AUTOMATIC SPEECH SEGMENTATION WITH LIMITED DATA

DANIEL RUDOLPH VAN NIEKERK

AUTOMATIC SPEECH SEGMENTATION WITH LIMITED DATA

by

D.R. van Niekerk

Dissertation submitted in fulfilment of the requirements for the degree

Master of Engineering

at the

Potchefstroom Campus

of the

NORTH-WEST UNIVERSITY

Supervisor: Professor E. Barnard

May 2009

AUTOMATIC SPEECH SEGMENTATION WITH LIMITED DATA

The rapid development of corpus-based speech systems such as concatenative synthesis systems for under-resourced languages requires an efficient, consistent and accurate solution with regard to phonetic speech segmentation. Manual development of phonetically annotated corpora is a time consuming and expensive process which suffers from challenges regarding consistency and reproducibility, while automation of this process has only been satisfactorily demonstrated on large corpora of a select few languages by employing techniques requiring extensive and specialised resources.

In this work we considered the problem of phonetic segmentation in the context of developing small prototypical speech synthesis corpora for new under-resourced languages. This was done through an empirical evaluation of existing segmentation techniques on typical speech corpora in three South African languages. In this process, the performance of these techniques were characterised under different data conditions and the efficient application of these techniques were investigated in order to improve the accuracy of resulting phonetic alignments.

We found that the application of baseline speaker-specific Hidden Markov Models results in relatively robust and accurate alignments even under extremely limited data conditions and demonstrated how such models can be developed and applied efficiently in this context. The result is segmentation of sufficient quality for synthesis applications, with the quality of alignments comparable to manual segmentation efforts in this context. Finally, possibilities for further automated refinement of phonetic alignments were investigated and an efficient corpus development strategy was proposed with suggestions for further work in this direction.

Keywords: phonetic speech segmentation, phonetic alignment, speech synthesis, text-to-speech, speech corpus development, resource scarce languages, Hidden Markov Models, Dynamic Time Warping.

OUTOMATIESE SPRAAKSEGMENTERING MET BEPERKTE DATA

Vir vinnige ontwikkeling van korpus-gebaseerde gesproketaalstelsels, soos aaneenskakelende spraaksintesestelsels in tale met beperkte hulpbronne, word ’n doeltreffende, konsekwente en akkurate wyse van fonetiese spraaksegmentering benodig. Die handmatige ontwikkeling van foneties geannoteerde korpusse is ’n uiters moeisame en tydsame proses wat dikwels mank gaan aan uitdagings met betrekking tot inkonsekwenthede en reproduseerbaarheid. Helaas is die outomatisering van hierdie annoteringsproses nog slegs gedemonstreer vir groot korpusse in ’n beperkte aantal tale waar tegnieke wat omvangryke en gespesialiseerde hulpbronne vereis, gebruik word.

In hierdie studie word die probleem van fonetiese segmentering binne die konteks van die ontwikkeling van klein, prototipiese spraaksintesekorpusse vir nuwe tale met beperkte hulpbronne onder die loep geneem. Dit word uitgevoer deur middel van ’n empiriese evaluering van bestaande segmenteringstegnieke op tipiese spraaksintesekorpusse van drie Suid-Afrikaanse tale. Die werkverrigting van hierdie tegnieke word onder verskillende data-omstandighede gekarakteriseer, terwyl doeltreffende toepassingsmoontlikhede vir hierdie tegnieke ook ondersoek word om die akkuraatheid en betroubaarheid van fonetiese belyningsuitslae te maksimeer.

Ons bevindings toon dat, selfs met uiters beperkte hoeveelhede data, die toepassing van basislynspreker-afhanklike versteekte Markovmodelle relatief betroubare en akkurate resultate oplewer. Die doeltreffende ontwikkeling en toepassing van sodanige modelle word ondersoek, en daar word aangetoon hoe dit tot aanvaarbare resultate kan lei. Ons bewys dat die gehalte van belynings voldoende is om die ontwikkeling van spraaksintesetoepassings te ondersteun en dat dit selfs ten opsigte van akkuraatheid vergelykbaar met handmatige belynings in die gegewe konteks is. Ten slotte word verdere moontlikhede vir die outomatiese verfyning van fonetiese belynings ondersoek, en ’n doeltreffende korpusontwikkelingstrategie (met voorstelle vir toekomstige navorsing) word voorgestel.

Sleutelwoorde: fonetiese spraaksegmentering, fonetiese belyning, spraaksintese, teks-na-spraak, spraakkorpusontwikkeling, hulpbron-arm tale, versteekte Markovmodelle, dinamiese tydverstelling.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE

INTRODUCTION

1.1 PROBLEM STATEMENT

Modern spoken language systems such as speech recognition and synthesis systems have become increasingly reliant on large corpora of annotated speech data in the form of audio recordings, the most significant of these annotations being the time aligned transcriptions identifying phonetic segments. Such transcriptions serve as an accurate indication of where individual *phones*, the acoustic realisation of *phonemes*, begin and end. This is important because phonemes are considered the smallest meaningful units of speech and thus most speech based systems need to process these basic units. It follows that corpus based systems rely on annotated corpora of this nature in order to construct acoustic definitions in some form or another. Examples of such definitions or representations vary from statistical models like Hidden Markov Models (HMMs), used extensively in automatic speech recognition (ASR), to phonetic catalogues employed by concatenative text-to-speech (TTS) systems.

Due to the reliance on phonetic definitions, the performance of most corpus based systems are directly dependent on the accuracy of phonetic transcriptions: When training statistical models, accurate transcriptions allow one to better initialise training procedures (such as the commonly used expectation maximisation (EM) algorithm) which leads to more successful models (Young *et al.*, 2005), while systems relying on more direct representations of acoustic units benefit even more significantly in terms of quality with more accurate transcriptions (Clark *et al.*, 2007). Such concatenative TTS models are the focus of the current research.

Unfortunately developing accurately annotated speech corpora is often a challenging task. Manual phonetic segmentation is an arduous and time consuming task requiring expert knowledge of the phonemic (and phonetic) constituents of the specific language, significant skill in order to correctly identify phonetic transitions and a high level of concentration to ensure consistent results. In addition

to the tedious and specialised nature thereof, the problem is exacerbated by the fact that boundaries between consecutive phones cannot always be unambiguously defined due to the phenomenon of co-articulation resulting in a gradual change in acoustic properties between adjacent phones. For these reasons, the manual development of high quality phonetically annotated corpora is a costly process usually involving groups of well trained individuals relying on well defined protocols. Even under ideal circumstances, manual segmentation still presents challenges pertaining to the consistency of transcriptions (Pitt *et al.*, 2005).

In contrast, automating phonetic segmentation promises fast, cost-effective and consistent results. This however comes at the cost of less accurate and occasionally erroneous results with the additional problem of not generalising well to all contexts including new language, voice or recording channel conditions. Thus simply applying a generic technique such as HMM-based ASR for the purposes of speech segmentation is not sufficient for the development of high quality corpora. Consequently most research into achieving quality automated segmentation has involved specialising generic techniques to suit specific language and speaker conditions amongst others (Toledano *et al.*, 2003). Despite advances in improving automatic segmentation accuracy, current solutions are often context specific and still require significant resources to begin with.

In most cases manual segmentation is prohibitively expensive and too time consuming for the rapid development of resources and systems in new scarcely resourced languages. This is especially true in the developing world, where there are many languages in dire need of spoken language technologies, while skills necessary to develop corpora and build systems are severely limited. Furthermore, current techniques cannot be indiscriminately applied to solve the problem of high quality automatic segmentation in this context, mainly because such techniques require the existence of systems and resources such as high quality speech recognition and speech synthesis systems as well as large, accurately (manually) segmented corpora for the application of machine learning techniques.

This prompts one to consider the problem of developing an automated, accurate, consistent and robust phonetic speech segmentation system suitable for the rapid development of small prototype corpora in new languages with minimal ideal resources. An overview of literature on the topic of phonetic segmentation (particularly in the context of developing TTS corpora) is thus presented here. Also of interest are methods for judging the quality of speech corpora with respect to phonetic alignments and techniques aiding in quality control of complete corpora with regards to alignment accuracy.

The next section presents an overview on the relevant literature pertaining to the general problems of phonetic segmentation and corpus construction and the subsequent and final section summarises and discusses the relevance of the current literature as well as elaborating on the specific research questions related to the context presented above.

1.2 LITERATURE REVIEW

In the following subsections we present different approaches to phonetic segmentation as well as techniques for high quality temporal alignment of phonetic boundaries and methods for ensuring corpus quality.

1.2.1 PHONETIC SEGMENTATION

Given an audio recording of speech in a certain language, the task of phonetic annotation can be regarded as the combination of two sub-tasks (Paulo and Oliveira, 2004):

1. Determining the underlying phonemic sequence, and
2. Obtaining the temporal locations representing boundaries between consecutive phones.

The first of the sub-tasks is fundamentally a speech recognition problem, while the latter concerns segmentation and temporal phonetic alignment. The topic of phonetic segmentation in the context of developing corpora for systems such as TTS is primarily concerned with the second task. Approaches to the problem of segmentation can usefully be categorised into two classes, namely text-dependent and text-independent segmentation.

1.2.1.1 TEXT-INDEPENDENT SEGMENTATION

Text-independent segmentation, also called unsupervised segmentation, attempts to partition speech signals without any linguistic knowledge (e.g. word or phonemic sequence). Thus from the perspective of phonetic annotation, the problem of explicitly determining the underlying phonemic sequence is abandoned in favour of an analysis based purely on the speech signal properties. Methods of this kind usually make the assumption that phonetic boundaries exhibit local changes in the signal or that phonetic segments are in some way coherent with regards to signal properties. Based on these assumptions text-independent segmentation can be done by boundary detection based on acoustic cues or by defining boundaries between segments identified by the application of unsupervised machine learning algorithms such as clustering (Andre-Obrecht, 1988; Šarić and Turajlić, 1995; Sharma and Mammone, 1996; Estevan *et al.*, 2007; Almpanidis and Kotropoulos, 2008; Golipour and O'Shaughnessey, 2007).

Due to the application of such methods without prior information on the number of or nature of segments represented in the signal, resulting boundary candidates are not guaranteed to be in agreement with meaningful phonetic boundaries. Thus although these approaches cannot be exclusively employed for the development of phonetically annotated TTS corpora, they present worthwhile options when refining boundaries obtained by text-dependent techniques (refer to Section 1.2.2 below).

