be summarised as follows:

- The effect of modelling single-state or multi-state HMMs.
- Using single Gaussian or multiple Gaussian mixtures as density estimators.
- Modelling context-independent mono-phones versus context-dependent triphones.

3.4 GENERAL APPROACH

To investigate the relationship of variability and data requirements for a particular acoustic model, we define a suitable stability estimation measure. As shown from the definitions in Chapter 2, the similarity of Gaussian mixture models can be estimated using the Bhattacharyya estimator. It follows that if carefully selected training data are utilised to train different acoustic models, the similarity values generated in this way can be applied to characterise subtle differences for the given set of training data.

3.4.1 MODEL SATURATION

Any acoustic model that is based on a density estimator incorporates two types of functions: memorisation and generalisation. Furthermore, there exists an interplay between these two functions. Maximising memorisation may lead to overfitting of training data (so that the system cannot accurately

deal with similar unseen data), while too much generalisation (anything goes) may lower discriminative capabilities. Assuming we have an endless amount of training data, as far as a variability analysis of the data are concerned, memorisation can be increased by progressively grouping the data and then training separate acoustic models for all of the selected subsets of data. To do this effectively, one needs to have insight into the major sources of variability present in the data, so that optimal subsets of differences within the main data set can be modelled. This seems like a promising path to take, because we can endlessly divide data into more and more subsets, eventually characterising all of the variability.

In practice however, only limited amounts of data are available and memorisation quickly results in problems with regard model estimation. Too little training data effectively cause overfitting of the type of variability that one intended to model. Before any further conclusions with regard to similarity measurements between models are made, understanding the effect of data on model estimation are key.

The behaviour of a certain model can be investigated, by repeatedly constructing models for increasing amounts of data containing the specific type of variability to be represented. It is conceivable that one needs at least two acoustic models for this experiment, to allow for the estimation of a single similarity value between the two acoustic models. This directly translates to two mutually exclusive training data sets, used as training data for the acoustic models. An increasing amount of data can then be used to train both acoustic models and the similarity values estimated for models trained on the same amounts of data. If the amount of training data is sufficient, the similarity values should converge until eventually no observable change can be noted in the limit.

This is the point of model saturation. Ideally, for an acoustic model with a certain level of complexity, this is the place where we have enough training data to make effective comparisons between different types of variability being modelled. In practice however, with a limited amount of data, we can define a threshold where the rate of change has dropped below a certain magnitude and can be considered as insignificant to the other comparisons we want to make.

3.4.2 A THREE-DIMENSIONAL MEASURE OF ACOUSTIC STABILITY

Within a speaker independent speech recognition environment, there are two interesting types of variability that act on the level of individual acoustic model construction. These are phone and speaker variability, respectively. It is possible to define a stability estimation measure to operate in this three-dimensional space. This could be done by the definition of three axes: Number of phone observations, number of speaker occurrences and estimated similarity, respectively.

The variability introduced into a measurement of two mutually exclusive data sets with regard to speakers can, however, be significant, especially if the number of speakers per data set is limited. Ideally, with an unlimited number of speakers, the measurable difference between two sets of data due to different speaker variances would disappear. Because of the fact that the variability introduced by having a relatively small number of different speakers per data set includes and dominates the

amount of variability introduced by different utterances of the same phone, we need to measure and compensate for the instability of the measurements between different sets of speakers.

This is effectively done using a sufficiently large number of mutually exclusive training data sets with regard to speakers. If we do this, the number of similarity values that can be calculated between the acoustic models estimated on the different sets of data, becomes progressively larger with n distance measurements, where

$$n = C_2^x = \frac{x!}{2!(x-2)!} \quad (3.1)$$

between the x selected sets of training data.

Note that the general stability measure can be used to analyse different variables (not only speaker and phone variability). For example, the effect of different genders, age groups or channels can be analysed in analogous fashion.

To construct the stability measure the following steps are taken:

1. x mutually exclusive training data sub-sets (with regard to the main type of variability) are chosen.
2. For each training sub-set, data selection is then performed so that we incrementally add a number of predefined acoustic units (separate observations of the main type of variability).
3. One acoustic model is then trained on the selected data of every subset incrementally. Note that if complex Gaussian mixture models are employed, the required number of samples needs to be established first following the procedure described in Section 2.4.3 and selecting specific acoustic models that have been trained on the training data to do so.
4. For every increment the unique set of similarity values is then estimated between all of the acoustic models, as described in Chapter 2.
5. At every increment the mean and standard deviation of the number of similarity values generated is calculated.
6. For a large enough number of increments the mean (see previous step) reach an asymptote which is the point of model saturation.
7. Lastly to evaluate the stability of the measure itself, a number of permutations of the main type of variability among the different sub-sets of training data can be evaluated at model saturation.

For the remainder of this dissertation the stability measure is mainly used to investigate the effect of additional speakers and phone observations.

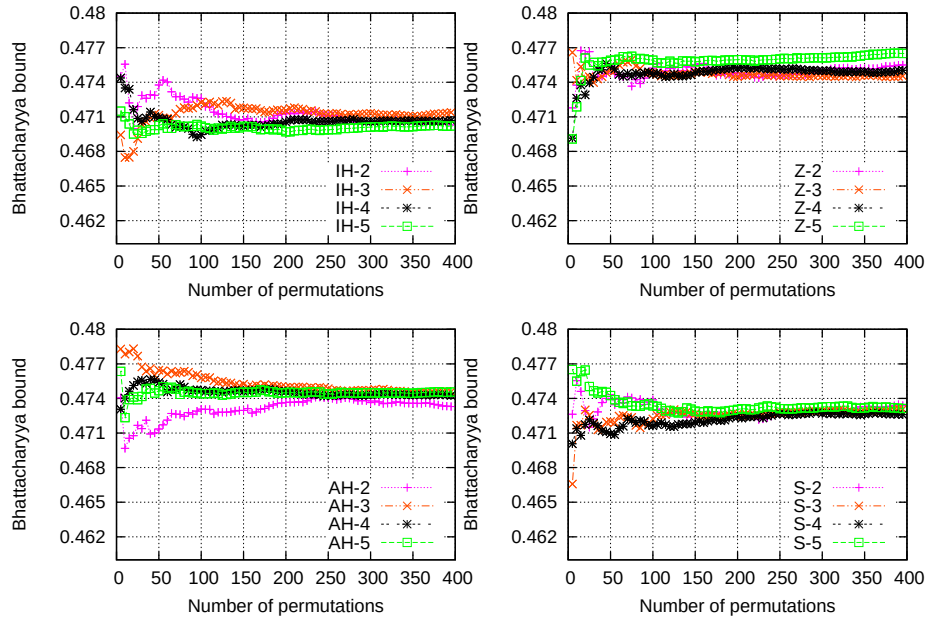


Figure 3.1: Mean of stability measurements for the phones /ah/, /ih/, /s/ and /z/

3.4.3 THE EFFECT OF SUBSET SELECTION

The drawback with application of the three-dimensional protocol to measure stability on limited numbers of speakers is the greater number of data sets being required. More data sets directly translate to more training data needed to investigate model stability on the same number of phone observations. In resource-scarce environments this can be a problem.

Also, depending on the specific experiment, different levels of accuracy may be required. If the goal is only to investigate the number of phone observations required for model saturation (speaker variability is not that important), then having fewer data sets may enable investigation of greater numbers of phone observations for this purpose. For experiments where the variability of different phones is being compared, a greater number of data sets would be required to more accurately represent the speaker variances.

Depending on the amount of variability introduced into the measurements due to the speaker variation, only a certain number of data sets is required for accurate estimation of the measurements taken. To determine the amount of data sets needed to compensate for the variability of our measurement strategy on an inadequate number of speakers we devise an experiment with increasing numbers of mutually exclusive data sets.

3.4.3.1 EXPERIMENTAL SETUP

To conduct the experiments the same data set as described in Section 4.2 is used. We follow the following protocol to investigate the effect of speaker variance using different numbers of data sets:

1. Our main data set consists of speech from 100 speakers. Each speaker is represented by 100

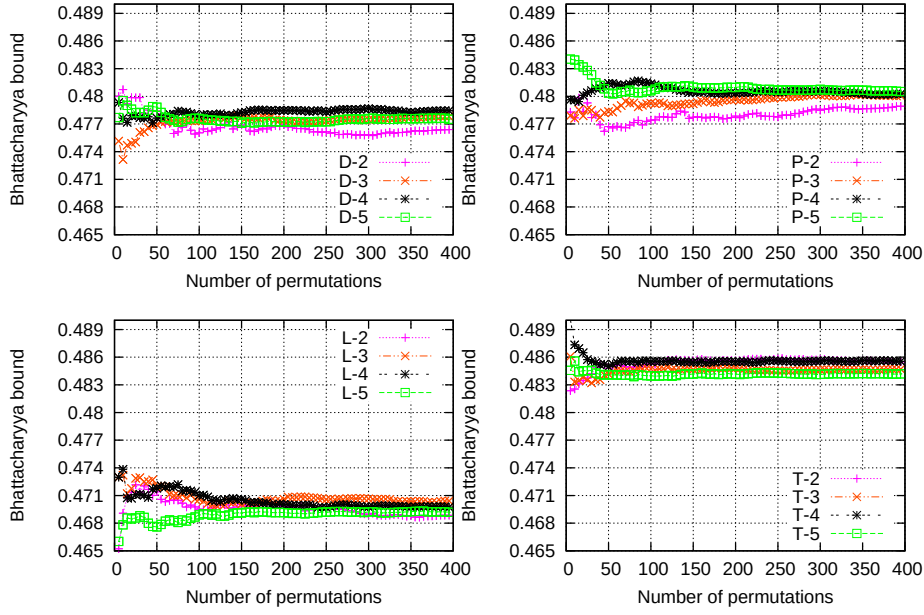


Figure 3.2: Mean of stability measurements for the phones /l/, /d/, /t/ and /p/

observations of a specific phone, randomly selected from the larger data set.

2. During each experiment a number of phone models (a single model for every data set) are constructed, with the number of phone models (denoted by i) ranging from 2 to 5.
3. Each of the i phone models is constructed by selecting 20 individual speakers randomly from the main data set. A phone model is also mutually exclusive with regard to speakers.
4. As the specific speaker selection may influence results, each experiment is repeated $J = 400$ times, each time using a different speaker permutation.
5. During each experiment, the Bhattacharyya bound between each model pair is calculated. If i models are used, N similarity values are calculated where $N = \{1, 3, 6, 10\}$ for $i = \{2, 3, 4, 5\}$. This bound is denoted by x_{ijn} with $2 \leq i \leq 5, 1 \leq j \leq 400, 1 \leq n \leq N$ respectively.

3.4.3.2 EXPERIMENT 1: MEAN OF STABILITY MEASUREMENT

Firstly, we investigate the mean of the bound generated for each number of data sets. This is done in order to verify that these means converge over various runs, when using different numbers of data sets. Calculation of these values is as follows:

1. For every experiment (i, j) , the mean

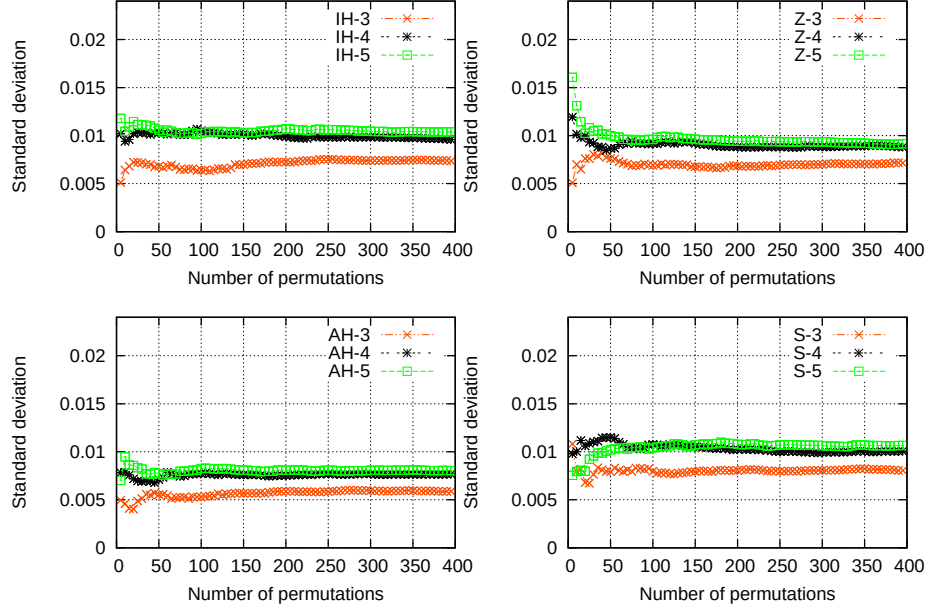


Figure 3.3: Cumulative standard deviation of the bound for the phones /ah/, /ih/, /s/ and /z/

$$\hat{\mu}_{ij} = \frac{1}{N} \sum_{n=1}^N x_{jin} \quad (3.2)$$

is estimated over the N Bhattacharyya bound values.

2. Next we calculate and plot the cumulative mean $\hat{\mu}_{iJ}$ over J permutations of the means obtained for a particular set of experiments

$$\hat{\mu}_{iJ} = \frac{1}{J} \sum_{j=1}^J \hat{\mu}_{ij} \quad (3.3)$$

where $1 \leq j \leq J$.

As noted in 3.4.3.1 each of the data sets is constructed by selecting 20 individual speakers and therefore we expect to observe a similar amount of variability on the bound. From Figures 3.1 and 3.2 it is indeed apparent that the cumulative means stabilise around very similar values (within a threshold of approximately 0.003) of the Bhattacharyya bound for the different numbers of data sets being investigated. It is, however, also apparent that the means generated for the different phones behave quite differently. As expected, the cumulative mean measurements of some phones take longer to stabilise and stabilise at lower bound values than others. Also, the cumulative mean measurements of experiments with a greater number of data sets stabilise more quickly for all of the investigated

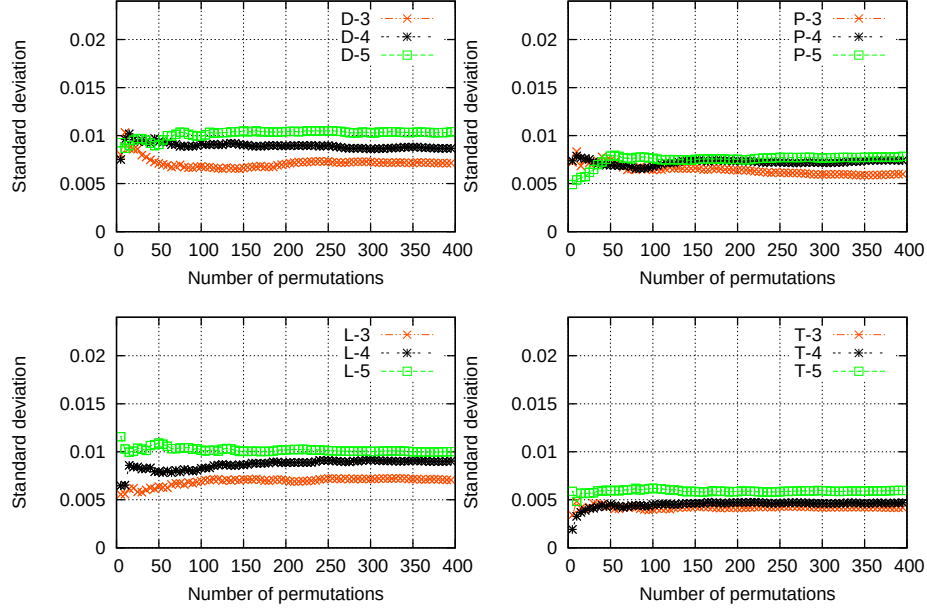


Figure 3.4: Cumulative standard deviation of the bound for the phones /l/, /d/, /t/ and /p/

phones.

For most phones it can be seen that the experiment with the least number of data sets (2) achieves slightly lower Bhattacharyya bound values than for experiments with more sets of data. This can be explained by the fact that more of the true variability is seen in general for a selection from a larger speaker set (2 data sets require selection of 40 speakers per experiment) compared to just the effective re-ordering (5 data sets requiring a selection of 100 speakers per experiment) of a limited set of speakers (only 100 speakers available in the larger data set). Note that this effect is small.

3.4.3.3 EXPERIMENT 2: CUMULATIVE STANDARD DEVIATION OF THE BOUND

Convergence can also be investigated through an analysis of the standard deviation of the bound, over a number of speaker permutations. The parameters required to do so are the following:

1. For every experiment (i, j) the standard deviation of the bound can be calculated as:

$$\hat{\sigma}_{ij} = \sqrt{\frac{1}{N} \sum_{n=1}^N (x_{ijn} - \hat{\mu}_{ij})^2} \quad (3.4)$$

where $i > 2$.

2. Next the cumulative standard deviation $\hat{\sigma}_{iJ}$ can be estimated over J experiments for every $i > 2$ data sets

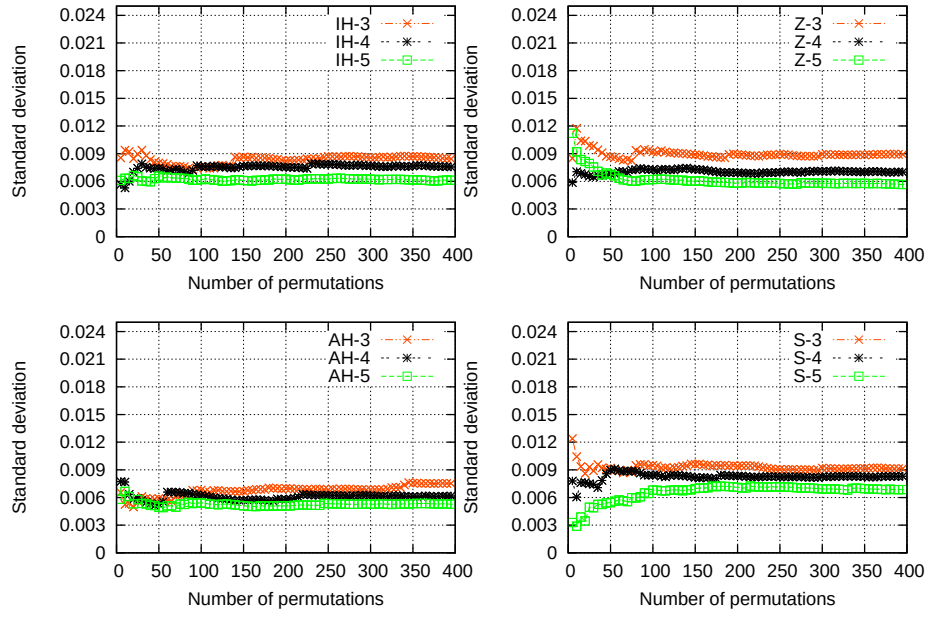


Figure 3.5: Cumulative standard deviation of stability measurements for the phones /ah/, /ih/, /s/ and /z/

$$\hat{\sigma}_{iJ} = \frac{1}{J} \sum_{j=1}^J \hat{\sigma}_{ij} \quad (3.5)$$

As expected from Equation 3.5 the standard deviations of the bound for experiments are better estimated cumulatively and convergence is achieved (Figures 3.3 and 3.4)¹. It can be seen that for the different numbers of data sets considered that the standard deviations for the greatest number of data sets (5) are still larger (for all phones) than the standard deviations of fewer data sets. Asymptotically, the standard deviations for different numbers of data sets are expected to converge, where J is large enough and all variability is seen.

3.4.3.4 EXPERIMENT 3: CUMULATIVE STANDARD DEVIATION OF THE MEAN OF THE BOUND

Finally the cumulative standard deviations of the mean of the bounds generated for different numbers of data sets and phones can be calculated as follows:

1. For every experiment (i, j) , the mean

¹Note that Equation 3.5 is a biased estimate. From the biased and un-biased estimates of $\hat{\sigma}_{iJ}$ in Table 3.1 it can be seen that the un-biased estimate converge even better.

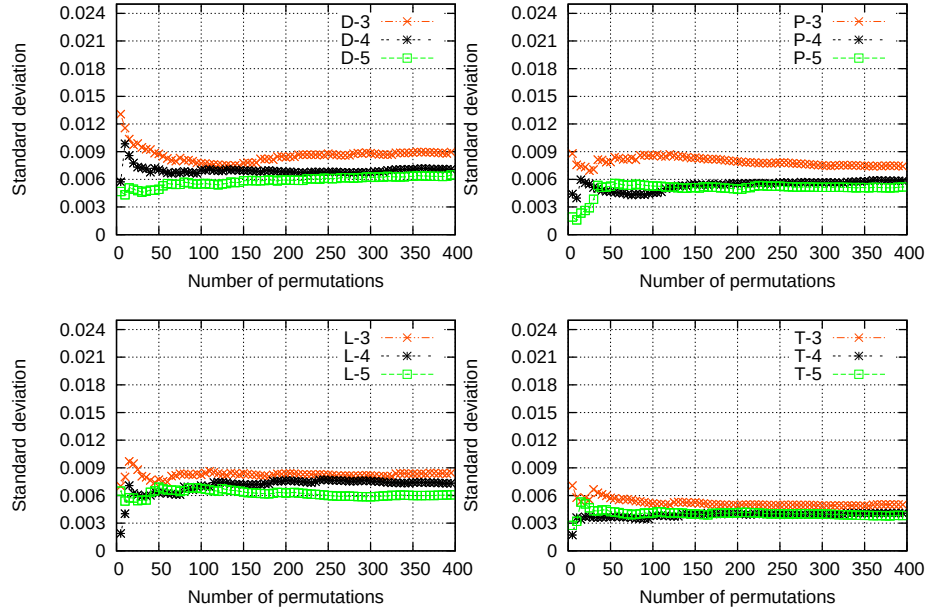


Figure 3.6: Cumulative standard deviation of stability measurements for the phones /l/, /d/, /t/ and /p/

Phone	Nr of data sets (i)	$\hat{\sigma}_{i,J}$	Un-biased $\hat{\sigma}_{i,J}$	Phone	Nr of data sets (i)	$\hat{\sigma}_{i,J}$	Un-biased $\hat{\sigma}_{i,J}$
ah	3	0.0059	0.0072	ih	3	0.0074	0.0090
	4	0.0077	0.0084		4	0.0096	0.0105
	5	0.0080	0.0085		5	0.0104	0.0109
s	3	0.0081	0.0099	z	3	0.0071	0.0087
	4	0.0101	0.0110		4	0.0088	0.0096
	5	0.0107	0.0112		5	0.0092	0.0097
l	3	0.0071	0.0086	d	3	0.0071	0.0087
	4	0.0090	0.0099		4	0.0087	0.0095
	5	0.0100	0.0106		5	0.0104	0.0110
t	3	0.0042	0.0051	p	3	0.0060	0.0073
	4	0.0047	0.0051		4	0.0074	0.0081
	5	0.0060	0.0063		5	0.0078	0.0082

Table 3.1: Difference between biased (Equation 3.5) and un-biased estimates of the cumulative standard deviations of the bound at 400 speaker permutations

$$\hat{\mu}_{ij} = \frac{1}{N} \sum_{n=1}^N x_{jin} \quad (3.6)$$

is estimated over the N Bhattacharyya bound values (as in Equation 3.2).

2. We calculate the cumulative standard deviation $\hat{\sigma}_{\mu i J}$ of the mean of the bounds over j permutations for a particular set of experiments

$$\hat{\sigma}_{\mu i J} = \sqrt{\frac{1}{J} \sum_{j=1}^J (\hat{\mu}_{ij} - \hat{\mu}_{iJ})^2} \quad (3.7)$$

where $1 \leq j \leq J$

It should be noted that the asymptotic value of $\hat{\sigma}_{\mu i J}$ is different for different i , as expected with more data sets providing more stable results (lower standard deviation of mean of bound).

Phone	Nr of data sets (i)	$\hat{\sigma}_{\mu i J}$	Phone	Nr of data sets (i)	$\hat{\sigma}_{\mu i J}$
ah	3	0.0075	ih	3	0.0084
	4	0.0062		4	0.0076
	5	0.0053		5	0.0061
s	3	0.0091	z	3	0.0089
	4	0.0083		4	0.0070
	5	0.0069		5	0.0057
l	3	0.0084	d	3	0.0090
	4	0.0073		4	0.0071
	5	0.0061		5	0.0065
t	3	0.0050	p	3	0.0074
	4	0.0041		4	0.0058
	5	0.0038		5	0.0052

Table 3.2: Standard deviations for phones of stability measurements at 400 speaker permutations

3.5 CONCLUSION

To effectively model various types of variability, given a specific set of ASR training data and different modelling choices, can be challenging. A way is needed to understand how a chosen model reacts to the variability of the investigated set of training data. Relevant to the experiments conducted in chapters to follow, a model stability measurement protocol is defined in order to analyse the effect of additional ASR training data on the learning behaviour of an acoustic model. To test the data requirements of an acoustic model when it is used to model the selected types of variability (speaker and phone variability respectively), model saturation is the key concept that is applied.

Trade-offs can be made with regard to measurement accuracies, depending on the goal of the particular experiments that are being performed. More subsets of data increase the consistency of stability measurements for different speaker permutations, with the true value of μ_i expected to be found in the range $[\hat{\mu}_j - \hat{\sigma}_{\mu i J}, \hat{\mu}_i + \hat{\sigma}_{\mu i J}]$. The measurement protocol can then be adjusted accordingly.

CHAPTER FOUR

DATA SUFFICIENCY ANALYSIS TECHNIQUES

4.1 INTRODUCTION

In Chapter 3 various aspects to consider with regard to the effective acoustic modelling of ASR training data are discussed. The effect that factors such as the amount of training data, type of investigated variation and model complexity have on model estimation are important to understand when any variability measurements are conducted. In this chapter we specifically analyse how different levels and types of variability in training data impact the data sufficiency requirements of acoustic models in terms of the previously defined model stability measurement protocol. This section contains a number of experiments that we use to develop the techniques required for data sufficiency analysis. We also obtain some preliminary results (In Chapter 5 more tools will be applied to analyse a full corpus).

The chapter is structured as follows:

- Training data and experimental parameters are introduced in Section 4.2;
- In Sections 4.3 and 4.4, an initial analysis employing single Gaussian models is conducted;
- The effect of additional model complexity is investigated (Section 4.5);
- Section 4.6 deals with context dependence;
- Bootstrapping is applied as a model refinement (Section 4.7); and
- Lastly in Section 4.8 the effect of Cepstral Mean Normalization is also evaluated.

4.2 DATA AND EXPERIMENTAL SETUP

To conduct the analysis, the November 1992 ARPA Continuous Speech Recognition Wall Street Journal Corpus is used as training data [33]. The dataset consists of 102 speakers recorded over the same acoustic channel. This enables experimentation with up to 20 hours of data. Pronunciations for the training data are obtained from the CMU pronunciation dictionary [34]. We train a complete speech recognition system and use forced alignment to label all of the utterances. This has two very specific purposes namely:

1. Pronunciation variant selection (identify correct pronunciations based upon acoustic evidence); and
2. Obtaining phone boundaries (used in Section 4.7).

Speech recognition is performed using standard HMMs with three emitting states, tied across models, each containing up to 12 GMMs. The GMMs are trained on 39-feature MFCC-based speech vectors (13 MFCCs, deltas and double-deltas with cepstral mean subtraction).

To estimate model estimation stability, the protocol as previously described in Section 3.4.2 is used. The data set is separated into 5 disjoint subsets with regard to speakers. With this protocol the mean of 10 similarity values are thus computed for every stability measurement. Keeping the speaker selection constant with regard to each of the subsets, the variability of different phones can then be compared, as the size of the training set is slowly increased.

4.3 INITIAL ANALYSIS

An initial analysis is performed to investigate the stability measurement space at various parameters. Three-dimensional speaker-and-utterance plots are developed for the acoustic models of a number of different phone types. Each plot indicates the value of the stability measure (the Bhattacharyya bound), in accordance with the applied protocol, as a function of both the number of speakers in the training corpus and the number of phone observations per speaker. An example of such a plot for the phone /ah/ is shown in Figure 4.1. As the stability value shown is an estimate of the Bhattacharyya bound, this value should approach 0.5 once a model is fully trained on an optimal set of data.

From this initial analysis the following are observed:

1. A specific number of speakers and phone observations result in significantly different results for the different phones.
2. While phones from the different broad phone categories (such as vowels, plosives or fricatives) exhibit varying learning behaviour, phones within a specific phone category follow remarkably similar trends.
3. Similar trends are also observed when using either simple or the more complex acoustic models.

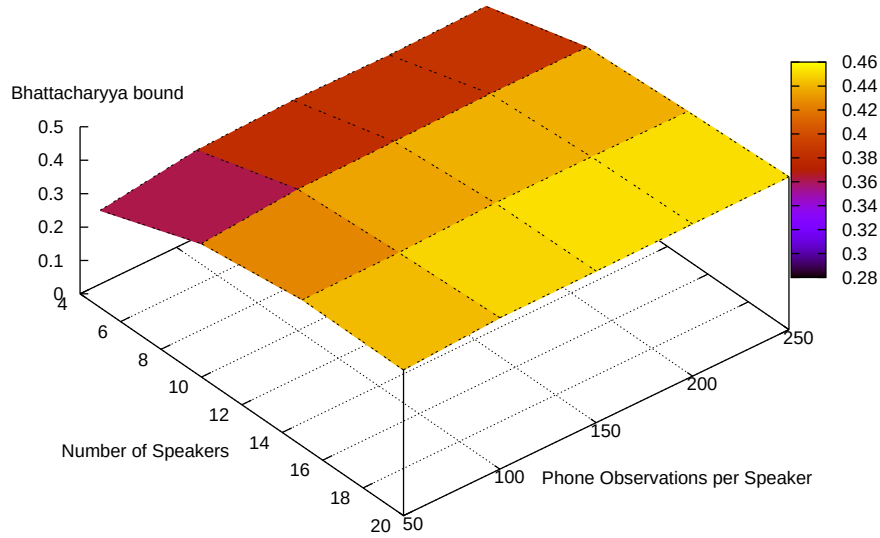


Figure 4.1: *Speaker-and-utterance three-dimensional plot for the phone /ah/*

For each broad phone category, a number of representative phones are selected to illustrate the trends observed. Specifically, the following phones are selected after verifying that these phones are indeed representative of the larger groups:

1. /ah/ and /ih/ (vowels)
2. /n/ (nasals)
3. /l/ and /r/ (liquids)
4. /d/ (voiced plosives)
5. /t/ and /p/ (unvoiced plosives)
6. /z/ (voiced fricatives)
7. /s/ (unvoiced fricatives)

The investigations to follow utilise these selected phones for comparative purposes.

4.4 SINGLE GAUSSIAN MODEL

In this section the aim is to understand whether a saturation point is reached, as described in Section 3.4.1, after a specific amount of data has been added for single Gaussian models. To do so, cross-sections of the initial 3-D plots as shown in Figure 4.1 are taken.

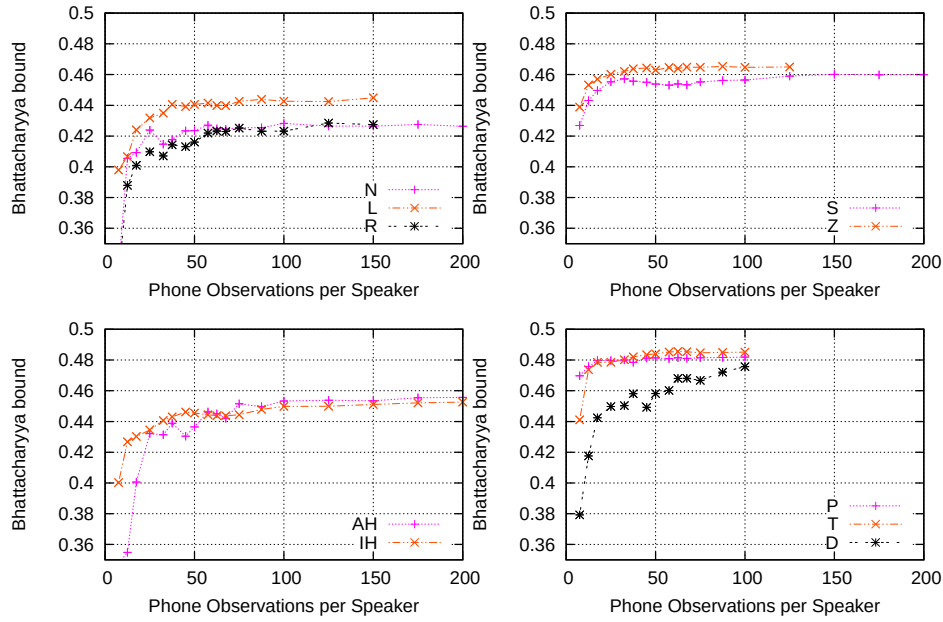


Figure 4.2: Effect of number of phone utterances per speaker on mean of Bhattacharyya bound for different phone groups using data from 20 speakers

4.4.1 NUMBER OF PHONE OBSERVATIONS REQUIRED PER SPEAKER

The saturation point, after which additional examples of phones by a specific speaker no longer improve the accuracy of the speaker independent acoustic model for that phone can be deduced from Figure 4.2. The plots show cross-sections of the initial 3-D analysis at a specific number of speakers (20). Evaluating increased numbers of phone observations per speaker leads to saturation.

As clearly demonstrated in Figure 4.2, the stability measurements of all phones reach an asymptote quite quickly and for 20 speakers, this asymptote does not yet approach the ideal 0.5 level for most of the phone types. When this experiment is repeated with 50 speakers (see Figure 4.3), even fewer phone observations per speaker are required to reach the asymptote, and all the asymptotes are also nearer to the ideal level of 0.5. Interestingly though, the total numbers of phone observations necessary for the model of a phone to reach the asymptote are comparable for the 20 and 50 speaker cases. Note that in order to be able to evaluate the effect at 50 speakers, only 2 models could be trained and 1 similarity value estimated, in comparison to the 5 models and 10 similarity values possible at the 20-speaker level. From Section 3.4 it is clear that more data sets of 50 speakers each would improve measurement accuracy. It is expected that the true 50-speaker stability measure would be slightly lower (order of 0.001).

For the different phone types we observe that vowels are the slowest to reach the saturation point (at approximately 100 phone observations per speaker in the 20-speaker case) while unvoiced plosives and fricatives are the quickest to stabilise, reaching this point at only 35 phone observations for /s/, 45 phone observations for /z/ and 25 phone observations for /p/ or /t/. There is a clear difference

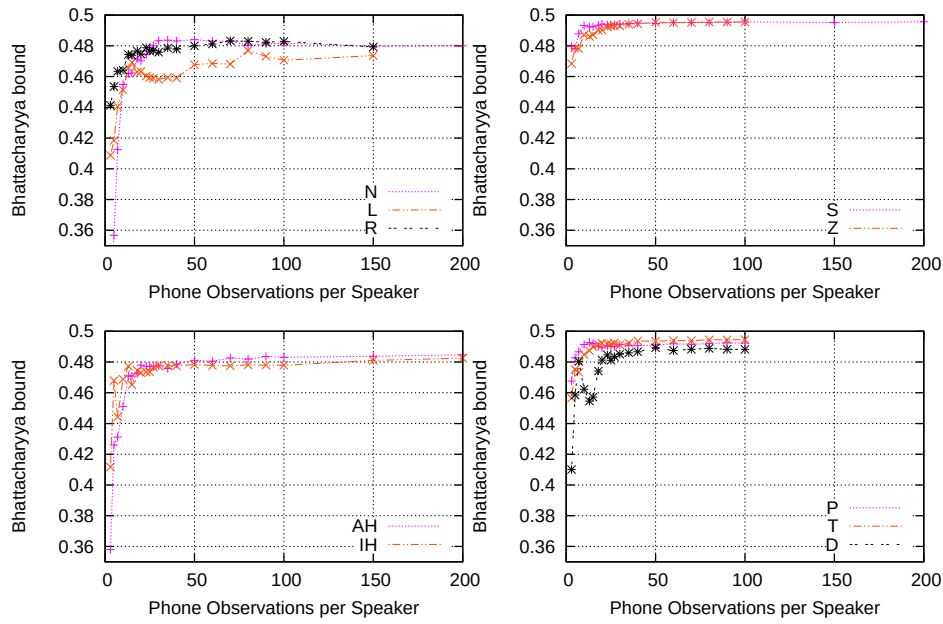


Figure 4.3: Effect of number of phone utterances per speaker on mean of Bhattacharyya bound for different phone groups using data from 50 speakers

between the unvoiced and voiced versions of the plosives, with voiced versions taking significantly longer to stabilise (compare /d/ at 85 phone observations with /t/ at 25 phone observations). For most phones, those that saturate more quickly achieve a higher bound (closer to the ideal 0.5). However for some phones, such as /d/, a large number of phone observations are required per speaker, but the higher bound indicates that fewer speakers are required to obtain an accurate estimate. Similarly, the fricatives (/s/ and /z/) reach their asymptote very quickly, but this asymptote is fairly low, indicating low intra-speaker but high inter-speaker variability for this phone.

4.4.2 NUMBER OF SPEAKERS REQUIRED PER PHONE

In this section the aim is to understand the effect of adding additional speakers to a training corpus during acoustic model construction. A number of phone observations per speaker (100) is selected, where the asymptote has already been reached for all of the investigated phones if 20 training speakers are employed. For each new speaker an additional 100 phone observations are then systematically added. The results of this experiment are shown in Figure 4.4.