1.2.1.2 TEXT-DEPENDENT SEGMENTATION

When developing speech corpora for TTS, some form of linguistic knowledge is invariably available. This is often in the form of orthographic transcriptions which can then be used to predict the phonetic sequence via a pronunciation lexicon or letter-to-sound rules (also known as grapheme-to-phoneme rules) which has to be developed as part of TTS systems. Speech data to be segmented is generally acquired through the recording of readings of carefully selected text in a controlled way in order to minimise mismatches between the audio and orthographic transcriptions. This allows for the application of a linguistically constrained process namely text-dependent segmentation.

Two methods based on dynamic programming algorithms have commonly been applied to provide estimates of boundary locations between consecutive phones (Adell *et al.*, 2005):

1. Dynamic Time Warping (DTW) of the target signal to match a signal with the same underlying phonetic sequence of which the phonetic boundaries are known, and

2. Hidden Markov Models (HMMs) applied in forced alignment via the Viterbi algorithm.

The first method has its origins in the early days of speech recognition where word recognition was performed by pattern matching (Ney, 1984). It relies on the existence of a signal with acoustic properties similar to the signal to be segmented of which the phone boundaries are known. In the TTS application domain, such a signal can sometimes be generated relatively easily by synthesizing a waveform from the available transcriptions. This should ideally produce boundary placements that are highly appropriate for the purposes of TTS and as a result, this method is often employed in this context (Malfrère and Dutoit, 1997).

The second method essentially uses an HMM-based phone recogniser applied to forced alignment, meaning that the language model is constrained by the available transcriptions. This method requires the training of HMMs for each phoneme that occurs in the specific language (or dialect) and subsequently using these models to find the most likely locations and durations of segments according to these models. This results in approximate phonetic boundaries between segments.

Comparing the requirements of these two methods, the application of DTW requires the existence of an appropriate speech synthesiser while the HMM based technique is model based and is thus dependent on the availability of training data. Studies have shown DTW based segmentation outperforms the HMM based technique with regards to fine accuracy, but falls short in comparison when considering robustness, with more gross segmentation errors attributed to the DTW technique (Kominek *et al.*, 2003; Malfrère *et al.*, 2003). No comparisons can be found between these techniques when applied to small speech corpora, although the DTW method has been applied towards the building of a speech synthesiser with limited resources (Louw *et al.*, 2006).

1.2.2 SEGMENT BOUNDARY REFINEMENT

A popular approach to achieving highly accurate segmentation is to imitate the expert human transcriber's procedure (Toledano *et al.*, 1998). The procedure whereby human experts perform phonetic

segmentation can be viewed as a two-stage process, where the transcriber initially identifies segments based on the acoustic properties (aided by visual representations thereof) and subsequently refines boundary placements between contiguous segments by considering sets of consistent acoustic cues based on the transition context (usually determined by broad phonetic classes). The application of HMMs to phonetic segmentation can be likened to the first stage of this procedure, which basically entails the recognition of segments without explicitly considering optimal boundaries. Thus, a large amount of research has been done on further reducing the discrepancies between HMM based and manually obtained boundaries (i.e. "boundary refinement") (Toledano *et al.*, 2003, 1998; Sethy and Narayanan, 2002; Kim and Conkie, 2002; Saito, 1998; Lo and Wang, 2007; Park *et al.*, 2006; Jarifi *et al.*, 2008).

Numerous techniques from the pattern recognition and machine learning fields have been applied to improve the accuracy of alignments resulting from either DTW or HMM-based techniques. This typically involves:

- Using any of the range of text-independent techniques mentioned earlier in a complementary fashion (e.g. using initial alignments to limit the scope of these methods to the refinement of existing boundaries) (Saito, 1998),
- Applying models which are trained to explicitly identify boundaries based on training examples of human placements (Sethy and Narayanan, 2002; Lo and Wang, 2007), or
- Combining or selecting boundary estimates in various ways obtained from different sources in order to increase accuracy in different phonetic contexts (Park *et al.*, 2006; Jarifi *et al.*, 2008).

This has proved successful, with researchers reaching levels of accuracy comparable to discrepancies between independently verified alignments by experts (Toledano *et al.*, 2003).

1.2.3 CORPUS QUALITY CONTROL

Another important avenue of research regarding automatic segmentation concerns the definition of confidence measures and other means of ensuring the accuracy and consistency of alignments. This is essential from the perspective of corpus and system development, because such techniques can elucidate problems efficiently and serve as a very specific indication of where some manual supervision might be needed. Attempts to ensure corpus quality have involved:

- Detecting erroneous alignments by flagging individual segments which are statistical outliers when considering specific properties. Segment properties which have been considered include spectral consistency (Barnard and Davel, 2006) and the mean duration of specific phone classes (Kominek *et al.*, 2003).
- The definition of confidence measures, usually derived from information generated by the segmentation process itself (e.g. a DTW or HMM-based technique) (Paulo and Oliveira, 2004).

The improvement in quality of systems and alignment accuracy reported by researchers employing the above methods, make the establishment of similar techniques as a standard part of corpus development a sensible proposition.

1.3 SCOPE OF RESEARCH

It is clear that the problem of automatic phonetic segmentation has been considered in a number of different contexts. In the overview presented, two particular contexts feature extensively:

- Segmentation of large general purpose corpora, requiring speaker independent segmentation, and
- Rapid segmentation of large speaker specific TTS corpora.

With the exception of a few cases, the majority of research has been on achieving accuracy of segmentation, specifically when comparing to manual alignments, at all costs. Towards achieving these goals, the following fundamental resources have been applied:

- Speaker independent speech recognition systems,
- High quality speech synthesis systems, and
- Large, accurately (manually) aligned speech corpora.

Due to the demanding requirements in basic resources, it follows that these techniques have only been applied to a select few languages. These are languages which already possess significant resources. This limits the applicability of these results and raises questions on the feasibility of such techniques in widespread, practical and especially unfavourable scenarios.

While more efficient techniques, such as the application of speaker specific speech recognition, are widely used towards constructing TTS systems, the subject has not received much attention in its own right. Thus questions regarding the following points are still unanswered:

- Appropriateness of the segmentation and refinement techniques discussed here considering data scarcity.
- The sensitivity of these techniques to the size of the speech corpus.
- The considerations when applying various segmentation techniques in different language and speaker conditions.
- Ensuring accurate and consistent results efficiently.
- Ensuring quality output while minimising manual intervention.

The aim of this work is thus to characterise the feasibility of current methods in a scarce resourced scenario, from the selection of an appropriate segmentation technique to considering efficient refinement methods and effective means of ensuring the quality of resulting corpora. This will be done within the context of developing small prototype TTS corpora for a number of local South African languages.

In the following chapter typical speech corpora to be used in this study are introduced as well as a discussion on general methodology and how phonetic alignments can be evaluated. Chapter 3 is concerned with the establishment of a baseline segmentation system based on an evaluation and comparison of the feasibility of the predominant text-dependent techniques mentioned. In Chapter 4 the refinement of the segmentation process is considered. Chapter 5 presents work on further refinement of boundary placements and the final chapter aims to make suggestions on the options regarding the implementation of a segmentation system and development of annotated speech corpora for the purpose of developing spoken language systems under the circumstances based on the results obtained in this study.

CHAPTER TWO

APPROACH

In order to present a quantitative analysis of speech segmentation techniques in a "scarce resourced" scenario, it is necessary to define measures of success and clarify the notion of "scarce resourced" by characterising the context of the the problem. The following sections introduce measures for evaluating results as well as motivations for using each of these and to characterise typical data sets that are considered representative and which will form part of this work, as well as to discuss the high level design of a segmentation system adopted here and the methodology to be followed throughout.

2.1 MEASURES OF SUCCESS

The evaluation of phonetic alignments generally fall into one of two broad categories:

1. Objective evaluation, which involves comparison between a set of alignments and an ideal reference set (usually manually obtained). This method of evaluation is system-independent and is relatively cost-effective.

2. Subjective evaluation, involving the incorporation of the resulting alignments into an end-system (such as a TTS system) and evaluating the results as part of the system, usually through some form of perceptual evaluation of the resulting speech in the case of TTS.

The most widely used measure of success employs the first approach by means of comparison with manual alignments which serve as the definitive case. This is justified by the observation that manual segmentation generally represents the most accurate solution, generalising well to new languages and speaker idiosyncrasies. One reason for the popularity of this method of evaluation is its relative cost-effectiveness when manually segmented corpora are available. However, results obtained in this way are not entirely deterministic as manual segmentation does not yield completely

consistent results mainly due to ambiguities in the definition of phone boundaries (as a result of co-articulation) and consequently comparisons suffer from similar inconsistencies. Studies comparing independent manual segmentation by experts have found discrepancies to range from 93 to 97% in terms of boundary placements in agreement at a 20ms tolerance level (Adell *et al.*, 2005; Toledano *et al.*, 2003). These levels of inter-transcriber variability where expert transcribers are involved can be considered a measure of the inherent ambiguity in the process and as such should represent an upper limit of what can be expected when comparing techniques with manual alignments.

Other researchers have criticised this form of evaluation from the perspective of system building (especially concatenative TTS), arguing that consistency and reproducibility is as important as accuracy and pointing out that on these grounds manual alignments cannot be considered to be inherently superior to automatic alignments or definitive in nature considering the target application (Clark *et al.*, 2007). From this perspective, alignments should be evaluated in the context of an end-system, e.g. perceptual experiments in the case of developing TTS corpora. While a number of perceptual experiments have reported that manually corrected segmentation results in more intelligible and natural sounding speech synthesis compared to baseline automated methods (Saito, 1998; Adell *et al.*, 2005), some researchers have found that automatic procedures can yield superior results in this context (Makashay *et al.*, 2000). This suggests that manual alignments can be useful as an initial benchmark, but should not necessarily be considered optimal or definitive.

In the following two sections, methods for judging alignments based on comparison with manual results as well as perceptual experiments are briefly presented and discussed.

2.1.1 MEASURING AGREEMENT BETWEEN ALIGNMENTS

The most prevalent method of comparing alignments between two different sources involves accepting individual boundary placements to be in agreement if they fall within a certain time threshold of one another. By accumulating boundaries in agreement, one can express this as a fraction of all boundaries to obtain what is termed the “boundary accuracy”. Boundary accuracies are most often reported for a range of thresholds from 5 to 50ms, with 20ms being the most often cited. Some authors also combine these accuracies to form a single mean accuracy (Adell *et al.*, 2005).

Another compact way of representing the discrepancies between reference and automatically obtained alignments is by simply taking the square-root of the mean squared time differences between all boundaries - i.e. the Root Mean Square Error (RMSE), which results in an indication of the mean difference represented in the units of measurement (e.g. milliseconds).

While the boundary accuracy and RMSE can serve as a rough guide regarding success on the corpus level, it is less appropriate when considering specific boundary contexts and gauging how well phonetic segments are identified. Differences in durations of various phones and phone classes necessitate a duration independent measure. Such a measure is proposed by (Paulo and Oliveira, 2004) and serves to determine the overlap of segments in proportion to the segment durations. This is termed the “overlap rate” (OR).

Briefly, the overlap rate is given by:

$$OR = \frac{D_{com}}{D_{max}} \tag{2.1}$$

$$= \frac{D_{com}}{D_{ref} + D_{auto} - D_{com}} \tag{2.2}$$

where D_{com}, D_{max}, D_{ref} and D_{auto} are the *common*, *maximum*, *reference* and *automatic* durations respectively (see Figure 2.1).

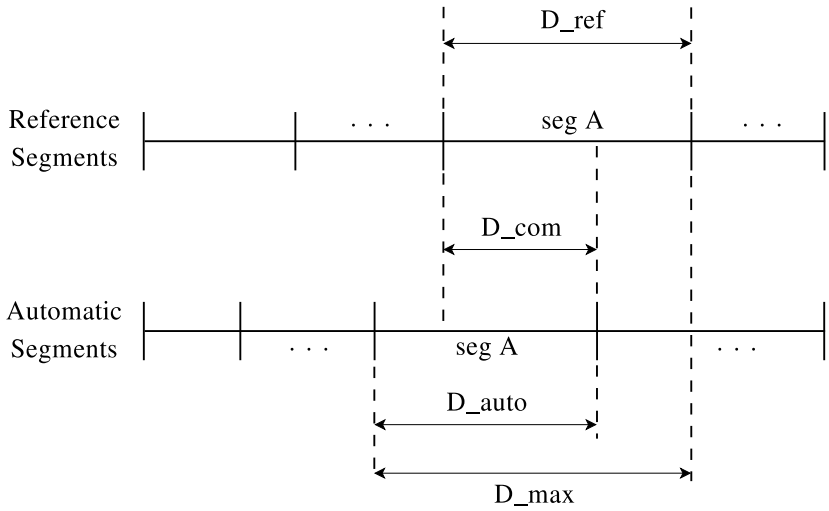

Figure 2.1: *"Overlap rate" definition (Paulo and Oliveira, 2004).*

It is important to note here that the phonetic sequence is known and as such, it is known exactly which reference segment to compare with a particular automatic segment. Thus even when no overlap occurs or when multiple segments overlap with incorrect reference or automatic segments, this merely results in $D_{com} = 0$ and thus $OR = 0$. This results in a measure of comparison which gives an effective indication of the relative importance of accuracy for segments of various durations.

2.1.2 PERCEPTUAL EXPERIMENTS

Any of the number of experiments designed to evaluate TTS systems can also be employed to establish the effects of phonetic alignments on the output of these systems and thus evaluate alignment procedures implicitly. Perceptual experiments are generally either designed to gauge the intelligibility of the speech output (e.g. the Diagnostic Rhyme Test, Modified Rhyme Test or Semantically Unpredictable Sentences approaches) where the user is required to recognise synthesised words, or are broad preference tests where two similar speech output signals are compared or scores are assigned to speech output samples (e.g. Mean Opinion Scores).

Specifically towards evaluating alignments, preference tests have been widely employed. This includes direct comparison between samples (Adell *et al.*, 2005; Kawai and Toda, 2004; Kim and

Conkie, 2002; Kominek and Black, 2004) and MOS scores (Jarifi *et al.*, 2008; Kominek and Black, 2004; Makashay *et al.*, 2000).

2.2 REFERENCE CORPORA

The widely spoken languages of the world have understandably received a lot of attention from language technology specialists developing large corpora of accurately annotated speech data, enabling high quality and optimised language technologies such as ASR and TTS. When considering the automation of corpus development for any of these languages, one has many resources to call upon in aid of such processes. However, when developing corpora and systems for new languages (especially highly dissimilar languages, such as languages of African origin), there are no analogous resources to build upon. Furthermore, skills shortages and economic viability of the lesser spoken languages hamper any prospects of developing large, high quality, manually constructed solutions. For similar reasons, towards the construction of systems such as TTS, speech corpora have often been minimally designed (Louw *et al.*, 2006). Speech segmentation is investigated within this scenario here.

Three sets of speech recordings used in the construction of prototypical TTS systems in South African languages are employed. These data sets represent minimally designed single speaker speech corpora, where text is selected carefully in order to cover all the appropriate phonetic constituents (diphones) of each language. The languages represented constitute three of South Africa's eleven official languages and importantly come from distinct family groups (see Table 2.1 for specific details of each corpus). As the majority of the country's official languages belong to one of these groups, it is hoped that the results will be highly relevant.

The corpora listed here were developed by manually correcting phonetic alignments based on baseline text-dependent techniques as these baseline techniques did not result in sufficiently accurate alignments to support intelligible concatenative synthesis systems. This work was largely performed by inexperienced transcribers with limited training, the initial Afrikaans and isiZulu alignments based on the DTW technique implemented in the *Festvox* software package (Black and Lenzo, 2007), while the initial pass for the Setswana set was obtained from a baseline HMM-based forced alignment procedure implemented using the *HTK* (Young *et al.*, 2005) package.

Language	Lang. group	Gender	Utterances	Duration	Phones
Afrikaans	Germanic	Male	134	21 mins.	12341
isiZulu	Nguni	Male	150	20 mins.	8559
Setswana	Sotho	Female	332	46 mins.	26010

Table 2.1: *Properties of the reference data sets.*

2.2.1 LEVEL OF CONFIDENCE

Although the comparison of alignments with manually obtained reference alignments are ideally a very convenient and relatively simple solution to obtaining an indication of alignment accuracy, it

does in practice present some challenges (especially in a context where skilled individuals required for the manual alignment of new corpora are not available).

Due to the limited level of experience and expertise involved in manual checking of the local corpora identified above, it can be expected that the consistency and accuracy of the reference alignments be somewhat less ideal than generally encountered in studies employing expert transcribers.

In order to quantify the level of confidence in the data presented from the perspective of the measures of comparison between alignments to be used in this work, we set up an experiment to measure inter-transcriber variability in this context.

Firstly we select a subset of utterances from each corpus, ensuring coverage of all the distinct phones present in each language (see Table 2.2).

Language	Lang. group	Gender	Utterances	Duration	Phones
Afrikaans	Germanic	Male	20	213 sec.	2125
isiZulu	Nguni	Male	20	158 sec.	1143
Setswana	Sotho	Female	10	185 sec.	1879

Table 2.2: *Properties of the subsets used to determine inter-transcriber variability.*

Each of these subsets are subsequently manually aligned independently of the reference data by correcting alignments generated by a baseline HMM-based forced alignment procedure. The alignments obtained in this way are directly compared to the reference alignments for the particular subsets, using the measures of comparison introduced in Section 2.1.1. Table 2.3 represents the results obtained in terms of both the boundary accuracy and OR.

Language	Boundary comparisons				OR	
	< 5ms	< 10ms	< 20ms	RMSE	μ	σ
Afrikaans	54.58%	73.35%	88.84%	16.40ms	79.41%	18.90
isiZulu	49.33%	74.35%	89.49%	17.62ms	81.16%	17.82
Setswana	58.05%	77.85%	90.64%	17.36ms	82.18%	16.54

Table 2.3: *Inter-transcriber agreement statistics.*

The boundary agreement rates, root mean square error (RMSE) between boundaries, as well as the mean OR values obtained are largely comparable between the three corpora, with the agreement on the Setswana corpus being slightly higher than the other two languages. Boundary comparison results can be directly compared with similar results from other studies mentioned above, as expected, the inter-transcriber discrepancies are somewhat higher than cases reported for expert transcribers. This is especially the case in the lower tolerance ranges (e.g. < 5ms where figures close to 70% have been measured). These higher levels of discrepancies might be attributed to inexperienced transcribers producing both inherently less accurate as well as more variable results and represent an approximate ceiling for reliably benchmarking alignments based on these reference sets.

Also of importance is considering the ORs per phone type, Figure 2.2 plots the mean OR (with error bars showing the standard deviations) for each phone type in each language (corresponding

International Phonetic Alphabet (IPA) representations (Ladefoged, 1990) for each phone can be found in Appendix B).

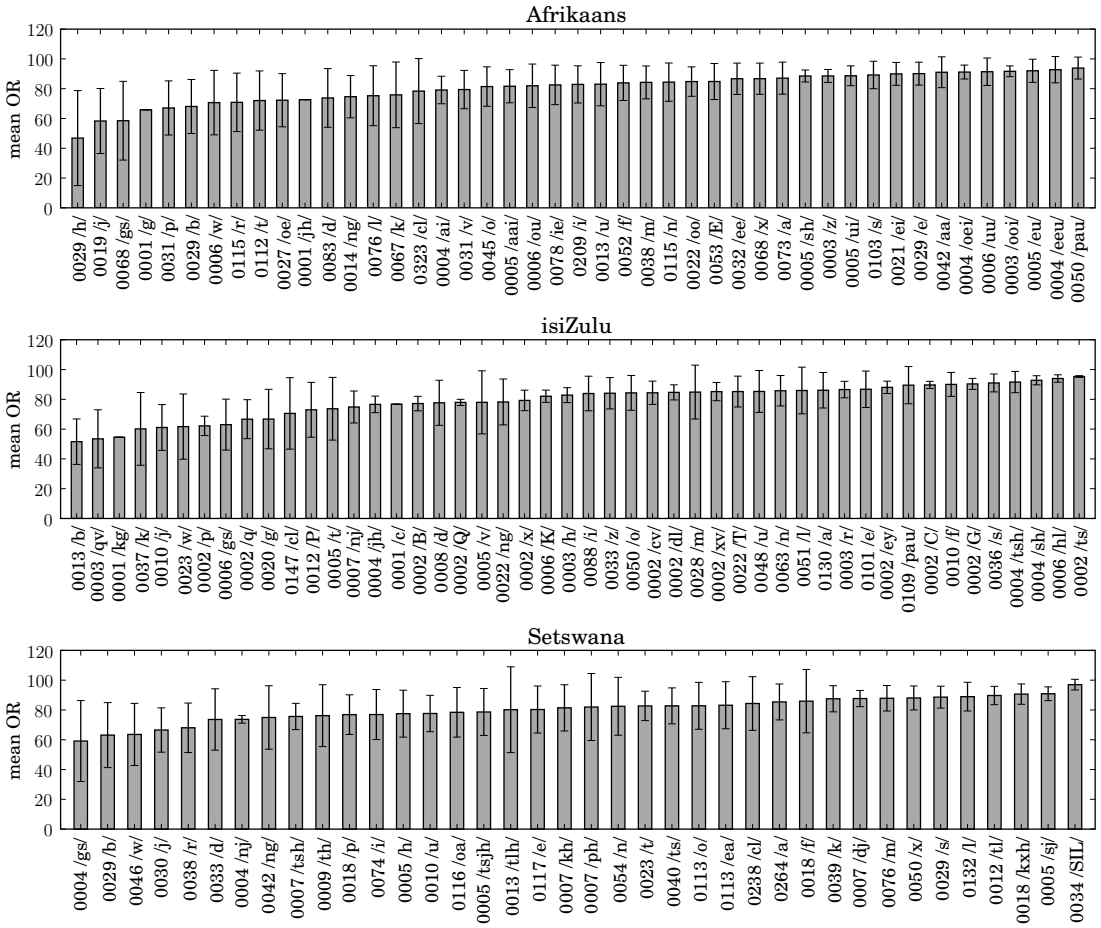

Figure 2.2: *Mean OR per phone between independent manual transcribers for each language. Error bars represent the standard deviations and values accompanying each phone label indicate the number of occurrences in the subset used.*

It seems, in addition to shorter phones having a lower mean OR (probably due to small errors distributed around each boundary placement having a greater effect on the OR of such segments), that some phone classes consistently presented greater difficulty to manual transcribers (e.g. the approximants /j/ and /w/).