This time, the asymptote is not reached, and it is clear that additional speakers would improve the modelling accuracy for all phone types. On theoretical grounds it is expected that the stability measurements should in all cases approach 0.5, and this expectation is supported by the observed trends. Again we observe that the unvoiced plosives and fricatives quickly achieve high values for the bound (close to the ideal 0.5). Low inter-speaker variability for the phone /d/ is also confirmed with high bound values. The high inter-speaker variability of the fricative phones are apparent in the unstable behaviour they exhibit (varying between 0.4 to 0.45 up to 20 speakers). Interestingly the

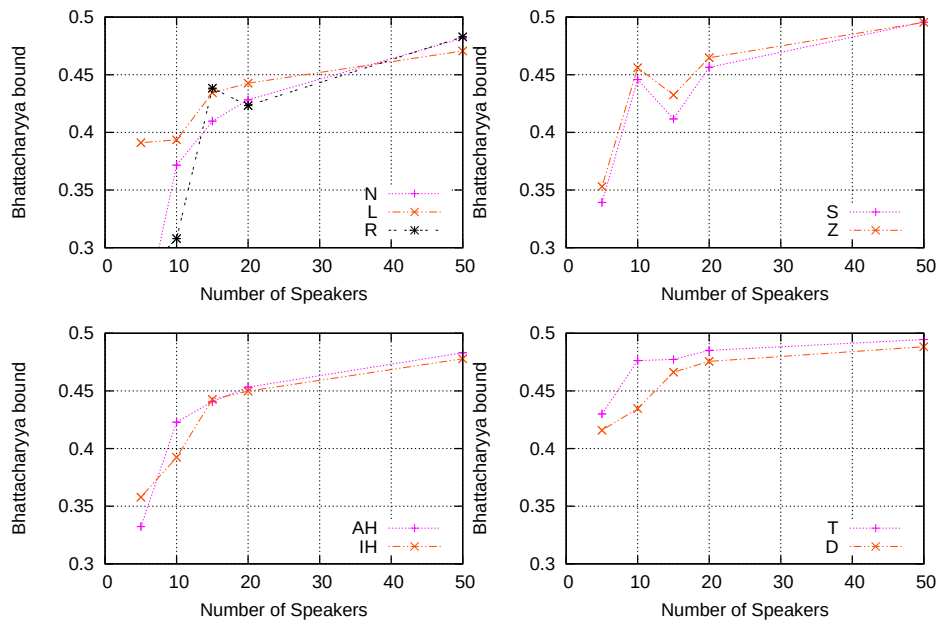


Figure 4.4: *Effect of number of speakers on mean of Bhattacharyya bound for different phone groups using 100 utterances per speaker*

vowels are not the slowest to reach large bound values as the speakers are increased: the phones /n/ (nasals) and /r/ (liquids) converge more slowly, signalling a higher inter-speaker variability for this group.

These results confirm the results obtained in Section 4.4.1 and comparative behaviour for the different phone types is summarised in Table 4.1.

Phone type	Inter-speaker variability	Intra-speaker variability
Unvoiced plosives	low	low
Voiced plosives	low	high
Unvoiced fricatives	medium	low
Voiced fricatives	medium	low
Vowels	medium	high
Nasals	high	medium
Liquids	high	medium

Table 4.1: *Comparative inter- and intra-speaker variability for different phone types*

4.5 GAUSSIAN MIXTURE MODEL

Evaluation of Gaussian mixture models on the same set of data is achieved using the derived similarity measure as defined in Section 2.4. Experiments are conducted at a model complexity of 6 mixture components. (The choice to utilise a 6-mixture GMM was made to balance high speech recognition

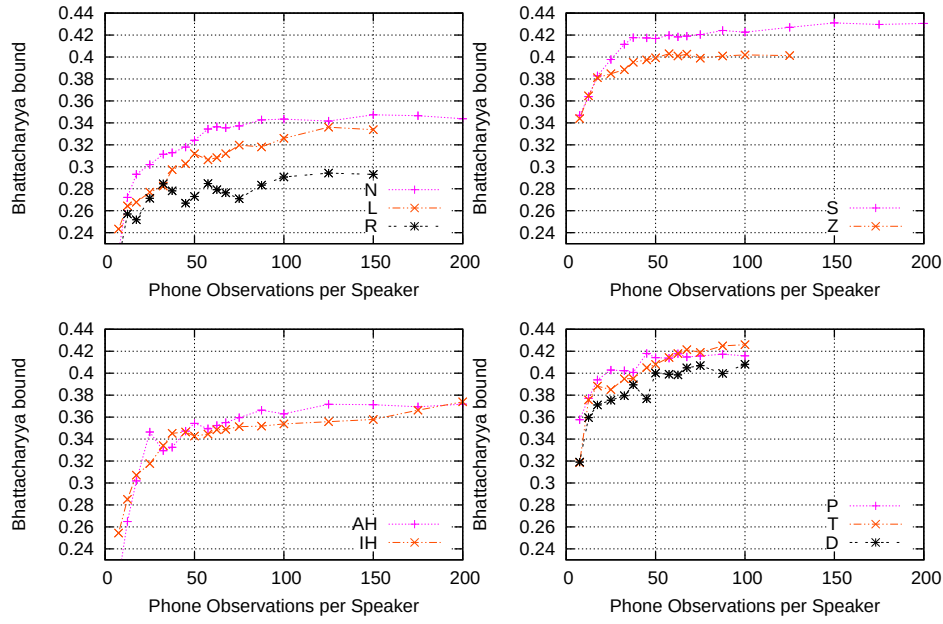


Figure 4.5: Effect of number of phone utterances per speaker on mean of Bhattacharyya bound for different phone groups using data from 20 speakers and 6 mixtures

accuracy for our data set with computational requirements during bound estimation.)

4.5.1 NUMBER OF SAMPLES

To implement the derived similarity measurements a minimum number of samples are required for accurate estimation. We estimate the number of samples required for a selected set of model pairs (see Section 2.4.3) and find that 20,000 samples are sufficient for our purpose. Table 4.2 summarises the number of samples required to keep standard deviations below a threshold of 0.0100 for the various model comparisons. With 20,000 samples, the standard deviation among different estimations of bounds between GMMs containing up to 6 mixtures are below 0.0020 for very similar phones and below 0.0061 for quite dissimilar phones. (Model pairs with similarity values of approximately 0.5 and 0.1 respectively). Also, with 20,000 samples and a single Gaussian model, these estimates are within 0.0002 and 0.0020 from the corresponding analytically calculated values.

4.5.2 EFFECT OF ADDITIONAL MODEL COMPLEXITY

To understand the effect of additional phone observations on the more complex 6-mixture Gaussian models, a similar experiment to Section 4.4.1 is performed (Figures 4.5 and 4.6). It is apparent that the stability measurements of all phone categories are significantly lower and consequently translates to a higher variability being seen for all phones with this experiment. However, similar trends than previously observed still hold across phone groups.

Evaluating increased numbers of phone observations still leads to sufficient model saturation for

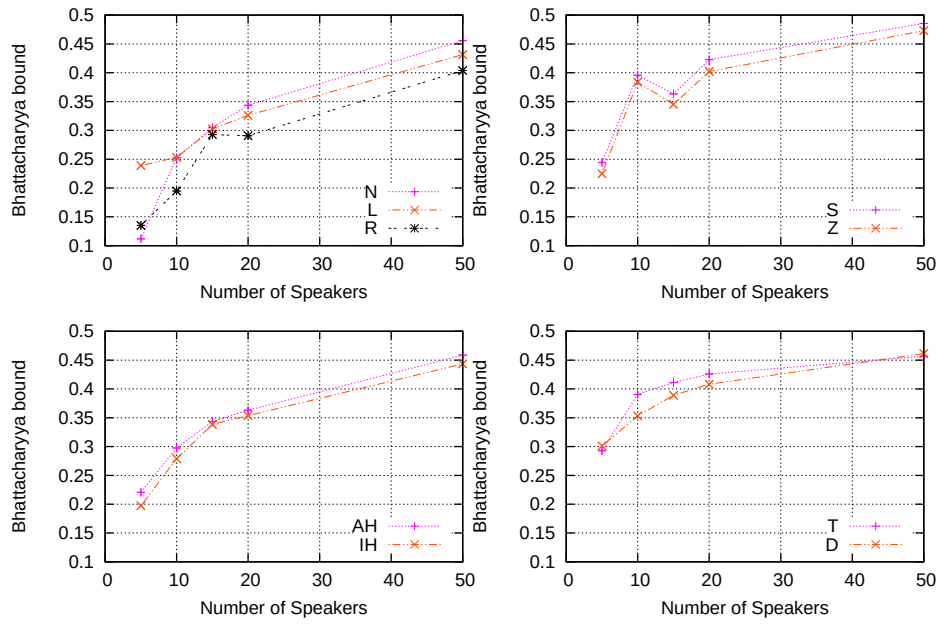


Figure 4.6: *Effect of number of speakers on mean of Bhattacharyya bound for different phone groups using 100 utterances per speaker and 6 mixtures*

Num mixtures	Samples required	ϵ	σ_s
1	5,000	0.1	0.0100
		0.3	0.0073
		0.5	0.0031
2	10,000	0.1	0.0062
		0.3	0.0060
		0.5	0.0018
4	20,000	0.1	0.0045
		0.3	0.0020
		0.5	0.0017
6	20,000	0.1	0.0061
		0.3	0.0045
		0.5	0.0020

Table 4.2: *Number of samples required for accurate estimation of similarity values*

most phones. For the different phone types it is clear that the nasals /n/ and liquids /r/ are the slowest to converge and even at 150 phone observations it still is not clear whether an asymptote is reached or not. The vowels also take much longer to stabilise (after approximately 130 phone observations) and exhibit similar variable behaviour than for the simpler models, confirming the previously observed high inter- and intra-speaker variability for this group. Interestingly, for the unvoiced plosives /t/ and /p/ a much higher intra-speaker variability similar to that of the voiced plosive /d/ is now being observed (This may be due to context-dependent co-articulation effects: see Section 4.6). The fricatives still reach their asymptote fairly quickly with /s/ and /z/ stabilising at 45 and 50 phone observations

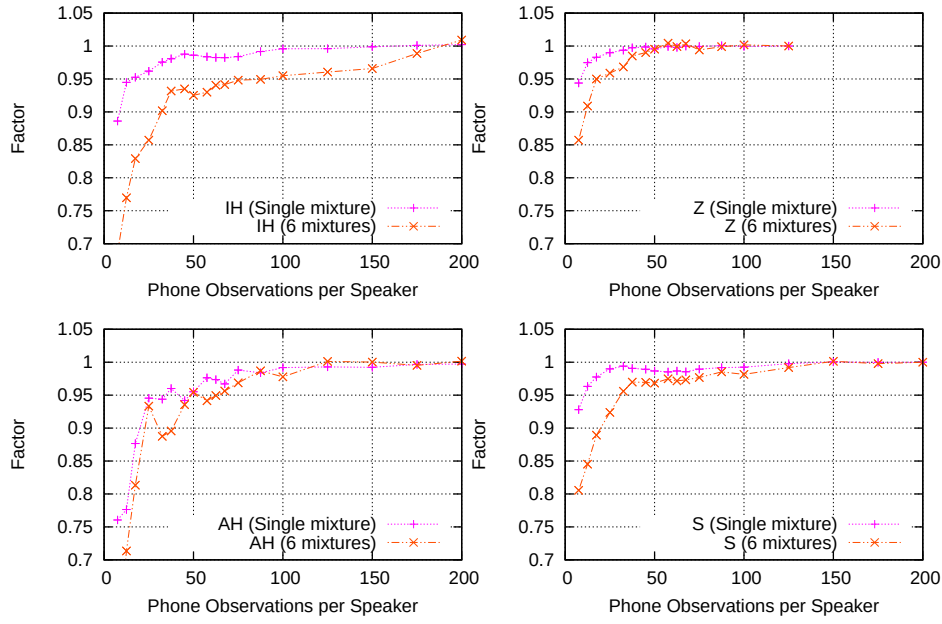


Figure 4.7: Comparing the effect of model complexity on the relative distance to asymptote for the phones /ah/, /ih/, /z/ and /s/

respectively, establishing a low intra-speaker variability for this group.

For the speaker section plots it is clear that the exact same trends as observed in Section 4.4.2 still hold for all of the investigated phones. As with the above experiment however, the extra variability for the more complex models translates to lower stability measurement values and for all phones it is clear that additional speakers will still improve the accuracy of the models. Again we observe that the plosives and fricatives quickly achieve high bound values. Low inter-speaker variability is confirmed for /t/ and /d/. For the fricatives /s/ and /z/ the high inter-speaker variability is also still apparent in the unstable behaviour they exhibit and as expected we observe a much higher inter-speaker variability for the nasals /n/ and liquids (/r/ and /l/) than for the vowels (/ah/ and /ih/) respectively.

The numerical values of the Bhattacharyya bound for different model types cannot be compared directly, since factors such as the existence of local minima during training increase the apparent variability of more complex models. We therefore compare such models by studying the observed bound values as a fraction of the observed asymptotic values. It can be seen that the more complex model requires additional samples before the asymptotes are reached. This is illustrated in Figures 4.7 and 4.8 where this fraction is shown, as the number of phone occurrences per speaker is increased. In these figures, data from 20 speakers is shown for both the simple single Gaussian model as well as the more complex 6-mixture GMM.

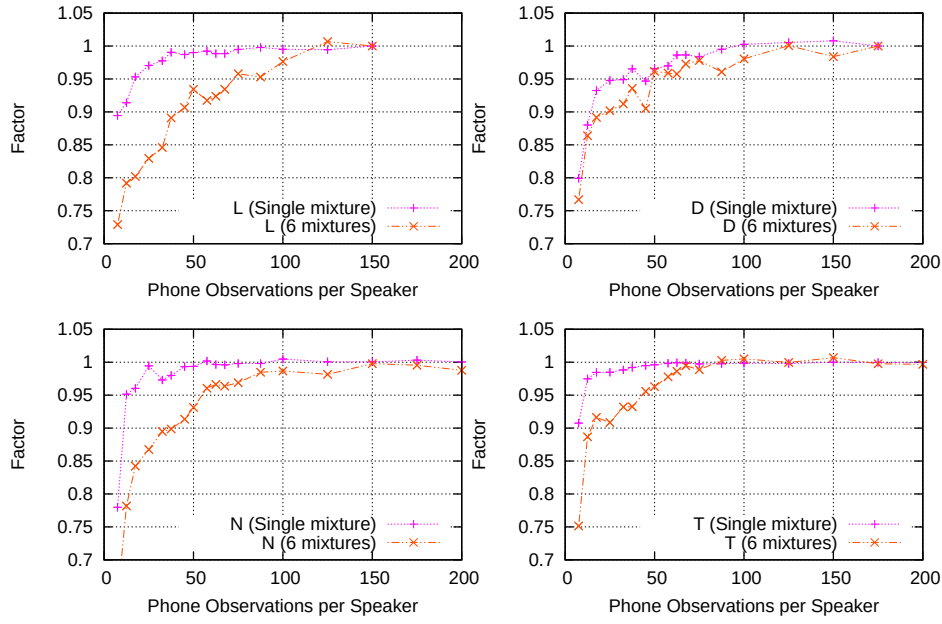


Figure 4.8: Comparing the effect of model complexity on the relative distance to asymptote for the phones /n/, /l/, /d/ and /t/

4.6 EFFECT OF CONTEXT-DEPENDENCE

To understand the effect of context-dependence on the stability measurements taken, 4 phones (/ah/, /n/, /s/ and /p/) are selected to cover the different phone groups as investigated in the previous sections. Only single Gaussian models are considered. To analyse the acoustics of the contexts, we train a complete set of tied triphone models for every phone and use a standard decision tree as described in Section 1.3.1.1, with a high acoustic threshold during tying. Specifically the threshold is set at such a level so that only the 5 acoustically most different triphone clusters are obtained for the nasal /n/ (expected to be highly variable from experiments in previous sections). A similar threshold value is then used for the three remaining phones to allow for comparability.

The triphone clusters obtained are analysed on a parent phone basis. Similar to the experiments of previous sections, the training data for each of the four parent phones are divided according to the applied protocol (5 disjoint subsets with regard to speakers). The same subsets of data are then also further utilised to analyse all of the triphone clusters as well.

Again Figures 4.9, 4.10, 4.11 and 4.12 represent a cross-section of the initial 3-D plot in Figure 4.1 at a specific number of speakers (20). The triphone clusters formed (up to a maximum of 5 covering the most data) as well as the corresponding parent phone is shown. Furthermore, the different contexts that represent the triphone clusters of the phones for the particular acoustic threshold can be found in Tables A.1, A.2, A.3 and A.4. Due to the large number of contexts, however, only a selected few are shown to convey the general idea.

It is immediately apparent that much less data are required for the triphone clusters to reach

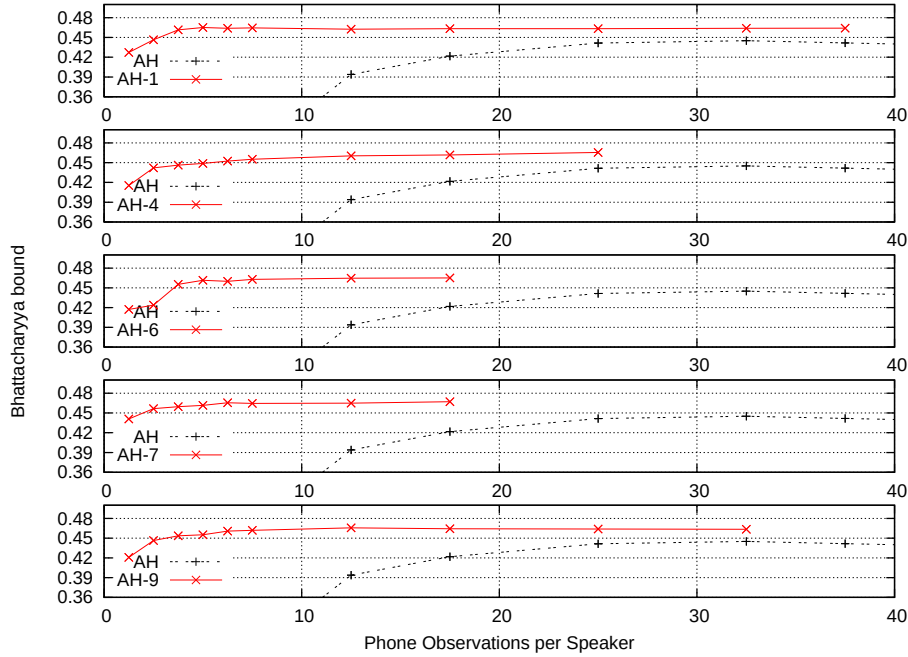


Figure 4.9: Effect of number of phone utterances per speaker on mean of Bhattacharyya bound for different contexts of the phone /ah/ using data from 20 speakers

saturation than the corresponding parent phone. For the different phone types we observe that all the triphone clusters for the nasal /n/ are slowest to reach the saturation point (at approximately 40 phone observations per speaker in the 20-speaker case), while all of the other phones now stabilise at approximately 15 phone observations per speaker.

The vowel /ah/ and nasal /n/ also consist of the most substantially different clusters (9 and 5 respectively) followed by the fricative /s/ (4) and interestingly the plosive /p/ which turned out to have two significantly different clusters. Although the vowel /ah/ now consist of more (9 triphone clusters) acoustically very different clusters than the plosive /p/ (2 triphone clusters), the clusters of both phones now stabilise with the same amount of data.