2.3 TWO-STAGE SEGMENTATION

In Chapter 1 a number of approaches toward the segmentation problem was presented. Towards annotated corpus development for the purpose of building systems, most approaches rely on a first stage of text-dependent segmentation (using one of the techniques mentioned), followed by a local refinement stage aimed at explicitly determining phone boundary placements (a local refinement strategy similar to one an expert human transcriber would follow). Thus most accurate results are obtained from a

process with high level design depicted in Figure 2.3.

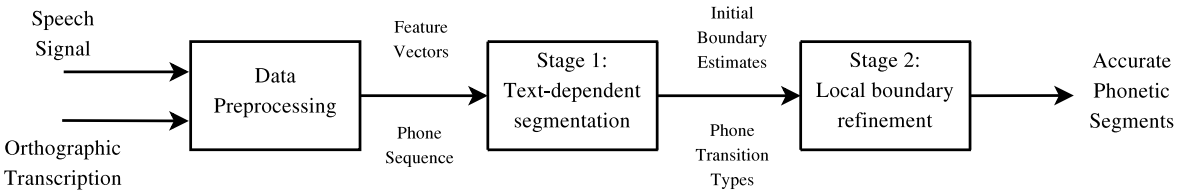

Figure 2.3: *Two-stage design toward accurate automated segmentation.*

This approach is adopted in this work and investigated in the stated context with the aim of understanding the appropriateness of techniques described in Chapter 1 and presenting quantitative results analysing specific possibilities of applying these with limited data. The focus is on improving results in a relatively language independent way. This will entail the design, implementation and application of a segmentation system and the systematic analysis of the constituent components.

We will start with the establishment of a baseline text-dependent process and proceed thereafter to investigate refinements to such a system based on analyses of the performance on the data presented in this chapter in order to implement an accurate and efficient system. We also investigate the feasibility of a refinement stage and discuss methods for ensuring quality corpora whilst minimising manual interaction (especially with regards to TTS applications).

CHAPTER THREE

ESTABLISHING A BASELINE TEXT-DEPENDENT SEGMENTATION SYSTEM

The first problem encountered during phonetic annotation of a speech signal involves speech recognition i.e. obtaining the underlying symbolic representation. General purpose speech recognisers are designed to recognise any valid spoken form in a particular language. This is often done by constructing statistical or other models representing the language structure or grammar. In this way the most likely underlying sequence of symbols can be obtained by matching the acoustic observations with pre-existing representations thereof in conjunction with the language model.

When developing speech corpora towards TTS system construction, it is customary to use read speech from carefully designed text and recordings. With this additional linguistic information in the form of the utterance orthography, the text processing front-end of the speech synthesis system is usually used to predict the corresponding phonetic sequence via pronunciation dictionary lookup or letter-to-sound (also known as grapheme-to-phoneme) rules. Assuming that this process of careful recordings and phone prediction yields consistent and accurate phonetic sequences reflected in the speech, one can effectively constrain the language model of general purpose speech recognisers to the known symbolic sequence, in effect reducing the problem of recognition to the assignment of segments of speech to a specific phone. This results in a temporal segmentation of the speech given the phone sequence, according to the recogniser. This process is referred to here as text-dependent segmentation.

3.1 BACKGROUND

Two approaches based on speech recognition techniques have been successfully applied to text-dependent segmentation:

1. Dynamic time warping matches a template signal with identical phone sequence with known phone durations to the input signal, taking into account variation in time in order to map the phone boundaries in the template signal to time locations in the input.

2. Viterbi forced-alignment aligns the input speech signal with a Hidden Markov Model representing the correct phonetic sequence.

The following sections discuss the applications of these two approaches specifically with regards to segmenting speaker-dependent TTS corpora.

3.1.1 DYNAMIC TIME WARPING

The use of DTW to perform automatic phonetic segmentation of a single speaker TTS corpus was first advocated by (Malfrère and Dutoit, 1997). The idea is that an existing synthesiser is used to not only predict the pronunciation, but also synthesise the template signal and that this be aligned to the input signal. This was shown to be successful despite potential mismatches in the qualities of the specific voice used (e.g. gender). DTW has found widespread use in this area of application, as it is a relatively simple, fast and convenient solution when building TTS systems.

The alignment process, once an appropriate template signal has been generated, involves parameterising both signals into sequences of feature vectors from frames with fixed window and step sizes. A dynamic programming algorithm then efficiently finds the path with minimum accumulated distance through a matrix representing the distances (the Euclidean distance is usually used) between each of these vector sequences. This path is used to map times corresponding to phone transitions in the reference signal to times in the signal to be segmented. Figure 3.8 shows an example of such a distance matrix, path and mapping.

3.1.2 HMM-BASED VITERBI FORCED-ALIGNMENT

Segmentation using Viterbi forced-alignment firstly requires the training of acoustic models in the form of HMMs, modelling each phoneme in the language individually. This has been attempted in a number of ways (including the application of speaker-independent models and speaker-independent models that have undergone speaker adaptation). An approach that has become common when developing TTS corpora simply involves training a speaker specific set of models on the same data to be segmented (Clark *et al.*, 2007).

Once the models are estimated, the alignment procedure involves applying the Viterbi algorithm by constructing a complete "composite HMM" from the provided phonetic sequence by concatenating single HMM models and finding the optimal path through this model given the parameterised speech vector sequence, i.e. maximising the likelihood of the model given the observations by the assignment of feature vectors to specific phone models (and states within each model).

3.2 CHOOSING A SUITABLE BASELINE SYSTEM

These two approaches are based on similar dynamic programming algorithms, with the difference being the reference representation which is either a relevant synthetic speech signal, or a model describing the phonetic sequence. Thus an important factor determining the accuracy of these techniques involves the construction of these reference representations or templates.

With DTW the template signal is constructed independently of the input data, which eliminates the concern of appropriate amounts of data to estimate models, but compromises the template signal in terms of acoustic relevance which might cause degradation in the accuracy of alignment. Furthermore, because the phonetic alignments are mapped from phonetic boundaries obtained from a synthesiser, this should presumably result in alignments which are similar in nature and thus highly relevant when building new synthesisers.

In contrast, the HMM-based approach, by training models from the data itself, can potentially better represent the acoustic realisations of phones provided that the training procedure is successful and the data is of sufficient quantity to train accurate models. These models represent the acoustic properties of individual phones and the alignment process places boundaries based on the interaction between the likelihoods of consecutive models (more precisely model states) given the observed feature vectors (see Section 3.1.2). This does not necessarily result in the most appropriate alignments.

Some existing studies comparing these two approaches have varied conclusions based on their specific contexts. (Kominek *et al.*, 2003) found that the majority (in excess of 70%) of DTW based segments were more accurate than the corresponding HMM based alignments, but that this system was more prone to gross errors in comparison. This comparison was however done by aligning an American English corpus by using a synthetic diphone synthesiser as reference, where the same speaker was the voice talent for the system and speaker in the corpus, which clearly mitigates the main disadvantages of using the DTW approach mentioned above. In contrast, (Adell *et al.*, 2005) found the HMM-based approach to outperform DTW conclusively. Other researchers have attempted to combine the strengths of these approaches in order to improve alignment accuracy in general (Paulo and Oliveira, 2004). An extensive comparison by (Malfrère *et al.*, 2003) on a number of European languages reports comparable results in terms of accuracy and also suggests using DTW as a bootstrapping stage prior to HMM-based alignment.

In the context of rapid development of prototype TTS systems in new languages, i.e. developing relatively small first-time corpora, the above techniques have to be applied under non-ideal circumstances. In this chapter the aim is to understand the relative merits and difficulties associated with the application of these two techniques in their basic forms (i.e. without considering optimal parameters for better performance) under these conditions. Specific questions and concerns that will be investigated are:

- The implications of using an English synthesiser to synthesise template signals for different languages on DTW performance,

- The implications of training and applying HMMs from minimally designed corpora where some phone occurrences are extremely limited, and
- General suitability of these approaches with respect to accuracy, robustness and practical implementability in the given context.

3.3 EXPERIMENTAL SETUP

For the purpose of comparison, two alignment systems based on the above-mentioned approaches are set up to produce alignments which are identical in phone sequence to the manually checked transcriptions which are part of the corpora described in Section 2.2. The implementation and setup of these systems are described below.

3.3.1 ALIGNMENT SYSTEMS

3.3.1.1 HMM-BASED PHONE RECOGNITION SYSTEM

For the HMM-based alignment system a simple phone recognition system based on the *HTK* software package was implemented by adapting a generic training strategy suggested in (Young *et al.*, 2005). For the purposes of this comparison, standard parameters judged appropriate for a baseline speech recognition system are used (Gouws *et al.*, 2004).

Briefly, this involves using Mel Frequency Cepstral Coefficients (MFCCs) as feature vectors with 12 coefficients, energy and the first and second order derivatives of these (39 coefficients in total). Feature vectors are calculated for Hamming windowed speech frames with length 20ms extracted at 10ms intervals. This is used to train tied-state, context-dependent HMMs consisting of three states with a standard "left-to-right" topology and a single mixture Gaussian Mixture Model (GMM) representing the state emission distributions from a "flat start" initialisation. Figure 3.1 depicts the basic system design and Table 3.1 summarises the parameters used.