For all four the parent phones the bound values of the triphone clusters at saturation are slightly higher than that of the corresponding parent phone at the same number of phone observations, indicating reduced intra-phone variability for the triphone clusters. Also, for the parent phones represented in Figure A.1, the triphone clusters with the most data seem to dominate, with the stability measurement values of the parent phones slowly drifting until they practically reach the same asymptote as the best represented triphone cluster.

4.7 BOOTSTRAPPING

In Section 4.3 a speech recognition system is used in forced alignment to label all of the utterances and consequently obtain suitable pronunciations that fit the acoustic evidence presented in the data. Such a system can also be used to estimate better phone boundaries. This is due to the fact that such

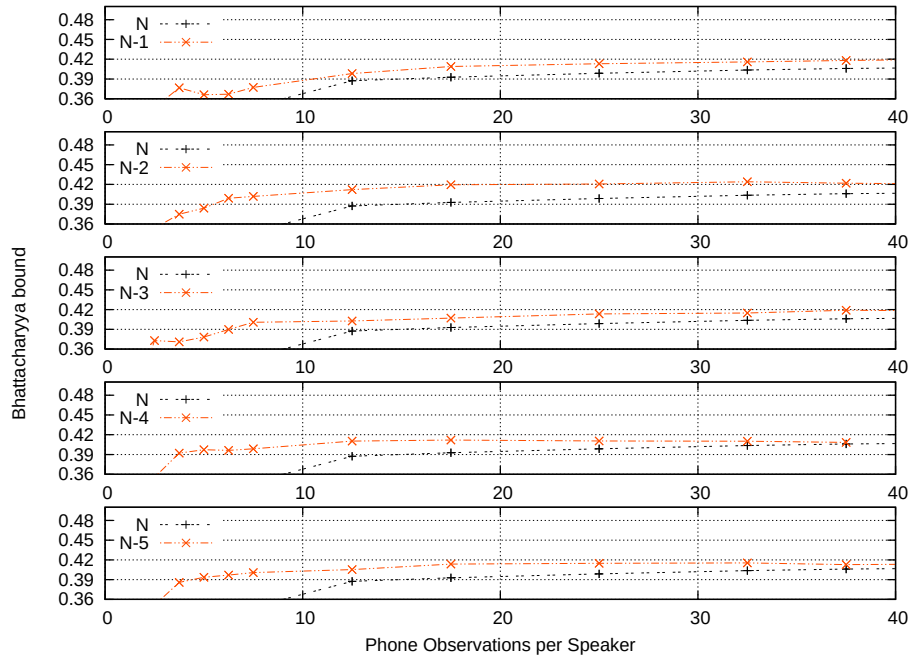


Figure 4.10: *Effect of number of phone utterances per speaker on mean of Bhattacharyya bound for different contexts of the phone /n/ using data from 20 speakers*

boundaries are generated more accurately using 3 state HMMs than single Gaussian models. To take advantage of this, a bootstrapping approach is followed, where the parameters of the single Gaussian models are estimated given the optimal phone boundaries of another optimised speech recognition system.

The previous results presented in Figures 4.2 and 4.4 are updated with Figures 4.13 and 4.14, where the effect of the better phone boundary estimation is apparent. As expected, the Bhattacharyya bound values are now consistently higher for nearly all of the phones (less intra-phone and inter-speaker variability). It is clear that the vowels and nasals are the most affected, while the unvoiced plosives show little or no improvement. Furthermore the unstable learning behaviour of the phones /ah/, /n/ and /r/ are mostly corrected for, although it still takes about 100 phones per speaker for the asymptotes to be reached.

The high intra-phone variability of the voiced plosive /d/ is still just as apparent although the stability measurements are higher. From the speaker section plot taken at a 100 phones per speaker, we also see the consistent improvement across the whole spectrum of phone types. All previous trends are preserved, but now with higher stability measurement values.

4.8 CEPSTRAL MEAN NORMALISATION

As discussed in Section 3.3.2, Cepstral Mean Normalisation (CMN) is a simple, but effective technique that compensates for long-term spectral effects. If the spectral mean is estimated for every speaker utterance and then subtracted in the log cepstral domain, normalisation of the features is

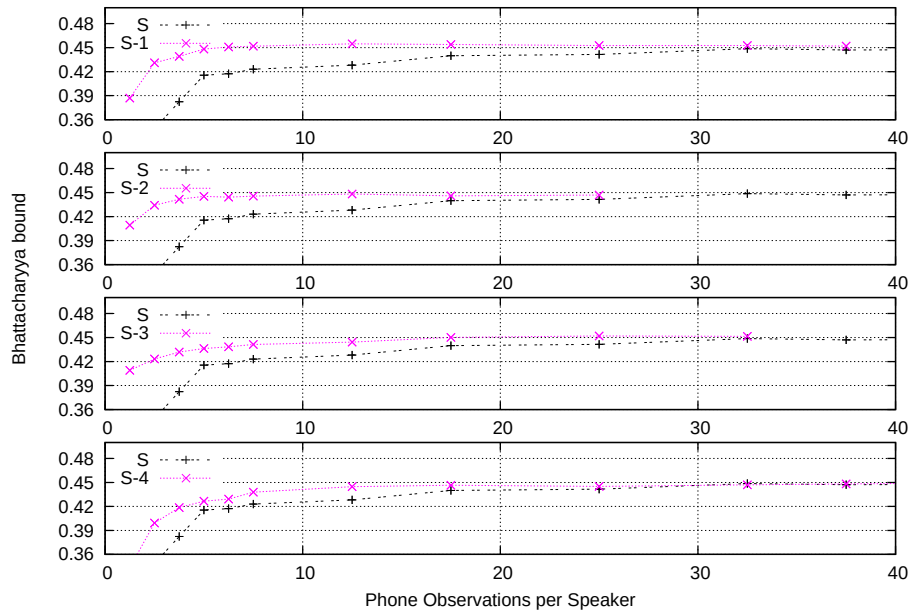


Figure 4.11: *Effect of number of phone utterances per speaker on mean of Bhattacharyya bound for different contexts of the phone /s/ using data from 20 speakers*

obtained.

The results obtained with bootstrapping in Figures 4.13 and 4.14 are updated to also include the added effect of CMN in Figures 4.15 and 4.16. For the 20 speaker case it can immediately be seen that the stability measurements for the vowels and nasals show great improvement (much higher bounds) and the plosives a small, but significant improvement. With the fricatives, CMN results reveal a degradation. From the speaker section plot taken at 100 phones per speaker, the difference between the voiced and unvoiced fricatives are apparent. The unvoiced fricative /s/ shows only degraded performance, while the voiced fricative /z/ (although more stable) seem to be a balance of counter effects and consequently does not show a significant difference.

4.9 CONCLUSION

In this chapter we developed a number of tools and techniques that can be used to analyse the data sufficiency requirements of different types of acoustic models. Specifically we found that:

- obtaining accurate phone boundaries is important (Section 4.7)
- CMN has a significant effect on data sufficiency, with less data required if CMN is applied (Section 4.8)
- single Gaussian models provide a good indication of the trends that can be expected if more complex modelling techniques are used including context-dependence (Section 4.6) and multiple mixtures (Section 4.5.2)

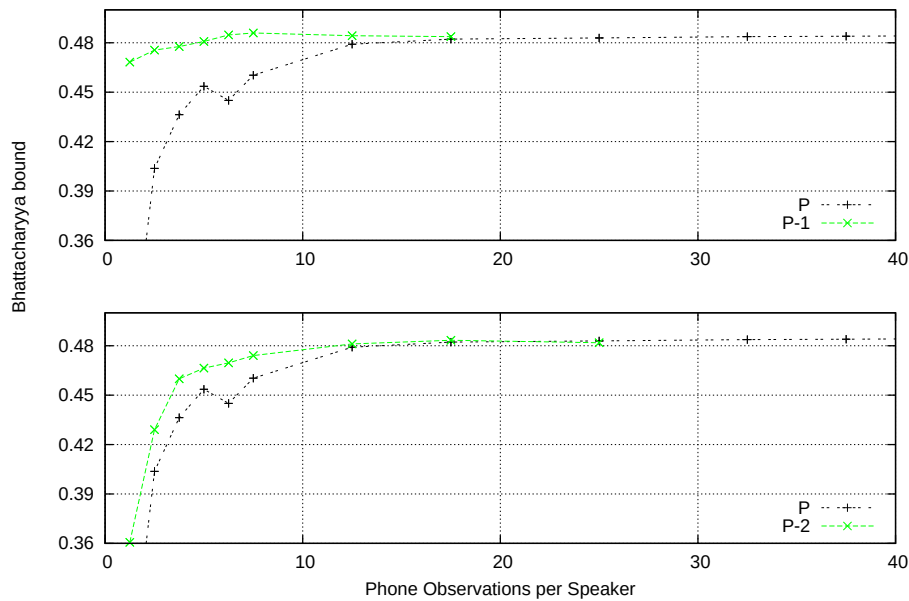


Figure 4.12: *Effect of number of phone utterances per speaker on mean of Bhattacharyya bound for different contexts of the phone /p/ using data from 20 speakers*

- we observe that different phone categories demonstrate different data sufficiency trends, which are related to the amount and type of variability (intra- or inter-speaker) for the specific phones.

These techniques are utilized in the next chapter to perform a data sufficiency analysis on a new speech recognition corpus: The Lwazi ASR corpus.

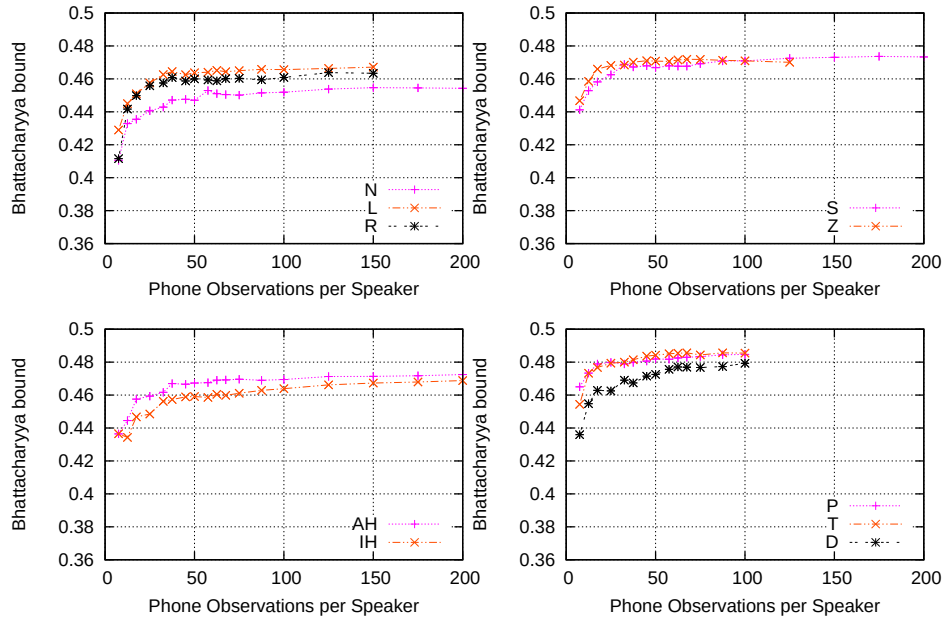


Figure 4.13: Effect of number of phone utterances per speaker on mean of Bhattacharyya bound for different phone groups using data from 20 speakers using bootstrapping approach

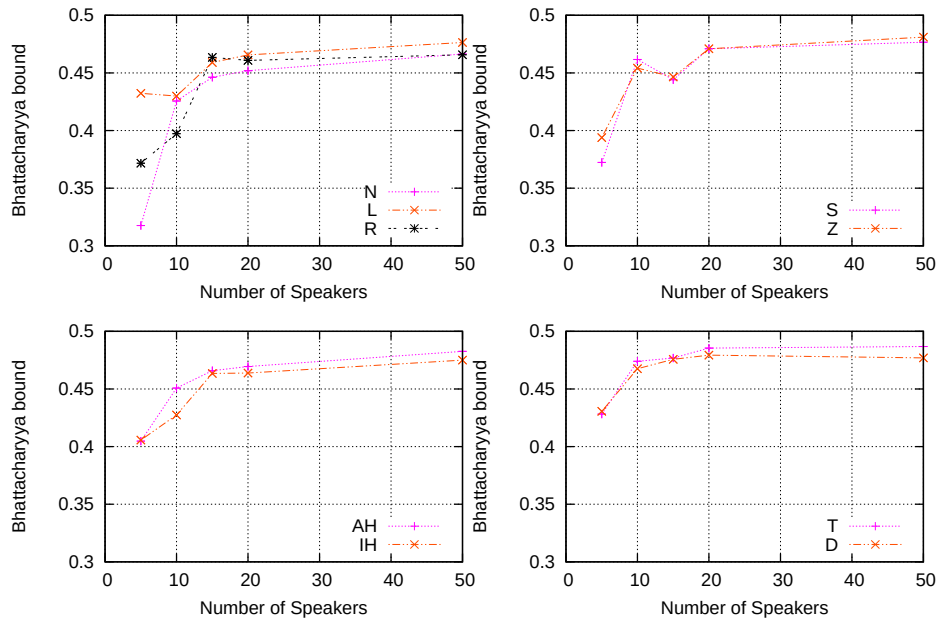


Figure 4.14: Effect of number of speakers on mean of Bhattacharyya bound for different phone groups using 100 utterances per speaker using bootstrapping approach

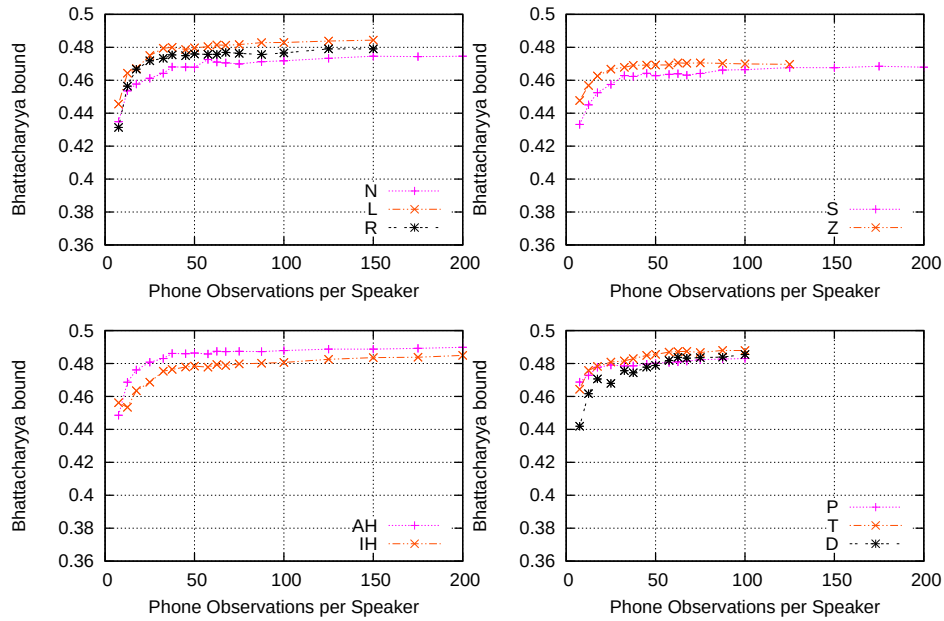


Figure 4.15: *Effect of number of phone utterances per speaker on mean of Bhattacharyya bound for different phone groups using data from 20 speakers applying bootstrapping and Cepstral Mean Normalisation*

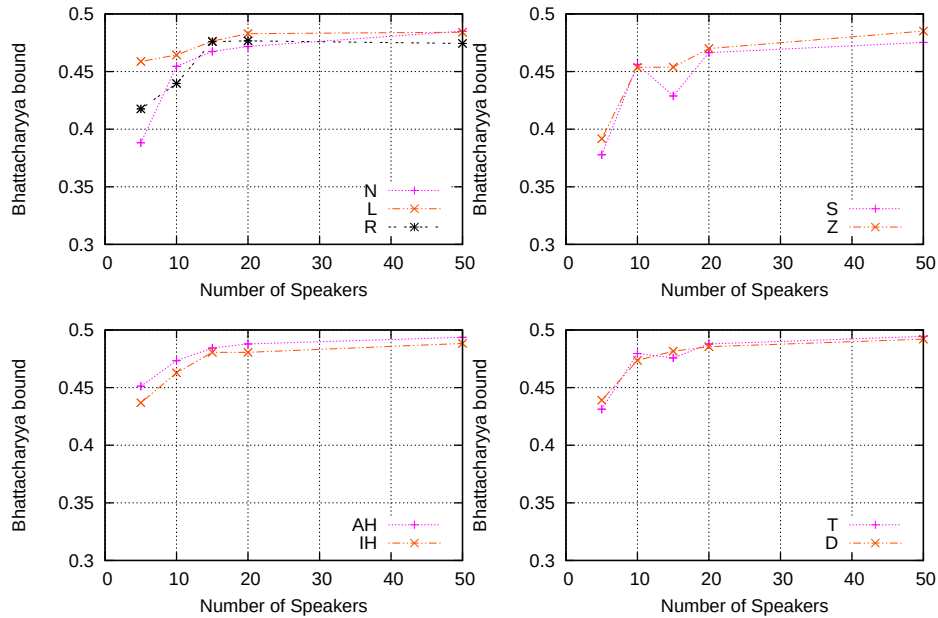


Figure 4.16: *Effect of number of speakers on mean of Bhattacharyya bound for different phone groups using 100 utterances per speaker applying bootstrapping and Cepstral Mean Normalisation*

CHAPTER FIVE

DATA SUFFICIENCY ANALYSIS OF THE LWAZI ASR CORPUS

5.1 INTRODUCTION

The work of the previous chapter successfully developed a number of techniques that can be used for data sufficiency analysis needed for further experimentation between different phones and even phones between languages. Armed with the knowledge of phone stability it becomes possible to estimate acoustic distances between languages. In this chapter we report on a data sufficiency analysis performed for a new speech recognition corpus of South African languages: the Lwazi ASR corpus.

The chapter is structured as follows:

- Training data and experimental parameters are introduced in Section 5.2;
- An analysis of the phone stability for different languages is conducted in Section 5.3;
- Section 5.4 presents the results obtained for experiments on distances between the phones of languages;
- In Section 5.5 ASR results to verify the usability of the Lwazi ASR corpus is given; and
- Lastly the impact of data reduction on ASR results is discussed (Section 5.6).

5.2 DATA AND EXPERIMENTAL DESIGN

5.2.1 LWAZI ASR CORPUS

The Lwazi ASR corpus consists of telephone speech data for all of the 11 official languages of South Africa. There are approximately 200 speakers per language and 30 utterances per speaker. The

speaker population was selected to provide a balanced profile with regard to age, gender and type of telephone. Read and elicited speech are covered for every speaker (16 sentences selected from a phonetically balanced corpus and 14 short words and phrases respectively).

Although we focus our analysis on five languages (isiNdebele, siSwati, isiZulu, Tshivenda and Afrikaans) that are part of the Lwazi corpus [35], Table 5.1 gives a summary of the amount of data available for 10 of the respective languages (9 Bantu and 1 Germanic) and also the number of distinct phones that is currently used for ASR purposes. Finally the last entry (NTIMIT) specifies these details for a telephone corpus in American English that we utilise as a control.

Language	total # minutes	# speech minutes	# distinct phonemes
Afrikaans	213	182	37
isiNdebele	564	465	46
isiXhosa	470	370	52
isiZulu	525	407	46
Tshivenda	354	286	38
Sepedi	394	301	45
Sesotho	387	313	44
Setswana	379	295	34
siSwati	603	479	39
Xitsonga	378	316	54
N-TIMIT	315	-	39

Table 5.1: A summary of audio data available for different languages

5.2.2 EXPERIMENTAL SETUP

For every phone in each of the selected target languages, all phone observations are extracted from the 150 speakers with the most utterances per phone. We use the technique described in Section 3.4.2 to conduct the phone stability measurements and a slightly different protocol is applied to evaluate model distances between phones. In both cases the data for each investigated phone is separated into 5 disjoint subsets with regard to speakers. For the stability measurements, the mean of the 10 resulting similarity values are calculated between all of the various intra-phone model pairs and the mean of 25 similarity values between the various inter-phone model pairs are taken when measuring phone distance.

ASR systems are trained on the data of every language and then used in forced alignment to label all of the utterances. We need these systems to verify recognition accuracy (Section 5.5) and also to obtain the accurate phone boundaries applied (see Section 4.7) throughout this chapter. Mel-frequency cepstral coefficients (MFCCs) with cepstral mean and variance normalisation are used as features, as described in Section 5.5.

5.3 ANALYSIS OF PHONE VARIABILITY

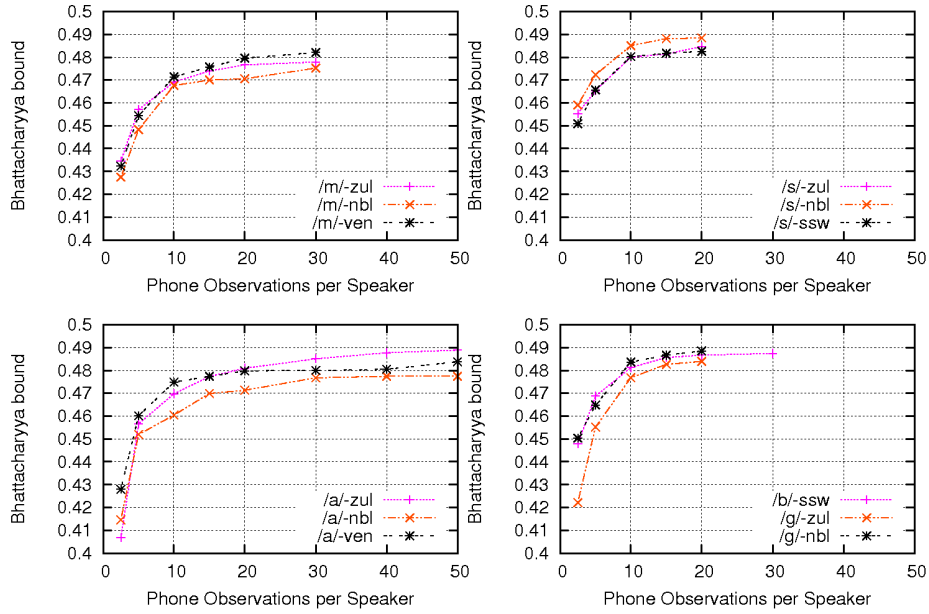


Figure 5.1: Effect of number of phone utterances per speaker on mean of Bhattacharyya bound for different phone groups using data from 30 speakers

The phone variability analysis as explained in Section 4.5 results in similar trends observed for different numbers of mixtures if Gaussian mixture models are employed. Because of the limited size of the Lwazi corpus, we only report on single-mixture, context-independent models.

As similar trends are also observed for phones within the same broad phone categories (see Section 4.3), only one or two examples from the several broad categories are selected, which occur in most of the target languages. Using SAMPA notation [36], the following phones are selected:

1. /a/ (vowel),
2. /m/ (nasal),
3. /b/ and /g/ (voiced plosives) and
4. /s/ (unvoiced fricative),

after verifying that these phones are indeed representative of the larger groups.

Figures 5.1 and 5.2 demonstrate the effect that variable numbers of phones and speakers have on the stability measurements for the selected phones. As stated in Section 4.4.1 this value should approach 0.5 for a model fully trained on a sufficiently representative set of data.

Again we see that the various broad categories of phones approach the asymptotic bound in different ways (Figure 5.1). The vowels and nasals require the largest number of phone observations to reach a given level, whereas the fricatives and plosives converge quite rapidly (with 10 observations

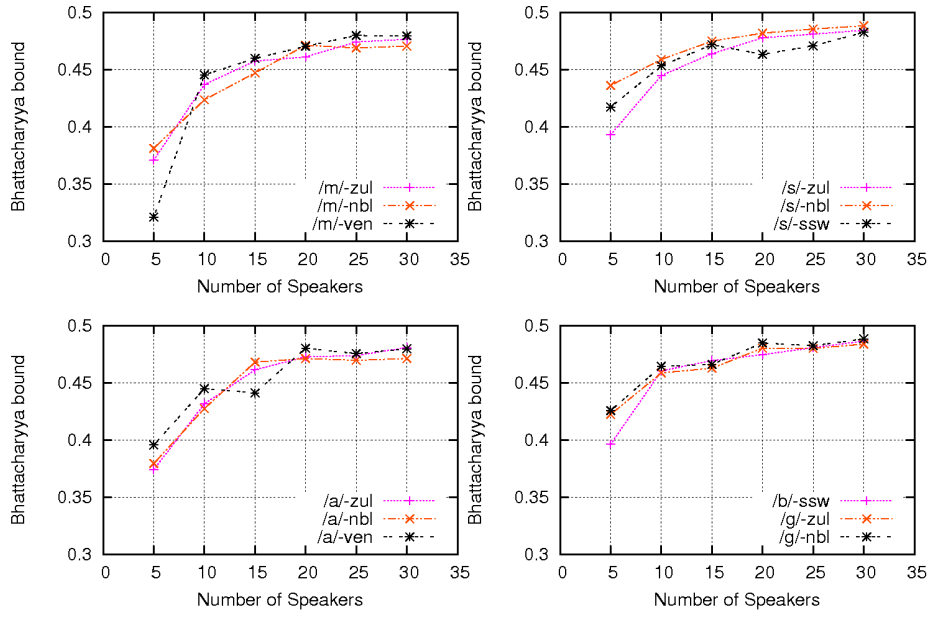


Figure 5.2: *Effect of number of speakers on mean of Bhattacharyya bound for different phone groups using 20 utterances per speaker*

per speaker, both the fricatives and plosives achieve values of 0.48 or better for all languages, in contrast to the vowels and nasals which require 30 observations to reach similar stability). Note that we employed 30 speakers per phone group, since that is the largest number achievable with the protocol.

For the results in Figure 5.2, we keep the number of phone observations fixed at 20 (this ensures sufficient data for all phones, and corresponds with reasonable convergence in Figure 5.1). It is clear that additional speakers would still improve the modelling accuracy for especially the vowels and nasals. The voiced plosives and fricatives also achieve high values for the bound rather easily (close to the ideal 0.5).

Figures 5.1 and 5.2 (as well as similar figures for the other phone classes and languages we have studied) suggest that all phone categories require at least 20 training speakers to achieve reasonable levels of convergence (stability measurement levels of 0.48 or better). The number of phone observations required per speaker is more variable, ranging from less than 10 for the voiceless fricatives to 30 or more for vowels, liquids and nasals. We return to these observations below.

5.4 DISTANCES BETWEEN LANGUAGES

In Section 1.3.3 it is pointed out that the similarities between the same phones in different languages are important predictors of the benefit achievable from pooling the data from those languages. Armed with the knowledge that stable models can be estimated with 30 speakers per phone and between 10 and 30 phone observations per speaker, we now turn to the task of measuring distances between phones in various languages.

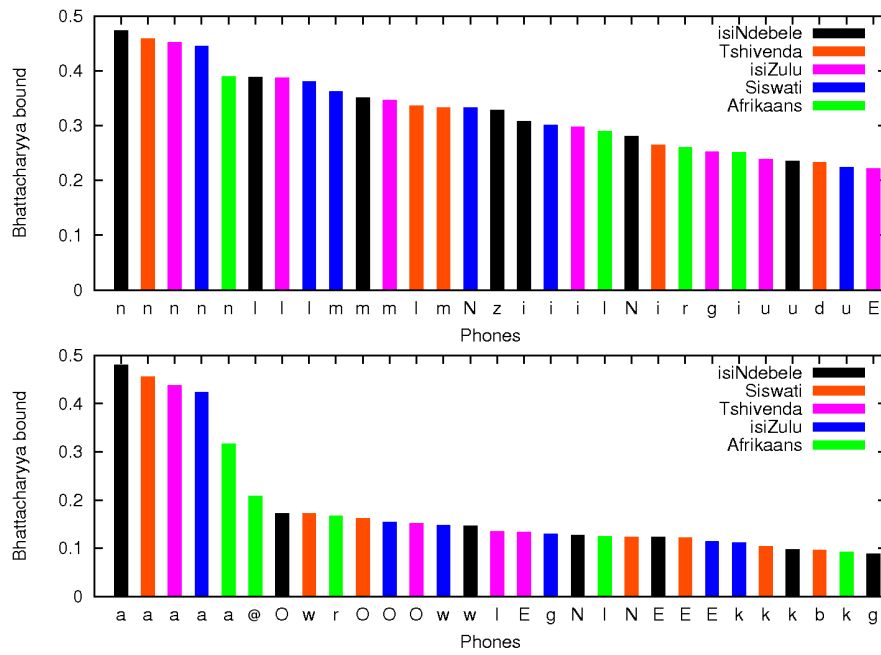


Figure 5.3: *Effective distances in terms of the mean of the Bhattacharyya bound between a single phone (/n/-nbl top and /a/-nbl bottom) and each of its closest matches within the set of phones investigated*

Again the similarity measurements based on the Bhattacharyya bound can be used to compare phones, and obtain values between all possible combinations of phones. In Figure 5.3 the results show these measurements for the isiNdebele phones /n/ and /a/ and all of the closest matches. As expected, similar phones from the different languages are closer to one another than different phones of the same language. However, the details of the distances are quite revealing: for /a/, siSwati is closest to the isiNdebele model, as would be expected given their close linguistic relationship, but for /n/, the Tshivenda model is found to be closer than either of the other Nguni languages. A second example where this happens is included in Annexure 1, where Figure A.9 shows that the siSwati /a/ is actually very close to the corresponding Tshivenda model.

For comparative purposes, we have included one non-Bantu language (Afrikaans), and we see that its models are indeed significantly more dissimilar from the isiNdebele model than any of the Bantu languages. In fact, the Afrikaans /n/ is about as distant from isiNdebele /n/ as the isiNdebele and isiZulu /l/ are.

Analysis of another isiNdebele vowel /i/ and nasal /m/ are shown in Figure 5.4. Interestingly all distances are found to be diminished and we conclude that the models for /i/ are indeed very similar across the investigated language borders. For /m/ all the Bantu languages are also even more closely related than for the model /n/.

To complete the picture, an isiNdebele fricative /s/ and plosive /g/ are also investigated. It can be seen in Figure 5.5 that immediately all of the closest matches are non-vowel sounds. We also find that models are no longer that dissimilar between the different plosive phones. For the isiNdebele /g/

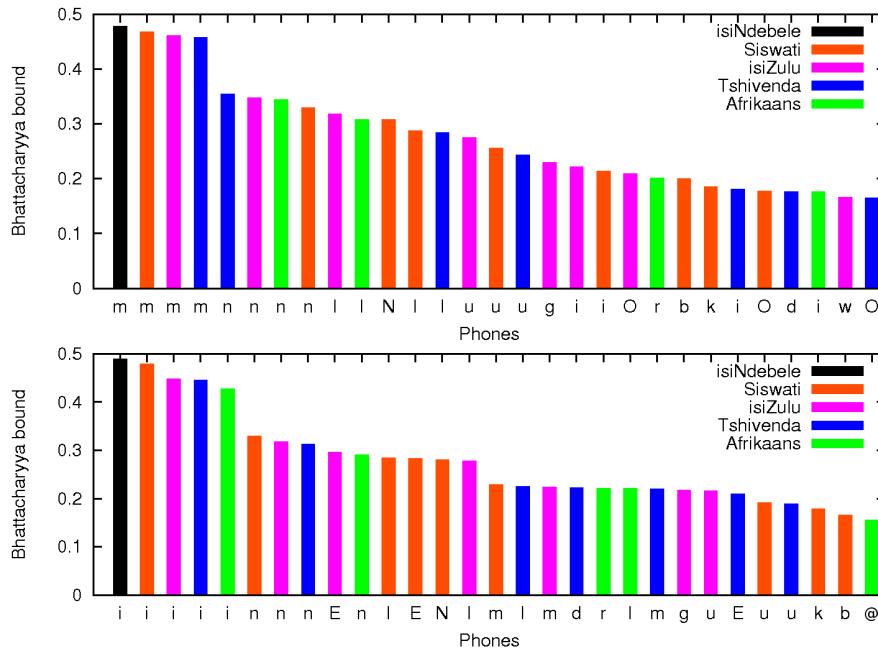


Figure 5.4: *Effective distances in terms of the mean of the Bhattacharyya bound between a single phone (/m/-nbl top and /i/-nbl bottom) and each of its closest matches within the set of phones investigated*

it can be seen that /k/ and the Afrikaans /d/ models are actually closer than models for isiZulu /g/ or the siSwati /b/.

The similarity of the fricative /s/ also proves interesting across language borders. Although the siSwati model is found to be closest, the model for Afrikaans is very similar. All of the fricative models are, however, distinct from the other investigated phones of the languages. Similar figures for a few comparative distances with regard to the other 4 languages are included as part of Annexure 1. The trends observed are similar to those described here.

5.5 ASR RESULTS

In order to verify the usability of the Lwazi ASR corpus for speech recognition purposes, initial ASR systems for 10 South African languages (Table 5.1) are constructed [37]. For these tests, data from 30 speakers per language are used as test data and the remaining data utilised for training. A summary of the phone recognition accuracies (including the control: NTIMIT) are shown in Table 5.2.

As the initial pronunciation dictionaries were developed to provide good coverage of the language in general, these dictionaries did not cover the entire ASR corpus. Grapheme-to-phoneme rules are extracted from the general dictionaries using the Default&Refine algorithm [38] and used to generate missing pronunciations.

We use HTK 3.4 to build a context-dependent cross-word HMM-based phoneme recogniser with triphone models. Each model has 3 emitting states with 7 mixtures per state. 39 features are used: 13

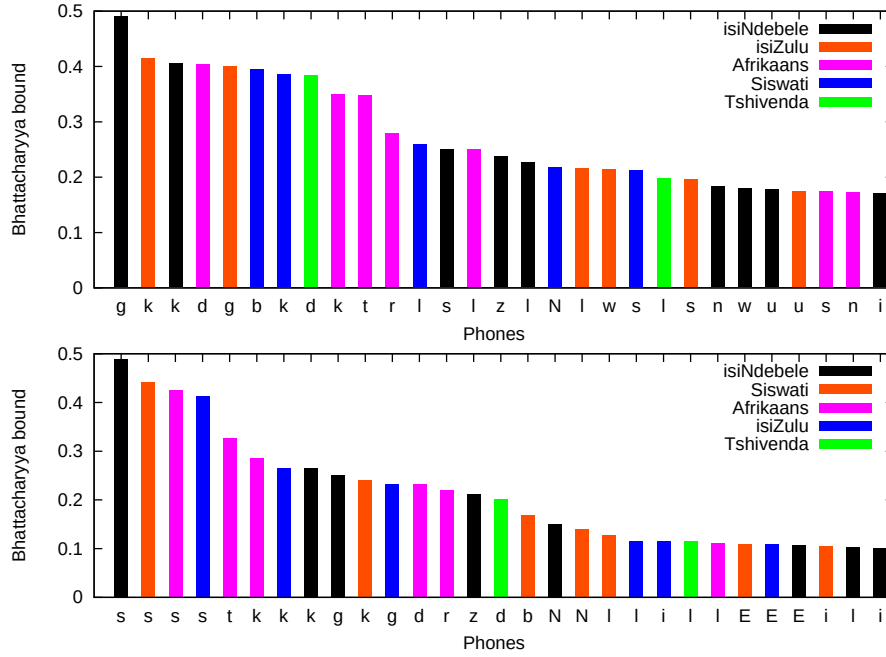


Figure 5.5: *Effective distances in terms of the mean of the Bhattacharyya bound between a single phone (/g/-nbl top and /s/-nbl bottom) and each of its closest matches within the set of phones investigated*

MFCCs together with their first and second order derivatives. Cepstral Mean Normalisation (CMN) as well as Cepstral Variance Normalisation (CMV) is used to perform speaker-independent normalisation. A diagonal covariance matrix is used; to partially compensate for the incorrect assumption of feature independence semi-tied transforms are applied. A flat phone-based language model is employed throughout [39].

As a rough benchmark of acceptable phoneme-recognition accuracy, recently reported results obtained by [40] on a similar-sized telephone corpus in American English (N-TIMIT) are also shown in Table 5.2. We see that the Lwazi results compare very well with this benchmark.

5.6 IMPACT OF DATA REDUCTION

An important issue in ASR corpus design is the trade-off between the number of speakers and the amount of data per speaker [41]. The figures in Section 5.3 are not conclusive on this trade-off, so the effect of reducing either the number of speakers or the amount of data per speaker when training the isiZulu and Tshivenda recognisers are also investigated.

At approximately 200 speakers per language, if 30 is kept separate as a test set and working with a division factor of up to 8, we end up with about 20 speakers per language as a minimum. Figure 5.6 shows the impact of both forms of reduction for the isiZulu and Tshivenda languages. Surprisingly reducing the number of speakers has approximately the same effect as reducing the amount of speech per speaker. This finding correlates well with the results of Section 5.3 where it can be seen that all

Language	% corr	% acc	avg # phons	total # speakers
isiNdebele	74.21	65.41	28.66	200
isiXhosa	69.25	57.24	17.79	210
isiZulu	71.18	60.95	23.42	201
Tshivenda	76.37	66.78	19.53	201
Sepedi	66.44	55.19	16.45	199
Sesotho	68.17	54.79	18.57	200
Setswana	69.00	56.19	20.85	207
siSwati	74.19	64.46	30.66	208
Xitsonga	70.32	59.41	14.35	199
N-TIMIT	64.07	55.73	-	-

Table 5.2: Initial results for South African ASR systems. The column labelled “avg # phones” lists the average number of phone observations for each phone for each speaker

investigated phone groups already achieve high phone stability values at 20 speakers.