Features	
Type	MFCCs (39 coefficients)
Window function	Hamming
Window size	20ms
Step size	10ms
Models	
Initialisation	Flat start
Topology	3-state left-to-right
State distributions	1 mixture GMM
Context-dependence	Tied-state triphones

Table 3.1: *Summary of parameters used in the HMM-based system.*

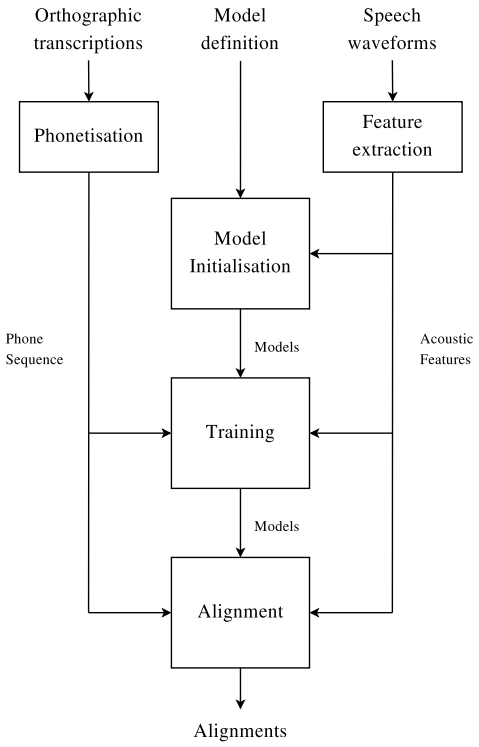

Figure 3.1: *Basic HMM-based alignment system.*

3.3.1.2 TTS-DRIVEN DTW ALIGNMENT SYSTEM

The DTW procedure used is based on the implementation in the *Festvox* software toolkit (Black and Lenzo, 2007) which is based on the freely available *Festival* synthesis software (Taylor *et al.*, 1998). The signals are compared by calculating the Euclidean distance between frames of feature vectors extracted for the input and reference signals. Feature extraction used the default parameter values used in Festvox i.e. MFCCs with 12 coefficients and the first order derivatives (24 coefficients) calculated from Hamming windowed frames with length 25ms and 5ms frame shift. The synthetic signal was generated by the standard "KAL" diphone voice (based on a male speaker of American English) which is distributed with *Festival*. Figure 3.2 shows the basic design of the DTW-based alignment process.

3.3.2 DATA PREPARATION

The formulation of this experiment prompted a number of questions on how the reference signal should be synthesised for DTW. The following points were considered:

- The possibility of using different voices to attempt matching the acoustics of the different voices in the different corpora (e.g. possibly matching at least the speaker's gender).
- The manual phonetic transcriptions includes indications of "closures" (i.e. segments containing little energy encountered before plosive consonants). These are not generally labelled automat-

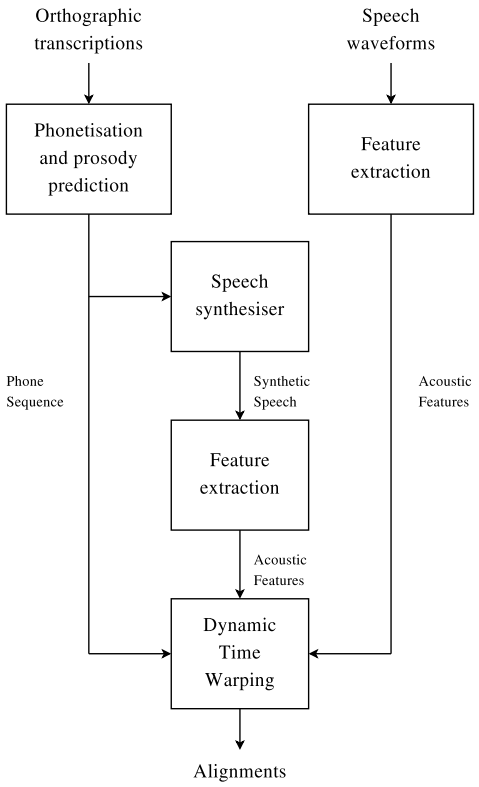

Figure 3.2: *Basic DTW-based alignment system.*

ically with DTW.

- The synthesis system can be set up to synthesise from the orthographic level or the phone sequence directly. Synthesising from the orthographic level allows a more complete analysis of the text towards applying prosodic information such as phone durations and pitch. This is not necessarily appropriate when synthesising a reference signal for alignment with a different language.

- Phone mapping between native phone sets and American English phone set used in the CMU Pronouncing Dictionary (Weide, 1998).

The focus of this chapter is not on optimising the techniques in question in order to obtain the best possible results, but rather on assessing the baseline performance. It is however important to obtain results which are at least representative of what can be expected in general.

Based on observations in (Malfrère and Dutoit, 1997), it is possible to achieve good results despite some level of mismatch between the reference and input voice qualities, however it was shown that segmentation with a cross-gender mismatch does degrade performance. In order to keep experimental parameter complexity to a minimum (e.g. keeping to the same synthesis techniques amongst other things), it was decided to align all corpora with the most stable, default male voice packaged with *Festival*.

To determine whether including closures in the segmentation procedure would result in a reasonable comparison, subsets of the corpora were aligned both including and excluding closure labels. These labels were removed by merging them into the subsequent plosive consonant. When the resulting segments were compared to manual segments, it was found that segmenting with closure labels resulted in significantly lower accuracy. This suggested that the baseline methods should be benchmarked without considering closures. The statistics for the reference data sets are thus slightly altered in terms of the number of segments (see Table 3.2).

Language	Lang. group	Gender	Utterances	Duration	Phones
Afrikaans	Germanic	Male	134	21 mins.	10028
isiZulu	Nguni	Male	150	20 mins.	7403
Setswana	Sotho	Female	332	46 mins.	22266

Table 3.2: *Properties of the reference data sets without "closure" and "glottal stop" segments.*

A similar test was done in order to determine whether the reference signal should be synthesised with basic English-based prosody (from the orthography) or simply using a sequence of phones with identical durations and flat pitch contour. Here the difference was less pronounced, but the synthetic signal from orthography did yield a slight improvement in alignment accuracy.

Phone mappings were developed based on perceived acoustic similarity, loosely motivated by place of articulation according to IPA phonetic definitions (e.g. click consonants from isiZulu were mapped to similar sounding plosive consonants and the velar fricative in Afrikaans and Setswana was mapped to the labiodental fricative in English). After attempting synthesis, further mappings had to be made as a result of the synthesiser not being able to render certain sequences of phones (due to missing diphone units). A description of the mappings and motivations can be found in Appendix B.

3.4 RESULTS

A comparison of the two baseline systems is achieved by using measures of agreement with the reference alignments (refer to Section 2.1). Results are first presented on the corpus level, followed by more detail for specific contexts.

3.4.1 BOUNDARY ACCURACY

The alignments resulting from each system are firstly viewed from the perspective of boundary placement. In Figure 3.3 a plot of the boundary accuracy values for each system compared to the reference segments are presented over a range of thresholds. From the information present in this figure one can assess the relative ability of each system to place phonetic boundaries within a small region around the reference boundaries (i.e. fine placements or accuracy) as well as an estimate of the nature of large discrepancies between automatic and reference alignments (i.e. gross errors). It is clear in this case that the alignments resulting from the HMM-based procedure are consistently closer to the reference

alignments, meaning that the this system results in both more accurate fine placements and fewer gross misplacements compared to the DTW technique.

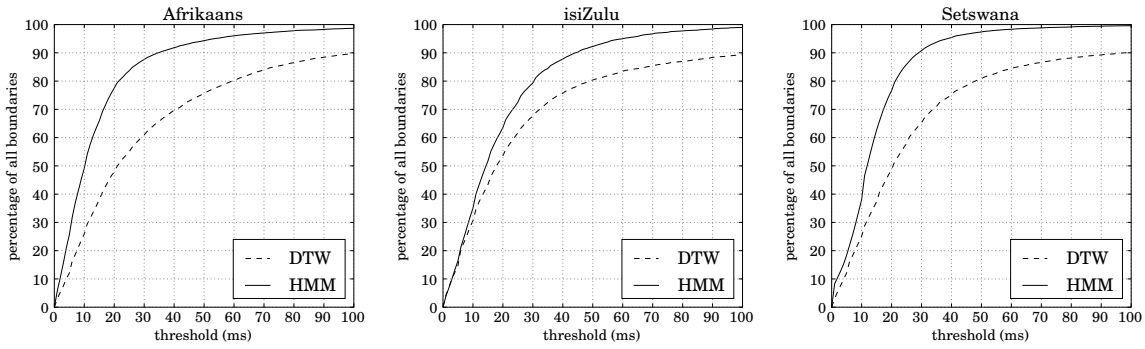

Figure 3.3: *A comparison of boundaries in agreement with the reference sets for a range of thresholds.*

The nature of boundary errors can be visualised by plotting histograms of automatically placed boundaries relative to reference boundary locations (see Figure 3.4). Negative differences represent placements where the automated alignments occur before reference alignments. Interestingly, in each plot one can observe peaks at regular intervals relative to reference location. This is a result of each set of reference alignments containing a proportion of boundary placements that was not modified from their original placements (based on automated methods) during manual correction (refer to Section 2.2). While these peaks mostly follow the same distribution as the remaining differences (e.g. in the case of Setswana aligned by the DTW system), there seem to be some clear biases toward the technique which was responsible for initial automated alignments. In the case of the Setswana HMM-based alignments and isiZulu alignments based on the DTW process, there are clearly higher peaks around the central point than would have been expected considering the remaining distribution of differences. Nevertheless, it should be evident from these plots that the HMM-based process generally results in boundary placements closer to carefully considered manual placements. The observation that both systems tend to place boundaries too early are in line with conclusions by (Kominek *et al.*, 2003), however while boundary placement discrepancies resulting from the DTW system are largely normally distributed, the HMM-based distributions tend to be skewed toward early placement.

Table 3.3 summarises the results in terms of boundary placements. Although results for each system are mostly comparable between the corpora, it is interesting to note that the DTW procedure performed relatively better on isiZulu while the HMM-based system achieved lower accuracy levels for the same language. Despite the fact that these results are not directly comparable with the inter-transcriber results in Table 2.3, due to the exclusion of closure and glottal stop segments here, it can nevertheless be seen that these baseline systems result in significantly less accurate and consistent alignments compared to manual segmentation.