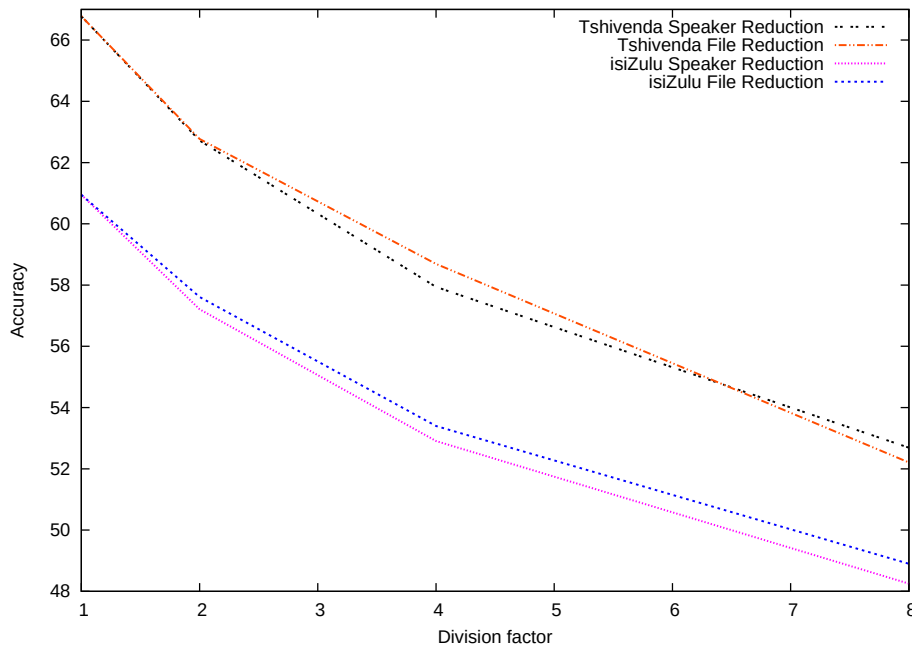


Figure 5.6: The influence of a reduction in training corpus size on phone recognition accuracy

5.7 CONCLUSION

Techniques defined in previous chapters are applied to successfully validate the data requirements for the telephone-based Lwazi ASR corpus. Initial phone recognition results correlate well with these findings and reveal important aspects related to ASR corpus design for resource scarce languages. Interestingly, for this particular corpus and the specific ASR parameters that are used, it is found that substantially fewer speakers are required for building initial recognition systems than previously

thought. Finally, the modelling measurement techniques of previous chapters are extended and used to successfully compare the variability captured by representative acoustic models for phones of different languages. Based on the findings, effective acoustic distances are deduced between these phones.

CHAPTER SIX

CONCLUSION

6.1 INTRODUCTION

As described in Section 1.2, extraordinary difficulties are encountered when building speech recognition systems for new or under-resourced languages. Compiling ASR audio corpora is a resource intensive process and because application-specific speech recognition systems require relevant data, the problem is only intensified. Multiple speakers are required to say specifically selected utterances, under the correct acoustic conditions and all of these utterances have to be orthographically transcribed.

Consequently a real need for the effective design of ASR audio corpora exists, a great deal of which can be explained acoustically. The amount of acoustic variation and similarities found within ASR training data beg the careful exploration and effective usage of any given corpus. Current phone-based speech recognition systems model phones as the basic acoustic units. These models serve to balance generalisation and memorisation of the variability seen in the features representing training data. Given a sufficient amount of accurately selected training data, an acoustic model will converge to represent the encountered variability.

However, sufficient data on its own still does not guarantee the effective representation of the information intended to be captured. Acoustic models are designed with various levels of complexity to best suit the practical purposes. In general, additional model complexity increases the memorising capabilities of an acoustic model, which then in turn may require more training data to effectively generalise for the seen variability. Sadly it is also the case that any real world acoustic channel is limited with regard to information throughput, so increased model complexity is only beneficial up to a certain point.

Modelling specific variability of training data is thus an art that balances model complexity and data requirements with the acoustics to be captured.

6.2 SUMMARY OF CONTRIBUTION

This dissertation explored the effectiveness of different ASR parameters (features, normalisation strategies and modelling schemes) to represent various types of ASR training data variability.

6.2.1 LITERATURE

A summary of how the current research relates to prior literature:

- In literature investigations with regard to different types of variability in ASR audio data are performed, but most approaches are based on making sound assumptions and building complete recognition systems to verify effectiveness of techniques with end results. Because of the phonetic similarities between related languages, these techniques are also applied to combine similar types of acoustic variability (phones) across different languages as described in Section 1.3.3.
- Data selection strategies do exist that identify and deal with different types of variability of training data 1.3.2.1 and are useful for ASR corpus design. Current ASR acoustic modelling strategies also accomplish similar purposes through the application of tree based clustering techniques. Because of the complex acoustic properties that co-articulation of phones produce, triphones are used that increase the number of models in an ASR system and also its capability to represent more subtle phonetic differences. Tree based clustering techniques then provide the control to ensure a sufficient amount of training data for each cluster and that similar triphones are clustered together. To test acoustic similarity, a distance metric is used between acoustic models representing the data of different triphones as mentioned in Section 1.3.1.1.
- Although several approaches to use distance metrics that measure the similarity between acoustic models have been tried (see Section 1.3.1.2), this is difficult to do in the context of ASR training data. As mentioned above, to accurately model the variability of speech data can be difficult and any calculated similarity values between acoustic models are subject to how representative the respective models are of the actual source of training data being measured.

6.2.2 MAIN CONTRIBUTIONS

With this dissertation we specifically set out to answer the question of how acoustic models react to additional training data, since such models can then be used as a tool to analyse the acoustic properties of speech data. In Section 1.2 it is explicitly stated that data sufficiency is a key concept of such analysis and has the practical implication that the usefulness (from an acoustic perspective) of newly developed ASR corpora can be verified. With regard to the initial objectives specified, the main contributions of this dissertation are summarised as follows:

- We defined a new stability measurement protocol that provides the capability to view the variability of ASR training data based on the acoustic modelling assumptions that are made. To investigate any specific type of variability a number of data examples are needed (eg. different speaker or phone occurrences). By selecting a number of disjoint subsets with regard to the specific type of variability being investigated and evaluating different permutations of the data examples that make up the subsets, a measurement stability analysis can be performed. Knowing the stability of these measurements, given the data and our modelling assumptions, then allow for the effective comparison of the different variations within the training data (eg. different phones).
- Acoustic model complexity is addressed in three ways: (1) complex Gaussian mixture models, (2) the effect of context dependence and (3) a bootstrapping approach. Investigation of acoustic stability for complex Gaussian mixture models required the implementation of a Monte Carlo sampling metric (see Section 2.4) due to the fact that calculation of the Bhattacharyya bound is analytically intractable for these densities. Computationally the design and implementation of this approach is quite challenging. To evaluate the bound for a number of 39-dimensional mixture components trained on the number of different training data subsets that we investigate, currently require cluster computing capabilities. We evaluated the effect of context dependence in terms of the applied stability measurement protocol. This requires a tree-based clustering approach (see Section 1.3.1.1) with appropriate acoustic thresholds to effectively divide the data into a number of significantly different data sets (acoustically). Finally, a bootstrapping approach is also implemented, that enables the use of optimal phone boundaries (obtained from an optimized speech recognition system). This effectively simulates one of the important characteristics of complex HMM models used in speech recognition systems.
- We investigated and discovered the similar trends that phones of the same phone categories follow when training acoustic models. Specifically, given the chosen type of acoustic models, phones are compared with regard to their data requirements. More complex models are also evaluated that still show the different phone categories to follow remarkably similar trends than obtained for simpler models. The effect of different ASR encoding parameters for the identified trends is also investigated, because channel normalization techniques (such as CMN) are important when working with telephone data.
- Importantly, for all of the different acoustic model complexities and ASR encoding strategies investigated, our stability measurement protocol provide a data sufficiency analysis that successfully describes the training data requirements for these various approaches. Strong indications with regard to the required number of speakers for different phones can be deduced from this analysis. The data sufficiency deductions that are made is then also tested with regard to phone recognition results of a standard phone-based ASR system. It is found that observed trends for the data sufficiency analysis on simple Gaussian acoustic models hold for the ASR

recognition accuracies.

- For telephone data, standard ASR system parameters and with the current ASR normalisation strategies, our results strongly suggest that as few as 30 speakers are needed for initial speaker independent ASR systems. These speakers are however still required to produce a sufficient amount of speech [39].
- Based on the data sufficiency analysis performed for the applied acoustic models in the context of the proposed stability measurement protocol, it is also possible to estimate distances between variations of speech data such as different phones. The stability measurements serve to characterise modelling effectiveness for the variation (phone) that is represented. Armed with this information we showed that estimated distances between different phones can be interpreted, even across language borders.

6.3 FURTHER APPLICATION AND FUTURE WORK

The presented dissertation serves as a starting point for the following different directions of future research and application:

- The investigation of new parameterisation (encoding) schemes to represent ASR audio data can be conducted similarly to the data sufficiency analysis. Tools can be developed to evaluate the data requirements and effectiveness of new features extracted from training data.
- Closely related to the above is the investigation and development of channel normalisation strategies. The stability measurement protocol can be applied to both the original and normalised data to characterise the specific changes in data variability.
- In Section 3.2 the most prominent research topics on variation in speech that impact ASR performance was pointed out. As for channel normalisation, the protocol defined in this dissertation can be applied to investigate the effect that these variations (eg. emotion, child speech, speaking styles, accents, etc.) have on the speech recognition parameters.
- The investigation of different pronunciations may eventually lead to the construction of tools to assist with the development of pronunciation dictionaries and the definitions of phone sets used with ASR systems. Based on the acoustic evidence of training data, suggestions of acoustically dissimilar words and sentences may be selected automatically for use in, for example, Spoken Dialog Systems (SDSs) as described in Section 1.3.4. Acoustic confusability (see Section 1.3.2.2) may be tested and systems may be built that interrupt the user appropriately to gain acoustic confidence [42].
- CALL (Computer Assisted Language Learning) approaches may benefit from data sufficiency analysis. The acoustic component of these applications typically relies on the effective mea-

surement of variations in audio data. A better understanding of the impact that the parameters of such a system have on these measurements can aid the future design of this software.

- Lastly the approaches to combine ASR training data for resource-scarce environments to build recognition systems, as discussed in Section 1.3.3, can be improved with an understanding of the acoustic properties that different sets of training data mandates.

6.4 CONCLUSION

With this dissertation a new way to analyse acoustic variability in the context of ASR training data is explored. It is shown that data sufficiency is closely related to the amount of variability in training data and that the similarity of different types of variations can be compared using similarity measures between acoustic models. Such an analysis is useful to understand the practical implications of working with limited ASR audio data in resource scarce scenarios.

APPENDIX A

ANNEXTURE

A.1 EFFECT OF CONTEXT-DEPENDENCE

The tables A.1,A.2,A.3 and A.4 represent the triphone clusters obtained in Section 4.6. Due to the large number of contexts, however only a selected few are shown to convey the general idea. In Figure A.1 the learning behaviour of the respective parent phones are also shown.

Triphone-cluster label	Triphones
AH-1	IY-AH+W, Y-AH+B, P-AH, G-AH+F, IY-AH+G, B-AH+F, K-AH+F, G-AH+JH, IY-AH+B, L-AH+HH, L-AH+NG, AH+NG, CH-AH+B, T-AH+CH, Z-AH+V, AH, D-AH+HH, HH-AH+W, SH-AH+V, IY-AH+CH, SH-AH+P, HH-AH+TH, D-AH+G, D-AH+F, UW-AH+T, AY-AH+P, R-AH+CH, V-AH+B, UW-AH+B, AY-AH+B, UW-AH, K-AH+HH...
AH-4	DH-AH+S, DH-AH
AH-6	HH-AH+D, AH+D, JH-AH+D, B-AH+D, D-AH+D, F-AH+D, L-AH+D, N-AH+D, P-AH+D, R-AH+D, T-AH+D, V-AH+D, Z-AH+D, TH-AH+D, G-AH+D, UW-AH+D, K-AH+D, S-AH+D, W-AH+D, IY-AH+D
AH-7	AW-AH+L, AY-AH+L, AO-AH+L, OW-AH+L, AH+L, DH-AH+L, OY-AH+L, ER-AH+L, CH-AH+L, JH-AH+L, B-AH+L, D-AH+L, F-AH+L, N-AH+L, HH-AH+L, P-AH+L, R-AH+L, T-AH+L, V-AH+L, Z-AH+L, TH-AH+L, SH-AH+L, G-AH+L, UW-AH+L, K-AH+L, S-AH+L, W-AH+L, IY-AH+L, Y-AH+L
AH-9	AY-AH+N, OW-AH+N, AH+N, ER-AH+N, CH-AH+N, JH-AH+N, B-AH+N, D-AH+N, F-AH+N, ZH-AH+N, M-AH+CH, M-AH+DH, M-AH+HH, M-AH+JH, M-AH+NG, M-AH+SH, M-AH+TH, L-AH+N, N-AH+N, HH-AH+N, P-AH+N, R-AH+N, T-AH+N, V-AH+N, Z-AH+N, TH-AH+N, G-AH+N, UW-AH+N, K-AH+N, M-AH+B, M-AH+D, M-AH+G...

Table A.1: *Triphone clusters for different contexts of the phone /ah/*

Triphone-cluster label	Triphones
N-1	AA-N, AE-N, AH-N, IY-N, AO-N, AW-N, AY-N, R-N, UW-N, EH-N, ER-N, EY-N, OW-N, OY-N, IH-N
N-2	AH-N+B, AH-N+F, AH-N+G, AH-N+K, AH-N+L, AH-N+M, AH-N+N, AH-N+P, AH-N+R, AH-N+V, AH-N+W, AH-N+Y, SH-N+ER, EH-N+F, EH-N+G, EH-N+K, EH-N+L, EH-N+R, EH-N+V, EH-N+W, EH-N+Y, AH-N+AA, AH-N+AE, AH-N+AH, AH-N+AO, AH-N+AW, AH-N+AY, AH-N+CH, IH-N+F, IH-N+G, IH-N+K, IH-N+M...
N-3	AH-N+T, EH-N+T, IH-N+T, AA-N+T, AW-N+T, R-N+T, OW-N+T, AY-N+T, EY-N+T, AE-N+T, OY-N+T
N-4	AH-N+S, AH-N+Z, EH-N+S, EH-N+Z, IH-N+S, IH-N+Z, AA-N+S, AA-N+Z, AW-N+S, AW-N+Z, R-N+Z, OW-N+Z, UW-N+Z, AY-N+S, AY-N+Z, EY-N+S, EY-N+Z, AO-N+S, AO-N+Z, IY-N+S, IY-N+Z, AE-N+S, AE-N+Z, OY-N+Z, ER-N+S, ER-N+Z
N-5	AH-N+D, EH-N+D, IH-N+D, AA-N+D, UH-N+D, AW-N+D, R-N+D, OW-N+D, AY-N+D, EY-N+D, AO-N+D, IY-N+D, AE-N+D, OY-N+D, ER-N+D

Table A.2: Triphone clusters for different contexts of the phone /n/

Triphone-cluster label	Triphones
S-1	S+AA, S+AE, S+AH, S+AO, S+AW, S+AY, S+EH, S+ER, S+EY, S+IH, S+IY, S+OW, S+OY, S+UH, S+UW, S+F, S+K, S+L, S+M, S+N, S+P, S+W
S-2	ER-S+K, ER-S+P, ER-S+W, AW-S+AH, AW-S+HH, AW-S+IH, AH-S+F, AH-S+K, AH-S+L, AH-S+M, AH-S+N, AH-S+P, AY-S+AH, AY-S+EH, AY-S+ER, AY-S+IH, AY-S+IY, OW-S+AA, OW-S+AH, OW-S+AO, OW-S+EH, OW-S+ER, OW-S+IH, OW-S+IY, EH-S+F, EH-S+K, EH-S+L, EH-S+P, OY-S+AH, OY-S+CH, AA-S+AA, AA-S+AH...
S-3	AA-S, AE-S, AH-S, IY-S, AO-S, AW-S, AY-S, F-S, TH-S, K-S, L-S, N-S, P-S, R-S, T-S, UW-S, EH-S, ER-S, EY-S, K-S+D, K-S+F, K-S+G, K-S+K, K-S+M, K-S+N, K-S+P, K-S+W, OW-S, OY-S, K-S+AA, K-S+AE, K-S+AH, K-S+AO, K-S+AY, K-S+CH...
S-4	NG-S+T, ER-S+T, AH-S+T, EH-S+T, M-S+T, IH-S+T, Z-S+T, AA-S+T, P-S+T, S-S+T, S+T, V-S+T, L-S+T, AW-S+T, B-S+T, R-S+T, OW-S+T, UW-S+T, K-S+T, N-S+T, AY-S+T, EY-S+T, D-S+T, AO-S+T, IY-S+T, AE-S+T, G-S+T, T-S+T

Table A.3: Triphone clusters for different contexts of the phone /s/

A.2 PHONE STABILITY

In this section an additional number of phone stability plots is shown to augment those of Section 5.3.