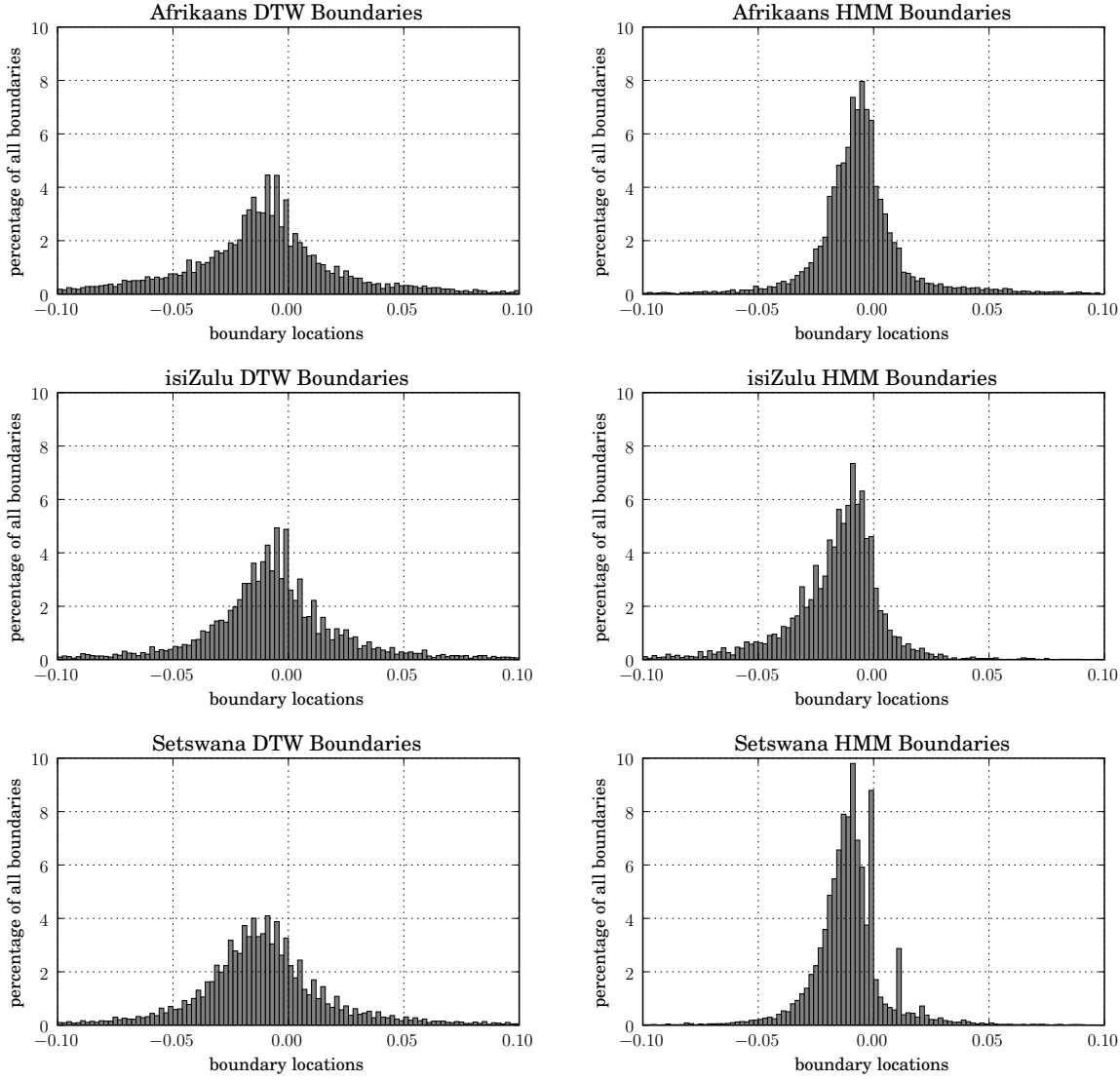

Figure 3.4: *Histograms representing the differences between automated and reference boundary placements. Each histogram consists of 100 bins for differences within 100ms from the reference placement (thus some boundaries are excluded here).*

3.4.2 PHONE OVERLAP RATE

Another perspective on the results can be obtained by considering to what degree each automatically labeled segment is an agreement with its corresponding reference segment via the overlap rate measure (see Section 2.1.1). By calculating this measure for each segment produced by the systems implemented here, one can visualise the phone overlap occurrences obtained by plotting histograms as in Figure 3.5. These plots confirm a high level of gross errors occurring during the DTW process as well as slight biases at the high overlap end for the cases of isiZulu DTW and Setswana HMM comparisons. The main observation to be made here is that in the case of HMM-based alignments, a higher number of segments tend to be concentrated in the higher overlap rate region than is the case for the DTW-based alignments.

Language	< 5ms		< 10ms		< 20ms		RMSE	
	DTW	HMM	DTW	HMM	DTW	HMM	DTW	HMM
Afrikaans	12.05%	25.58%	25.91%	49.16%	47.87%	77.50%	127.32ms	27.58ms
isiZulu	14.74%	16.43%	30.57%	35.03%	53.58%	63.57%	91.74ms	28.79ms
Setswana	12.15%	18.41%	25.08%	38.03%	48.66%	76.57%	185.76ms	21.32ms

Table 3.3: *Summary of the boundary accuracies obtained for each system.*

The advantage of this measure is the possibility of effectively comparing the performance of alignment systems for specific phone types. In Figure 3.6 the mean OR for each phone type is compared for both processes. As expected, the HMM-based system results in higher phone overlap rates for most phone types, especially for phones which frequently occur, and results are closer and less predictable for phones with low frequencies. Looking at the results for each language, in the case of Setswana where the most data is available and the number of distinct phones are relatively lower while a gender mismatch exists with the DTW system, the HMM-based procedure is clearly more successful. In the case of Afrikaans, where a few phones have very few occurrences, the HMM results are still relatively robust, probably because of the fact that the neighbouring phones occur sufficiently often and are thus well modelled, leading to acceptable placement. The results on the isiZulu corpus are however clearly more variable and this is presumably a result of the fact that a broader range of phones have very few occurrences.

A summary of the phone overlap results is found in Table 3.4. Also from these results based on the overlap rate measure, it is evident that neither of the baseline systems result in alignments as accurate or consistent as manual segmentation (see Table 2.3).

Language	DTW		HMM	
	μ	σ	μ	σ
Afrikaans	54.37%	26.69	74.42%	18.16
isiZulu	62.92%	26.09	71.81%	18.15
Setswana	55.27%	26.91	73.26%	15.76

Table 3.4: *Summary of the overlap rates obtained for each system.*

3.4.3 EFFECT OF CORPUS SIZE ON SEGMENTATION ACCURACY

On the corpora tested here, HMM-based segmentation results in more accurate and robust alignments compared to DTW on average as well as for most specific phone types. However, the performance of HMMs depends on the amount of data available for training (i.e. the corpus size) and should degrade for smaller corpora. In order to establish at which point segmentation via an HMM-based system such as implemented here become less reliable than DTW, an experiment was performed where utterances are segmented with both methods for subsets of each corpus ranging from only one utterance to the full size. For each outcome the mean OR is calculated, Figure 3.7 shows these values for each data set size.

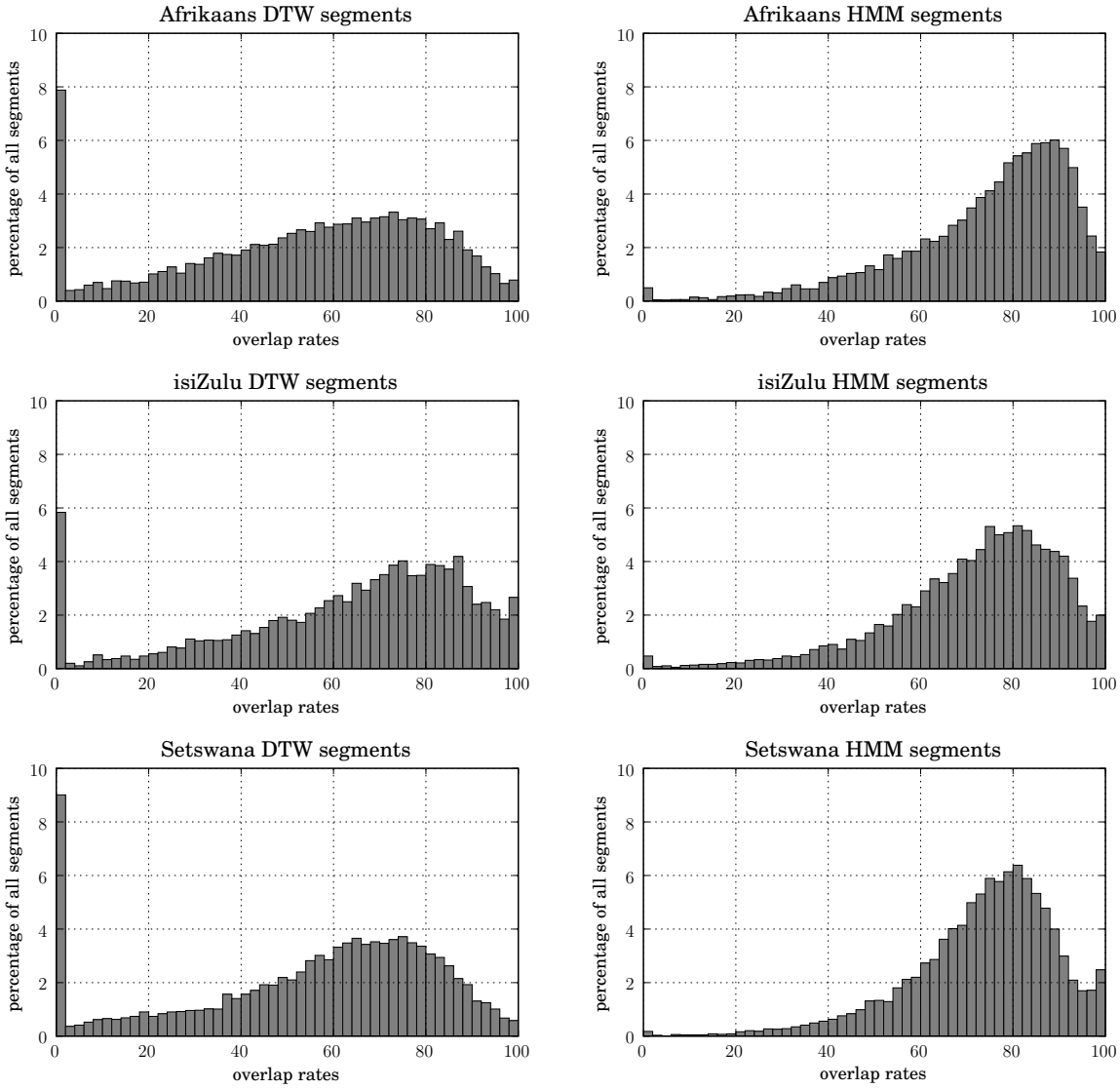

Figure 3.5: *Histograms depicting the number of automatically obtained segments falling into certain overlap rate ranges. Each histogram consists of 50 bins ranging from 0 to 100% overlap.*

Although the mean OR for the HMM-based procedure starts out very low, as expected, it is interesting to see that for all three cases the HMM-based curve is above the DTW curve by the time the mean OR values are relatively stable (at around 20 utterances). This means that the HMM-based procedure results in competitive results for all practical scenarios. The isiZulu curve is notably less stable than the other two cases, presumably because of the larger number of phone types with very few occurrences.

3.5 DISCUSSION

In this chapter two text-dependent segmentation methods were implemented and evaluated. The relative strengths and weaknesses of these methods are largely related to the nature of the "templates"

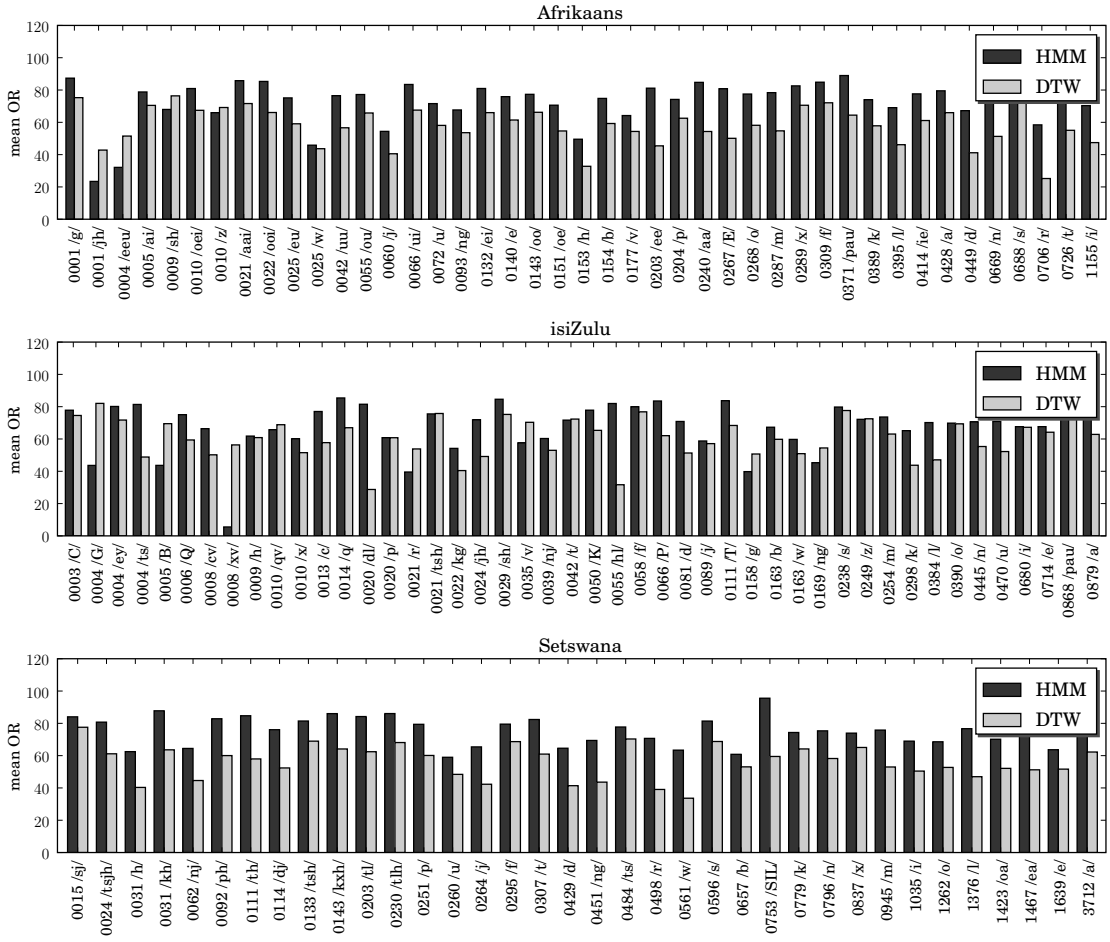

Figure 3.6: *A comparison of the mean overlap rates per phone type achieved by the two segmentation systems. The horizontal axis indicates the phoneme type, along with the number of occurrences of each.*

which are aligned with the input signal through dynamic programming algorithms. From the point of view of implementing these methods, the problems of interest involve generating these templates appropriately.

The HMM-based alignment system is largely data-driven enabling a generic implementation which can be applied relatively conveniently (with no manual intervention) to new languages and conditions of the speech data to be segmented, with the theoretical disadvantage of the performance being a function of the statistical properties of the input data (e.g. corpus size and phone distribution). In contrast while the DTW procedure should present consistent results independent of the statistical properties of the input data, the requirement of a TTS system for the generation of a template signal presents difficulties. In practice, there are a number of points that need to be considered for each new language and speech corpus to be segmented if one is to achieve quality output (see Section 3.3.2). This requires significant expert human interaction and in some cases it is difficult to determine the most prudent setup, especially when no native TTS systems exist for the a particular language. Map-

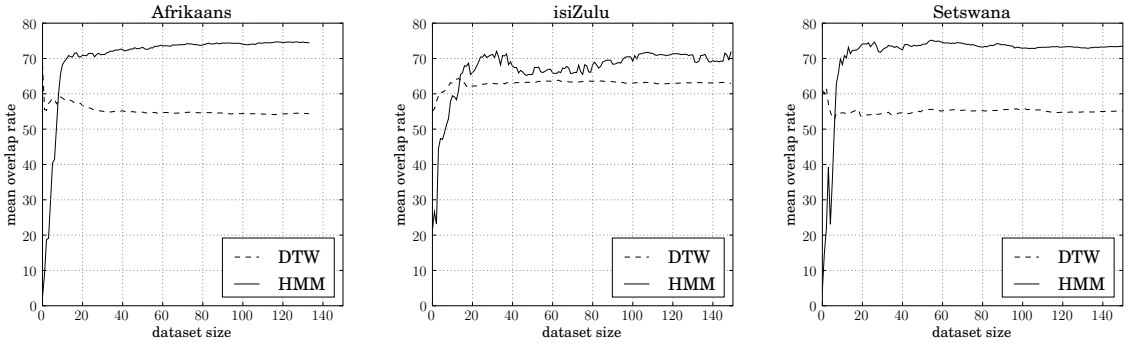

Figure 3.7: *Mean OR for each corpus with data set sizes ranging from 1 to 150 utterances.*

ping native phones to available phones is costly in terms of time and negatively impacts performance (accuracy). Figure 3.8 shows an example of a distance matrix that can be used to troubleshoot alignment when considering mappings. By comparing instances of such a distance matrix visually with various mappings, one can determine the relative success of different mappings by determining the relative acoustic distances between frames from known segments.

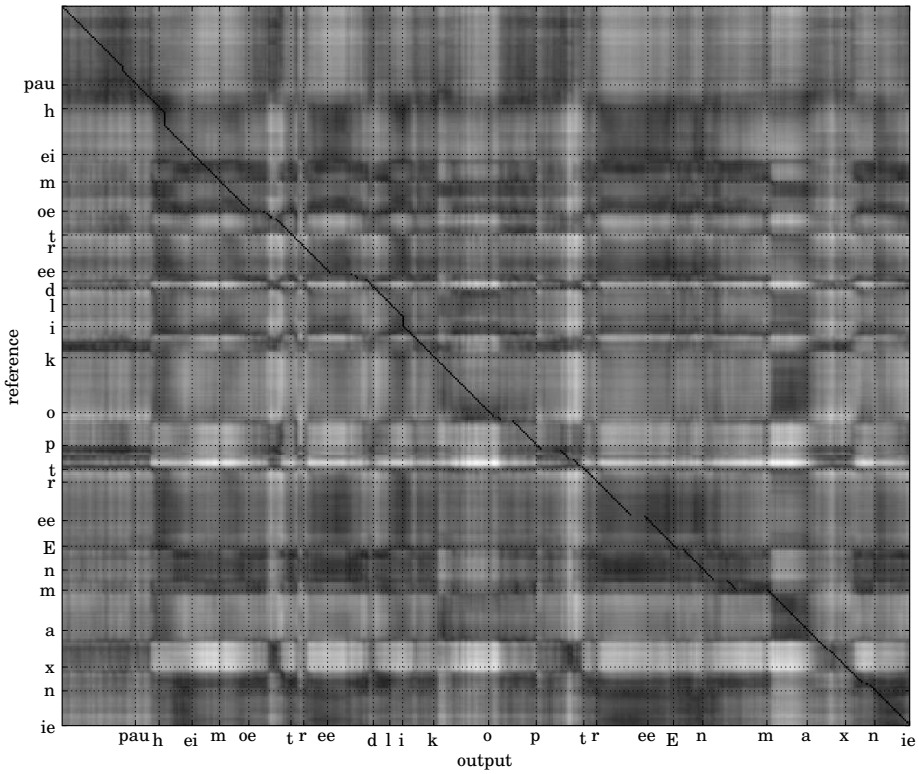

Figure 3.8: *An example of a distance matrix calculated for an Afrikaans utterance, with the corresponding path and label mappings. Darker areas represent lower distances.*

Based on the general results obtained in the previous section, it is clear that the HMM-based procedure performed more accurately and consistently than the DTW-based method, both in terms of

absolute boundary placement and phone overlap rates - in this regard the results are consistent with (Adell *et al.*, 2005). From the overlap rate histograms (Figure 3.5) it is evident that a large number of segments suggested by the DTW method have no overlap with the manually placed segments. On closer inspection of the results it was found that segments with no overlap with the reference segments usually occur in strings (see Figure 3.9 comparing consecutive overlaps for a specific isiZulu utterance for both methods).