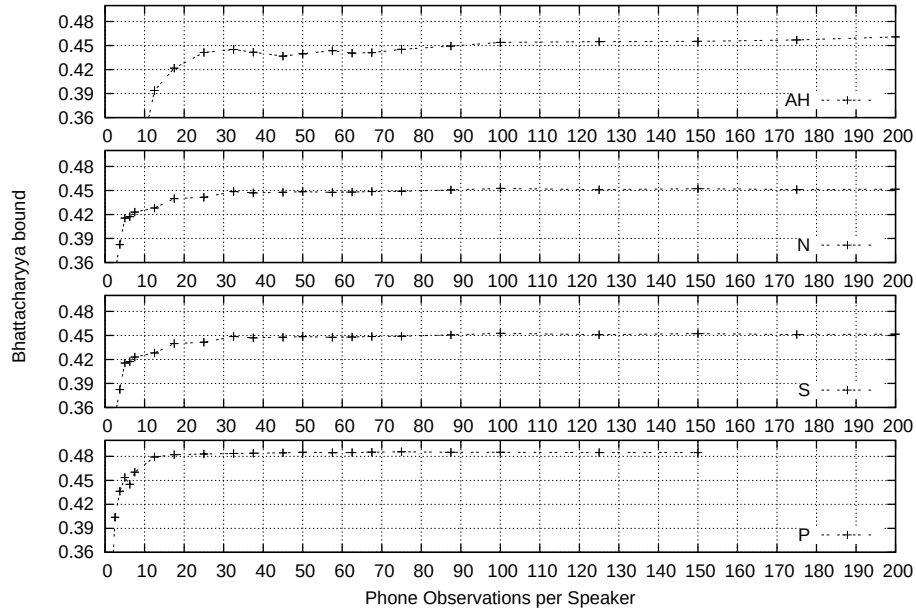


Figure A.1: Effect of number of phone utterances per speaker on mean of Bhattacharyya bound for the parent phones /ah/, /n/, /s/ and /p/ using data from 20 speakers

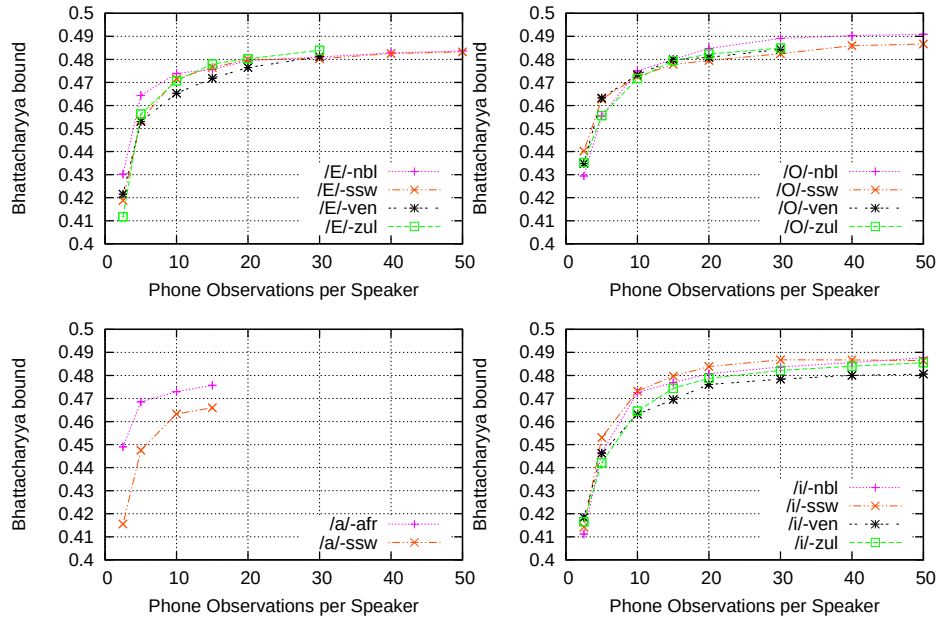


Figure A.2: Effect of number of phone utterances per speaker on mean of Bhattacharyya bound for different phone groups using data from 30 speakers

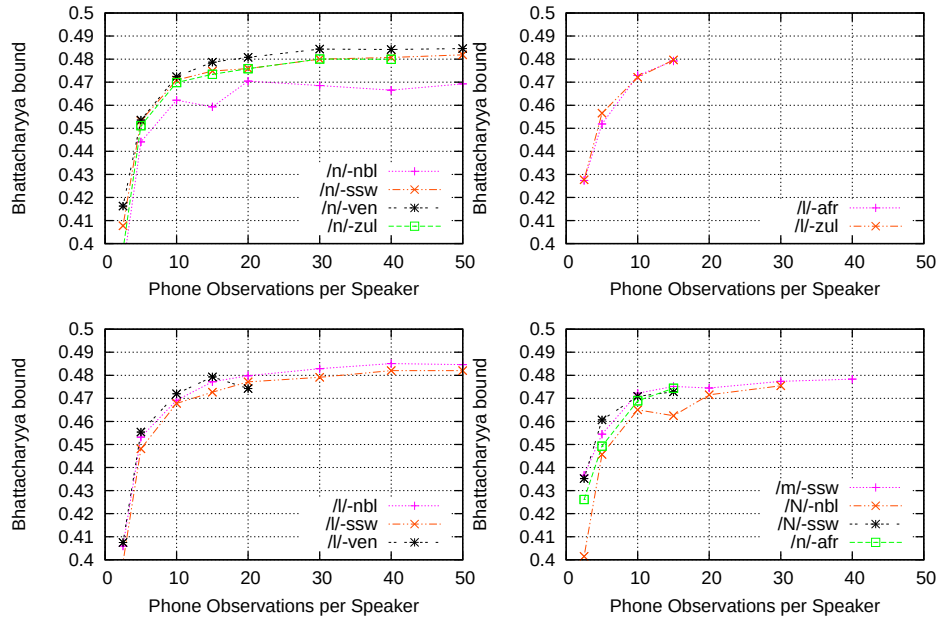


Figure A.3: Effect of number of phone utterances per speaker on mean of Bhattacharyya bound for different phone groups using data from 30 speakers

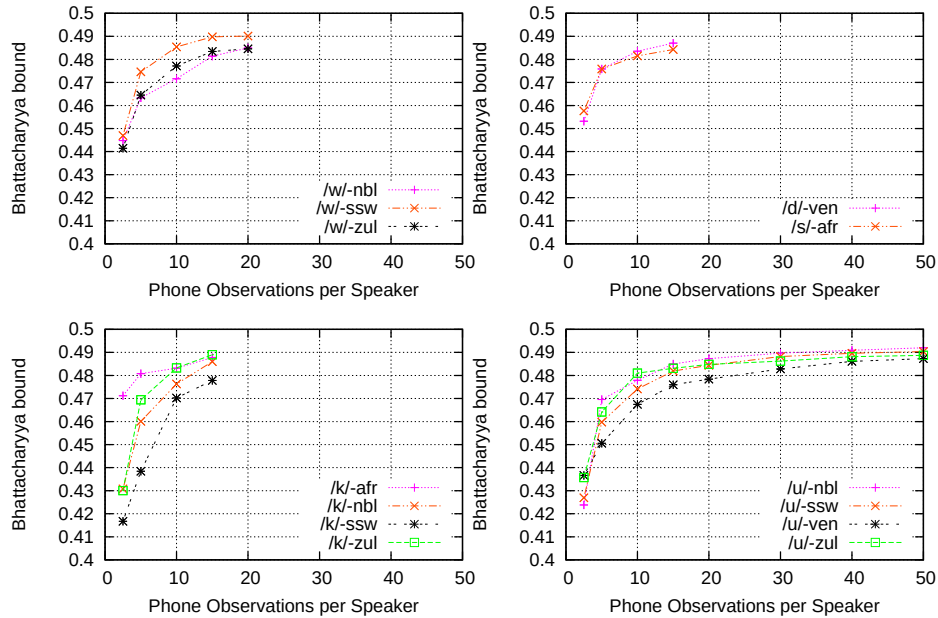


Figure A.4: Effect of number of phone utterances per speaker on mean of Bhattacharyya bound for different phone groups using data from 30 speakers

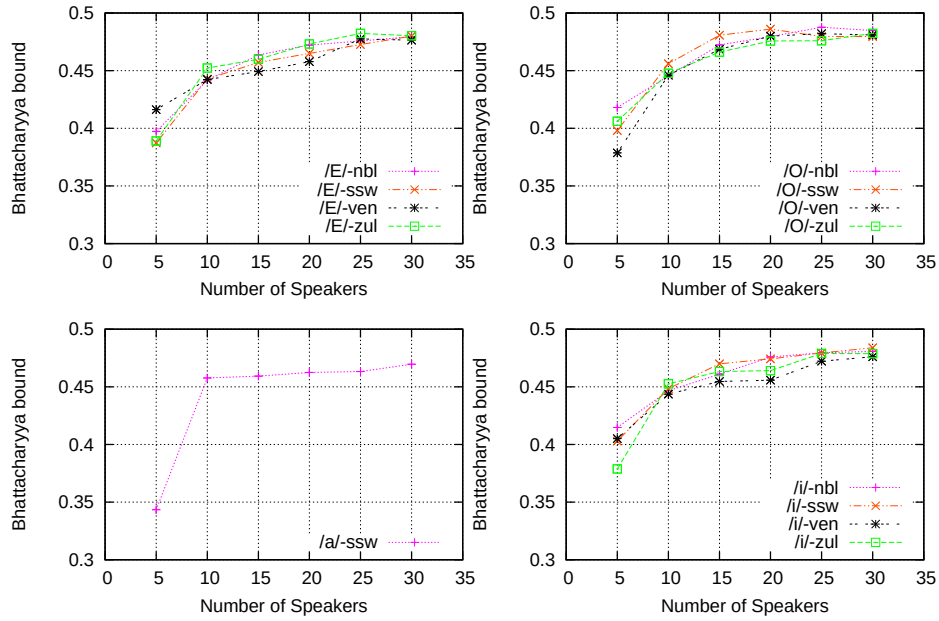


Figure A.5: Effect of number of speakers on mean of Bhattacharyya bound for different phone groups using 20 utterances per speaker

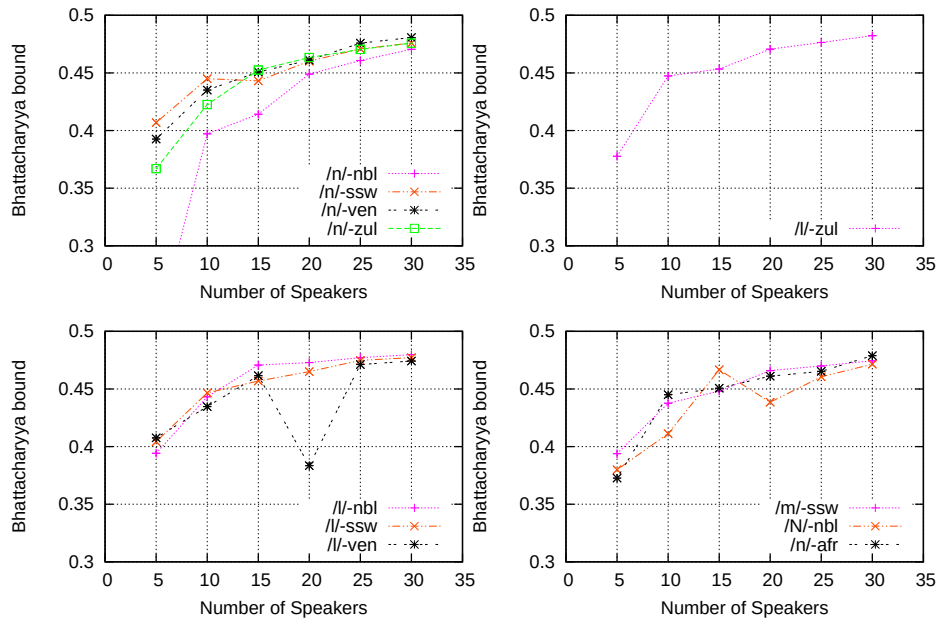


Figure A.6: Effect of number of speakers on mean of Bhattacharyya bound for different phone groups using 20 utterances per speaker

Triphone-cluster label	Triphones
P-1	AA-P, ER-P+L, ER-P+R, AH-P+K, AH-P+L, AH-P+M, AH-P+R, AH-P+S, AH-P+T, AH-P+W, AE-P, AH-P, R-P+AH, R-P+AO, R-P+ER, IY-P, T-P+AE, T-P+ER, T-P+EY, T-P+OW, T-P+UH, EH-P+R, EH-P+S, EH-P+T, EH-P+Y, AY-P, M-P+L, M-P+R, M-P+S, M-P+T, M-P+Y, IH-P+L...
P-2	AH-P+IH, EY-P+IH, K-P+IH, M-P+IH, S-P+IH, EH-P+IH, IY-P+IH, OW-P+IH, P+IH, AA-P+IH, AE-P+IH, IH-P+IH, L-P+IH...

Table A.4: Triphone clusters for different contexts of the phone /p/

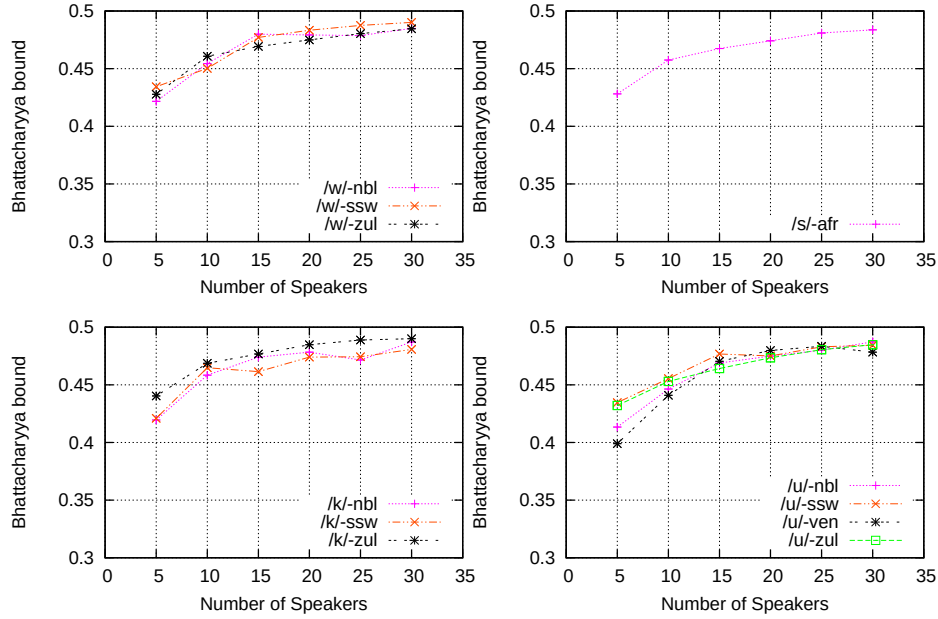


Figure A.7: Effect of number of speakers on mean of Bhattacharyya bound for different phone groups using 20 utterances per speaker

A.3 DISTANCES BETWEEN LANGUAGES

In Section 5.4 only the effective distances in terms of the isiNdebele (nbl) phones we shown. This section expand this distance plots to include the effective distances from the perspective of the other 4 languages as well.

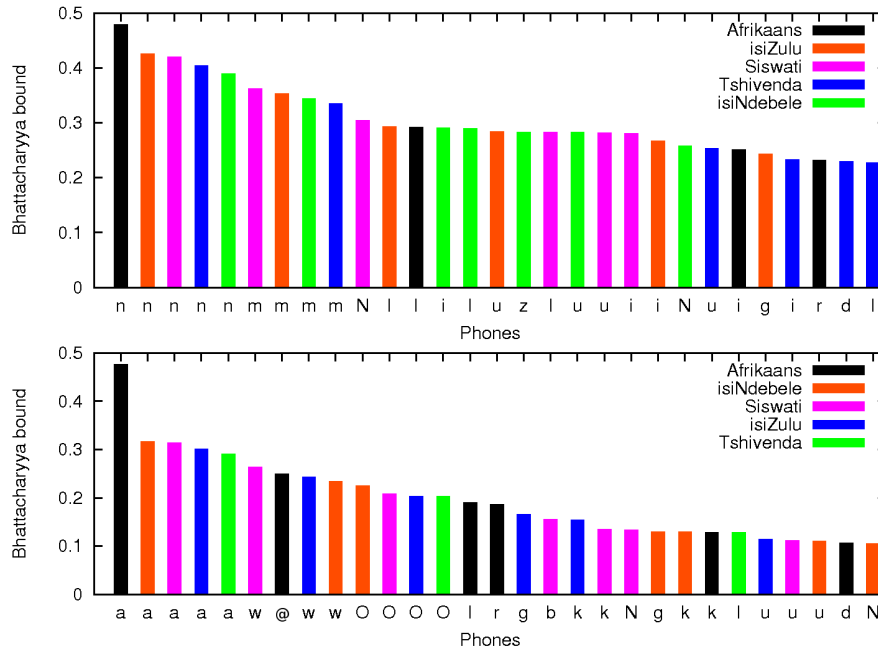


Figure A.8: Effective distances in terms of the mean of the Bhattacharyya bound between a single phone (/n/-afr top and /a/-afr bottom) and each of its closest matches within the set of phones investigated

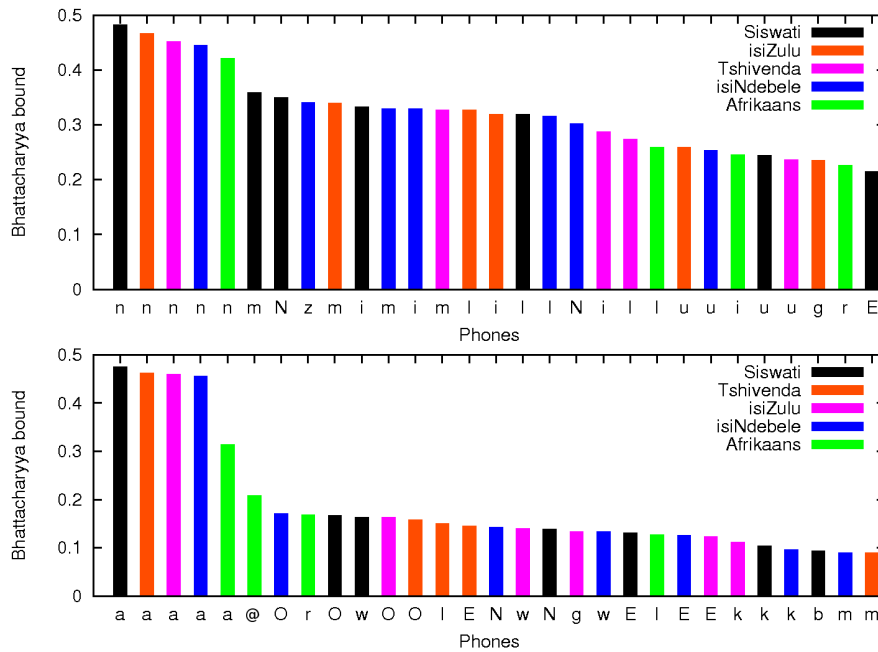


Figure A.9: Effective distances in terms of the mean of the Bhattacharyya bound between a single phone (/n/-ssw top and /a/-ssw bottom) and each of its closest matches within the set of phones investigated

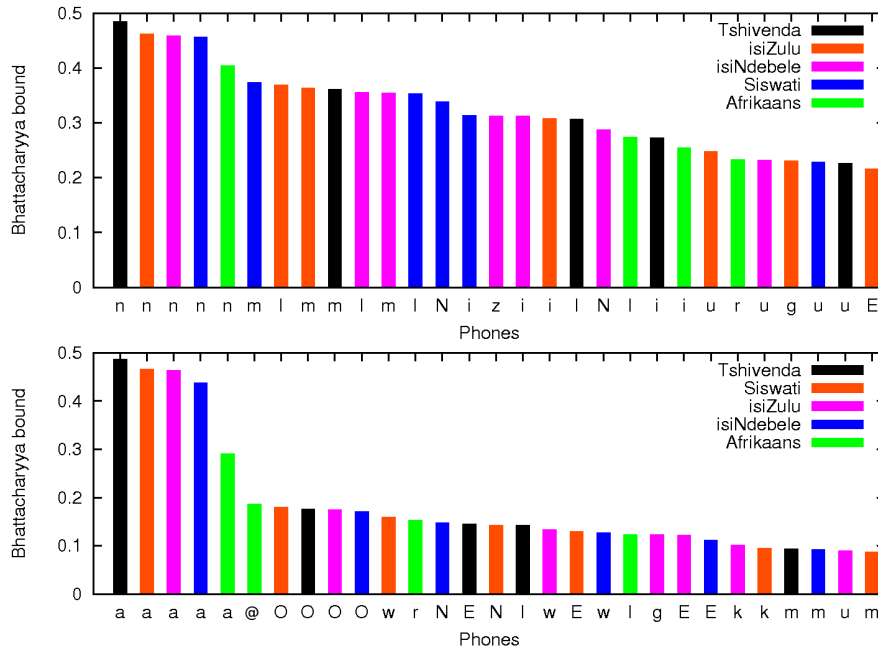


Figure A.10: Effective distances in terms of the mean of the Bhattacharyya bound between a single phone (/n/-ven top and /a/-ven bottom) and each of its closest matches within the set of phones investigated

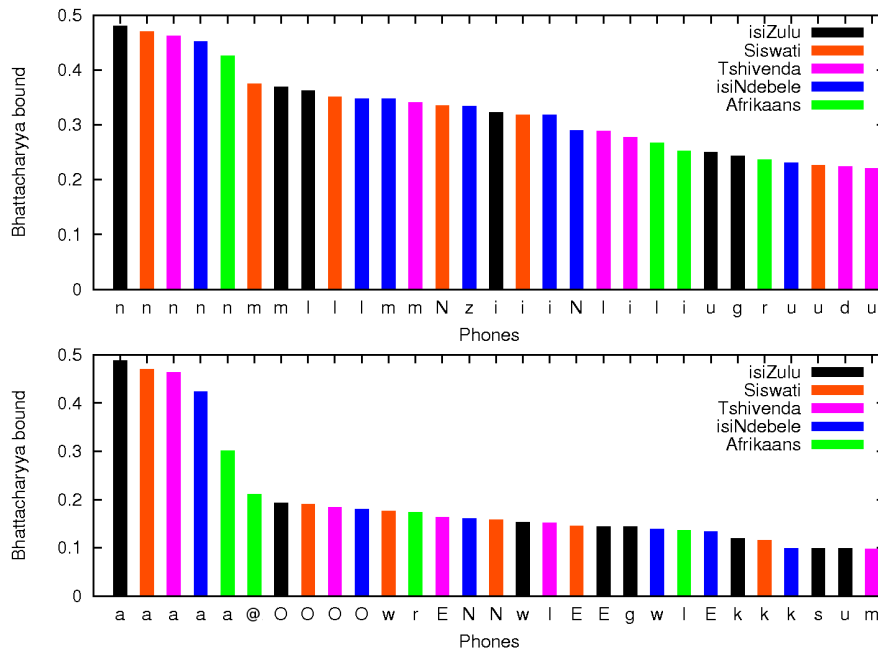
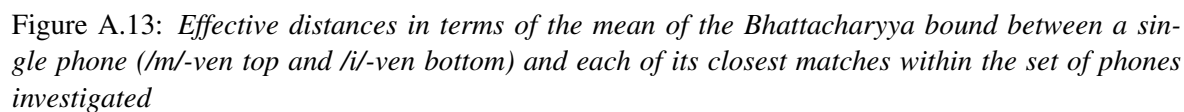
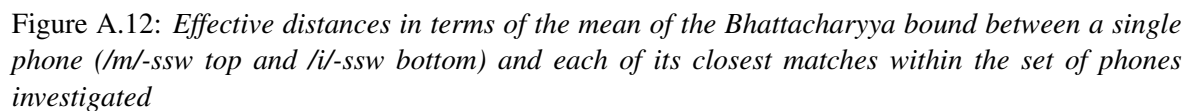


Figure A.11: Effective distances in terms of the mean of the Bhattacharyya bound between a single phone (/n/-zul top and /a/-zul bottom) and each of its closest matches within the set of phones investigated



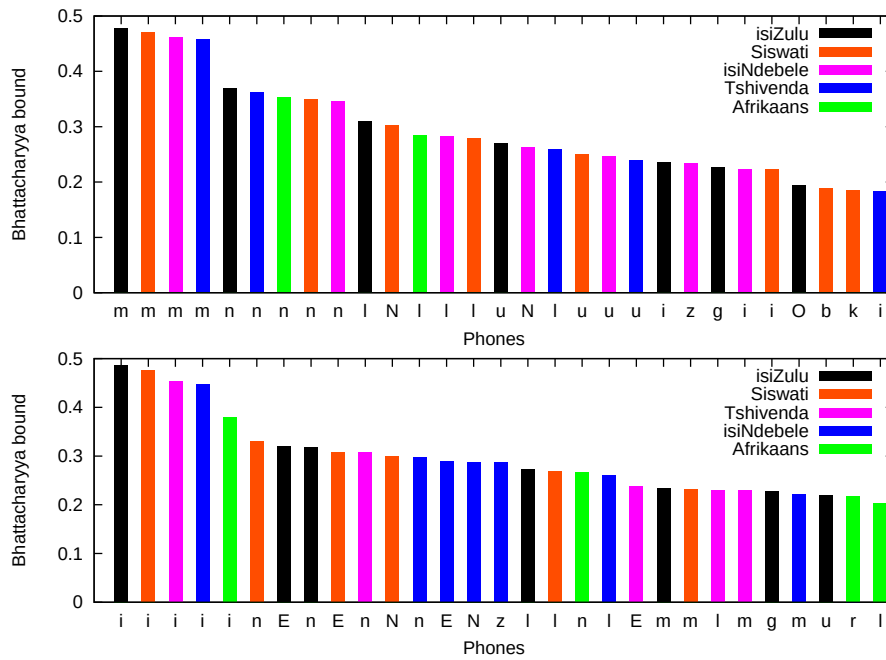


Figure A.14: *Effective distances in terms of the mean of the Bhattacharyya bound between a single phone (/m/-zul top and /i/-zul bottom) and each of its closest matches within the set of phones investigated*

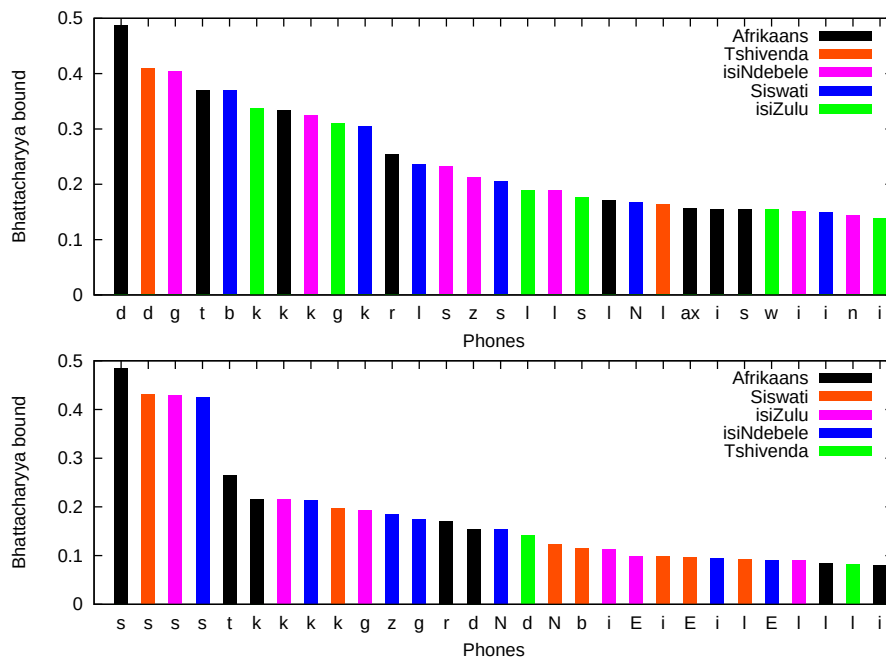


Figure A.15: *Effective distances in terms of the mean of the Bhattacharyya bound between a single phone (/d/-afr top and /s/-afr bottom) and each of its closest matches within the set of phones investigated*

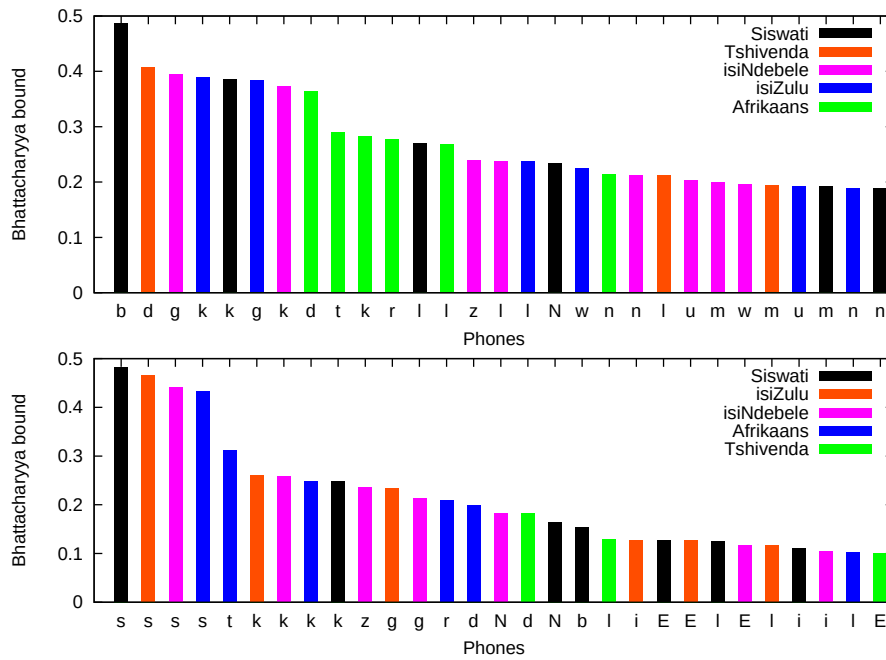


Figure A.16: Effective distances in terms of the mean of the Bhattacharyya bound between a single phone (/b/-ssw top and /s/-ssw bottom) and each of its closest matches within the set of phones investigated

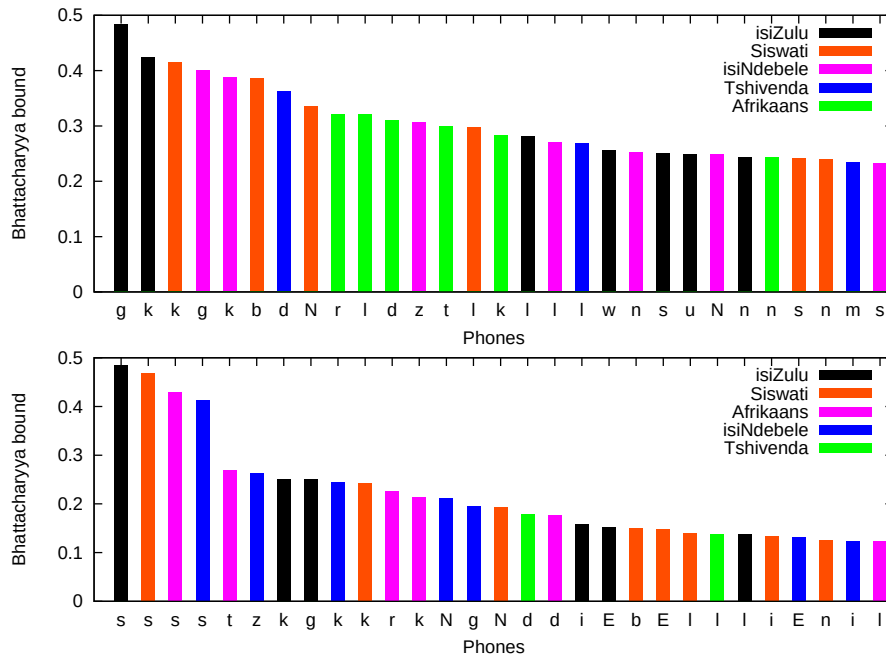


Figure A.17: Effective distances in terms of the mean of the Bhattacharyya bound between a single phone (/g/-zul top and /s/-zul bottom) and each of its closest matches within the set of phones investigated

REFERENCES

- [1] C.M. Bishop, *Pattern Recognition and Machine learning*, Springer Science and Business Media, "http://research.microsoft.com/ cmbishop/PRML", 2006.
- [2] C. Liu and L. Melnar, "Training acoustic models with speech data from different languages," in *Proceedings of ISCA Workshop on Multilingual Speech and Language Processing*, Stellenbosh, South Africa, April 2006.
- [3] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Veltchev, and P. Woodland, *The HTK Book*, Cambridge University Engineering Department, <http://htk.eng.cam.ac.uk/>, 2005.
- [4] S.J. Young, J.J. Odell, and P.C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *In Proceedings of ARPA Workshop on Human Language Technology*, Plainsboro, March 1994, pp. 307–312.
- [5] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, Inc., 2nd edition, 1990.
- [6] P.A. Olsen and J.R. Hershey, "Bhattacharyya error and divergence using variational importance sampling," in *Proceedings of Interspeech*, Antwerp, Belgium, August 2007, pp. 46–49.
- [7] B. Mak and E. Barnard, "Phone clustering using the Bhattacharyya distance," in *Proceedings of ICSLP*, October 1996, vol. 4, pp. 2005–2008.
- [8] J.-J. Aucouturier and F. Pachet, "Improving timbre similarity: How high's the sky?," *Journal of Negative Results in Speech and Audio Sciences*, vol. 1, no. 1, pp. 1–13, 2004.
- [9] Y. Qi, J.W. Paisley, and L. Carin, "Music analysis using hidden markov mixture models," *IEEE Transactions on Signal Processing*, vol. 55, pp. 5209–5224, November 2007.
- [10] M. Falkhausen, H. Reininger, and D. Wolf, "Calculation of distance measures between hidden markov models," in *In proceedings of Eurospeech*, Madrid, Spain, 18-21 September 1995, pp. 1487–1490.

- [11] M. Vihola, M. Harju, P. Salmela, J. Suontausta, and J. Savela, "Two dissimilarity measures for HMMS and their application in phoneme model clustering," in *In proceedings of ICASSP*, Orlando, Florida, 13-17 May 2002, vol. 1, pp. 933–936.
- [12] J. Silva and S.S. Narayanan, "A statistical discrimination measure for Hidden Markov Models based on divergence," in *Proceedings of Interspeech*, Jeju Island, Korea, 4-8 October 2004, pp. 657–660.
- [13] A. Nagroski, L. Boves, and H. Steeneken, "In search of optimal data selection for training of automatic speech recognition systems," *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 67–72, 30 November - 3 December 2003.
- [14] G. Riccardi and D. Hakkani-Tur, "Active and unsupervised learning for automatic speech recognition," in *Proceedings of EUROSPEECH*, Geneva, Switzerland, 1-4 September 2003, pp. 1825–1828.
- [15] Y. Wu, R. Zhang, and A. Rudnicky, "Data selection for speech recognition," *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 562–565, 9 - 13 December 2007.
- [16] E. Fosler-Lussier, I. Amdal, and H.J. Kuo, "A framework for predicting speech recognition errors," *Speech Communication*, vol. 46, no. 2, pp. 153–170, June 2005.
- [17] B. Raj R. Singh and R. M. Stern, "Structured redefinition of sound units by merging and splitting for improved speech recognition," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Beijing, China, 16-20 October 2000, vol. 3, pp. 151–154.
- [18] T. Schultz and A. Waibel, "Language-independent and language-adaptive acoustic modeling for speech recognition," *Speech Communication*, vol. 35, pp. 31–51, August 2001.
- [19] C. Nieuwoudt and E.C. Botha, "Cross-language use of acoustic information for automatic speech recognition," *Speech Communication*, vol. 38, pp. 101–113, September 2002.
- [20] T. Niesler, "Language-dependent state clustering for multilingual acoustic modeling," *Speech Communication*, vol. 49, pp. 453–463, June 2007.
- [21] B. Imperl, Z. Kacic, B. Horvat, and A. Zgank, "Clustering of triphones using phoneme similarity estimation for the definition of a multilingual set of triphones," *Speech Communication*, vol. 39, pp. 353–366, February 2003.
- [22] P. Beyerlein, W. Byrne, J. M. Huerta, S. Khudanpur, B. Marthi, J. Morgan, N. Peterek, J. Picone, and W. Wang, "Towards language independent acoustic modeling," in *Proceedings of the Acoustics, Speech, and Signal Processing, IEEE International Conference on*, June 2000, vol. 2, pp. 1029–1032.

-
- [23] P.C. Woodland, "Speaker adaptation: Techniques and challenges," in *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*, 2000, pp. 85–90.
- [24] D. Colibro, C. Vair, F. Castaldo, E. Dalmaso, and P. Laface, "Speaker recognition using channel factors feature compensation," *Proceedings of IEEE Workshop on Speaker and Language Recognition*, pp. 1–6, June 2006.
- [25] R. Kuhn, J. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Transactions on speech and audio processing*, vol. 35, pp. 695–707, November 2000.
- [26] R. Tucker and K. Shalanova, "The Local Language Speech Technology Initiative," in *SCALLA Conf.*, Nepal, 2004.
- [27] P. Nasfors, "Efficient voice information services for developing countries," M.S. thesis, Department of Information Technology, Uppsala University, 2007.
- [28] J. Sherwani, N. Ali, S. Mirza, A. Fatma, Y. Memon, M. Karim, R. Tongia, and R. Rosenfeld, "Healthline: Speech-based access to health information by low-literate users," in *IEEE Int. Conf. on ICTD*, 2007, pp. 131–139.
- [29] A. Sharma, M. Plauche, C. Kuun, and E. Barnard, "HIV health information access using spoken dialogue systems: Touchtone vs. speech," Accepted at IEEE Int. Conf. on ICTD, 2009.
- [30] E. Barnard, L. Cloete, and H. Patel, "Language and technology literacy barriers to accessing government services," *Lecture Notes in Computer Science*, vol. 2739, pp. 37–42, 2003.
- [31] 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 Communication*, vol. 49, no. 10-11, pp. 763–786, October 2007.
- [32] M. Westphal, "The use of cepstral means in conversational speech recognition," in *In proceedings of Eurospeech*, Rhodes, Greece, September 1997, pp. 1143–1146.
- [33] "November 1992 ARPA continuous speech recognition Wall Street Journal Corpus (WSJ)," 1992-2008, <http://www ldc.upenn.edu/Catalog>.
- [34] R. L. Weide, "The CMU pronunciation dictionary, release 0.6," 1998, <http://www.speech.cs.cmu.edu/cgi-bin/cmudict>.
- [35] Meraka-Institute, "Lwazi ASR corpus," 2009, Online: <http://www.meraka.org.za/lwazi>.
- [36] D. Gibbon, R. Moore, and R. Winski, *Handbook of Standards and Resources for Spoken Language Systems*, Berlin and New York: Mouton de Gruyter, www.phon.ucl.ac.uk/home/sampa, 1997.

- [37] E. Barnard, M. Davel, and C. Van Heerden, “ASR corpus design for resource-scarce languages,” in *Submitted to Interspeech*, Brighton, U.K., September 2009.
- [38] M. Davel and E. Barnard, “Pronunciation predication with Default&Refine,” *Computer Speech and Language*, vol. 22, pp. 374–393, October 2008.
- [39] J. Badenhorst, C. Van Heerden, M. Davel, and E. Barnard, “Collecting and evaluating speech recognition corpora for nine Southern Bantu languages,” in *Proceedings of the First Workshop on Language Technologies for African Languages*, Athens, Greece, March 2009, pp. 1–8, Association for Computational Linguistics.
- [40] N. Morales, J. Tejedor, J. Garrido, J. Colas, and D.T. Toledano, “STC-TIMIT: Generation of a single-channel telephone corpus,” in *LREC*, Marrakech, Morocco, 2008, pp. 391–395.
- [41] B. Wheatley, K. Kondo, W. Anderson, and Y. Muthusamy, “An evaluation of cross-language adaptation for rapid HMM development in a new language,” in *ICASSP*, Adelaide, 1994, pp. 237–240.
- [42] D. Schlangen and G. Skantze, “A general, abstract model of incremental dialogue processing,” in *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, Athens, Greece, March 2009, pp. 710–718, Association for Computational Linguistics.