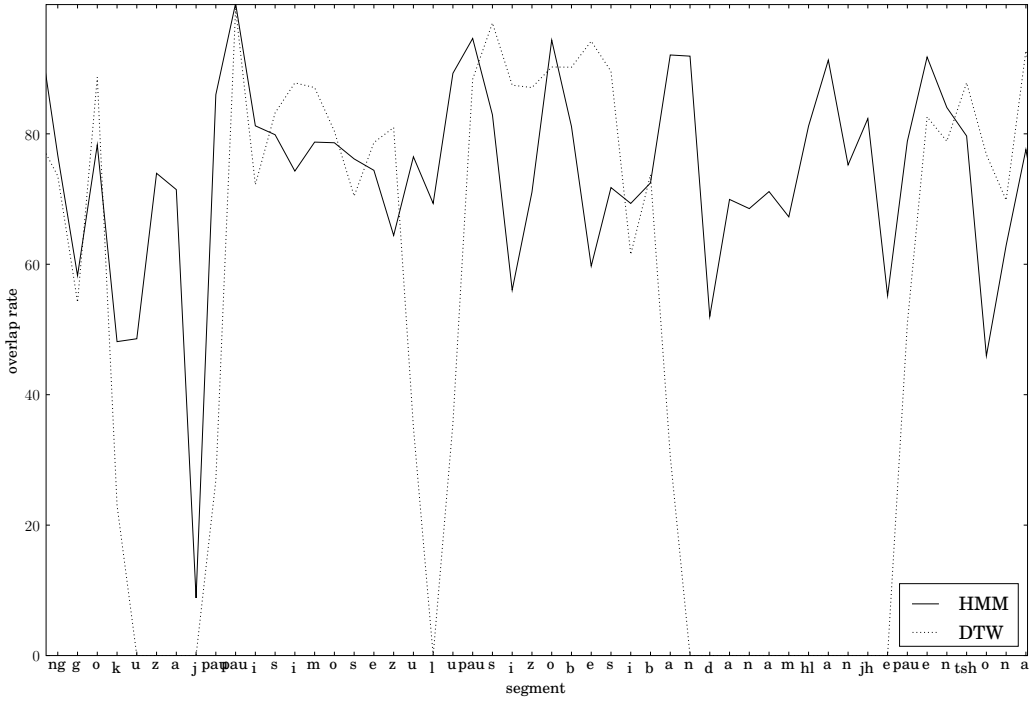

Figure 3.9: *An example of the nature of gross errors that occur during DTW.*

When considering the results presented in Figures 3.6 and 3.7 it is clear that although concerns about data sparsity on segmentation performance is justified, when training and applying speaker specific HMMs as demonstrated here, it is possible to get acceptable results with relatively little data.

3.6 CONCLUSIONS

From a practical perspective in this context, the approach of training a speaker-specific set of HMMs and subsequently applying these models towards phonetic segmentation is preferable over the alternative using TTS-driven DTW. This conclusion is based on the following results and observations from this chapter:

- HMM-based segmentation results can be achieved with minimal interaction (compared to DTW - see Section 3.3.2).
- HMM-based segmentation is more robust than DTW alignments (see Figures 3.3, 3.5 and 3.9).

- HMM-based segmentation is more accurate than DTW alignments (see Figures 3.3, 3.4, 3.5 and 3.6).
- Although sparseness of data (e.g. relatively low number of occurrences of certain phone types) affects the performance of alignments, results obtained are more accurate and robust than DTW in all practical scenarios (see Figures 3.6 and 3.7).
- Neither of the baseline methods investigated here achieve levels of accuracy comparable to manually obtained alignments (see Table 2.3).
- The generation of acoustic models from the data itself results in a useful representation which can be further analysed towards automatic segmentation refinements and quality control.

CHAPTER FOUR

REFINING A HIDDEN MARKOV MODEL-BASED SEGMENTATION SYSTEM

In the previous chapter speech segmentation was performed by applying a generic HMM-based phone recogniser using the maximum likelihood estimation (MLE) training criterion and Viterbi forced-alignment with a set of system parameters and training strategy appropriate for the task of speaker-independent speech recognition. The goals of segmenting a single-speaker speech corpus are however somewhat different from those for the speaker-independent recognition task.

For speaker-independent phone recognition, models are trained towards effectively classifying unseen segments of speech into different classes (phonemes), i.e. minimising the segment classification error rate given a series of observations. In contrast, for single-speaker phonetic segmentation, models are trained and applied on the same data with the goal of accurately determining the temporal positions of phones. This is a process which holistically resembles a form of temporal clustering of the observation sequences into a known sequence of classes.

Considering this, some researchers have redefined components of typical HMM-based recognisers in an attempt to be more aligned with the goals of segmentation. Interesting and successful approaches include the definition of a minimum boundary error (MBE) training criterion (Kuo and Wang, 2006) and the application of the forward-backward algorithm in place of the Viterbi algorithm during the alignment process (Laureys *et al.*, 2002). However, the majority of segmentation approaches successfully employ a relatively conventional phone recogniser such as implemented in the previous chapter.

Before considering methods of local refinement, the aim in this chapter is to investigate factors influencing the accuracy and consistency of alignments within a conventional HMM-based framework with the goal of achieving alignment output as accurate and consistent as possible. This is deemed a

sensible strategy toward a system capable of producing accurate alignments considering the successful results in the previous chapter and the fact that many sophisticated local refinement approaches based on explicitly modeling phone boundaries require a greater number of models and therefore data in order to produce good results. We start the process of investigating appropriate parameters and strategies within this framework by considering the system introduced in the previous chapter as an initial point and refining parameters based on the requirements for speaker-specific segmentation (supported by empirical results obtained on the three reference corpora).

4.1 CONSIDERATIONS FOR AN HMM-BASED SEGMENTATION SYSTEM

If we consider the general components of an HMM-based segmentation procedure, such as depicted in Figure 3.1, when HMMs are trained and applied on a small speaker-specific corpus, the following points need to be explored in order to determine appropriate system parameters:

- Modeling of "sub-phonemic" segments such as closure and burst sections of plosives, and glottal stops.
- The time-frequency resolution trade-off.
- Model generalisation requirements.
- Model and feature complexity.
- Training strategy.

In the following subsections we briefly review the system components related to the points raised and discuss the choices of parameters and motivation for revising some of these choices.

4.1.1 FEATURE EXTRACTION

The first stage in modeling speech signals involves extracting a sequence of feature vectors describing the signal. This is usually based on the short-time properties of the signal framed and windowed at consecutive time intervals, due to the non-stationary nature of speech signals. Widely used representations are based on cepstral or source-filter analyses of the speech signal, with the most successful of these (from the perspective of speech recognition accuracy) being MFCCs (Mel-Frequency Cepstal Coefficients) which represent the spectral properties on a perceptually motivated log-frequency scale (the mel-scale). Such features are most often extracted for window sizes ranging from 20-25ms, with a high degree of overlap between consecutive windows (typically at least half the window size). Overlap between consecutive windows is generally necessary in order to completely describe the signal properties because the application of a window function attenuates the signal at the edges of each frame to some degree, depending on the type of window function used. This practice leads to good results for the recognition task.

These extraction rates and window sizes in combination with the HMM model topology determine the effective time resolution obtainable from the system and should be revisited for the task of segmentation.

4.1.2 MODELS

HMMs are finite state machines which change state and emit an observation vector every time unit. These models are defined by state emission probability density functions and state transition probabilities. For the purpose of phonemic modeling in speech recognition, models with a three-state left-to-right topology are typically used (see Figure 4.1). Each state in the HMM effectively models a segment of a phone with homogeneous acoustic properties (e.g. transitional segments or segments with relatively stationary properties). The emission probability density functions for each state of an HMM are traditionally modeled by GMMs with diagonal covariance matrices.

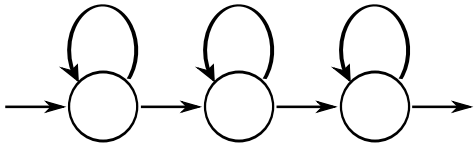

Figure 4.1: *An HMM with a three state left-to-right topology.*

Factors relating to the definition of the models that should be investigated here include:

- Context dependence: It is unclear whether context-dependent or context-independent models (e.g. triphones or monophones) will yield more accurate and consistent alignment results in the given context (especially considering a limited single-speaker corpus).
- State emission model complexity: Increasing the number of Gaussian mixture components typically increases recognition rates for speaker-independent phone models, especially when only diagonal covariance matrices are used. This parameter needs to be investigated for the case of speaker-dependent segmentation.
- Topology: Considering the modeling of very short "sub-phonemic" segments and the requirement of high time resolution output, the effectiveness of the standard topology needs to be investigated.

4.1.3 MODEL ESTIMATION

Estimating HMM parameters involves an unsupervised expectation maximisation process known as the forward-backward (or Baum-Welch) algorithm which is guaranteed to converge to a local maximum likelihood (under the MLE criterion). This algorithm is usually applied through a process of "embedded re-estimation" where individual models are first concatenated to form a compound HMM representing a complete utterance in the training set before re-estimating the parameters of this model

as a whole. The reliance on an EM process such as this where individual models parameters are estimated in context makes parameter initialisation an important consideration. Two different methods depending on the availability of data are generally used (and implemented in the *HTK* toolkit):

- The "flat start" initialisation approach simply initialises all individual model's state emission densities with the mean and covariance matrices calculated over all feature vectors in the training data and is useful when no explicitly labeled "bootstrap data" is available.
- Conversely, when phonemically labeled data is available, it is possible to initialise individual models via a first stage of iterative Viterbi alignment of feature vectors to HMM states (based on a uniform initial segmentation of states within a phoneme), before re-estimating model parameters individually using the forward-backward algorithm.

More appropriate initialisation leads to more accurate model estimation and presumably better system performance. However, especially when data is limited, inaccurate initialisation might degrade performance by forcing models into local maxima leading to decreased generalisation performance. Investigating practical possibilities for improving model estimation via appropriate initialisation in the context presented here is of importance.

4.2 APPLICATION EXPERIMENTS

In the following sections a number of experiments aimed at answering the points raised in the previous section are presented. Starting with the system and parameters used in the previous chapter, we discuss specific requirements of the system output for TTS along with the implications on the data characteristics that need to be considered.

4.2.1 MODELING CLOSURES AND GLOTTAL STOPS

In addition to phonemes which are conventionally modeled for recognition purposes, segment labels of speech for use in systems such as concatenative TTS ideally also contain explicit labels identifying the closure and burst portions of plosive consonants as well as the occurrence of glottal stops. Towards fully automatic segmentation, modeling of segments of this nature are included here within the HMM framework.

At this point we consider the implications of including such segments on the nature of the data and the effect on the results obtained from a generic system such as used in 3.3.1.1. Figure 4.2 depicts the overlap rates per phone type achieved by the segmentation system on the data sets including and excluding closures and glottal stop segments. If one compares the overlap rates of the plosive and affricate segments for these two cases, it is evident that the overlap rate is significantly lower where the burst portions are independently modeled from the closures. Overlap rates range from less than 20% to 60% compared to approximately 80% when not considering closures explicitly, with dental and bilabial plosives having the lowest overlap rates.

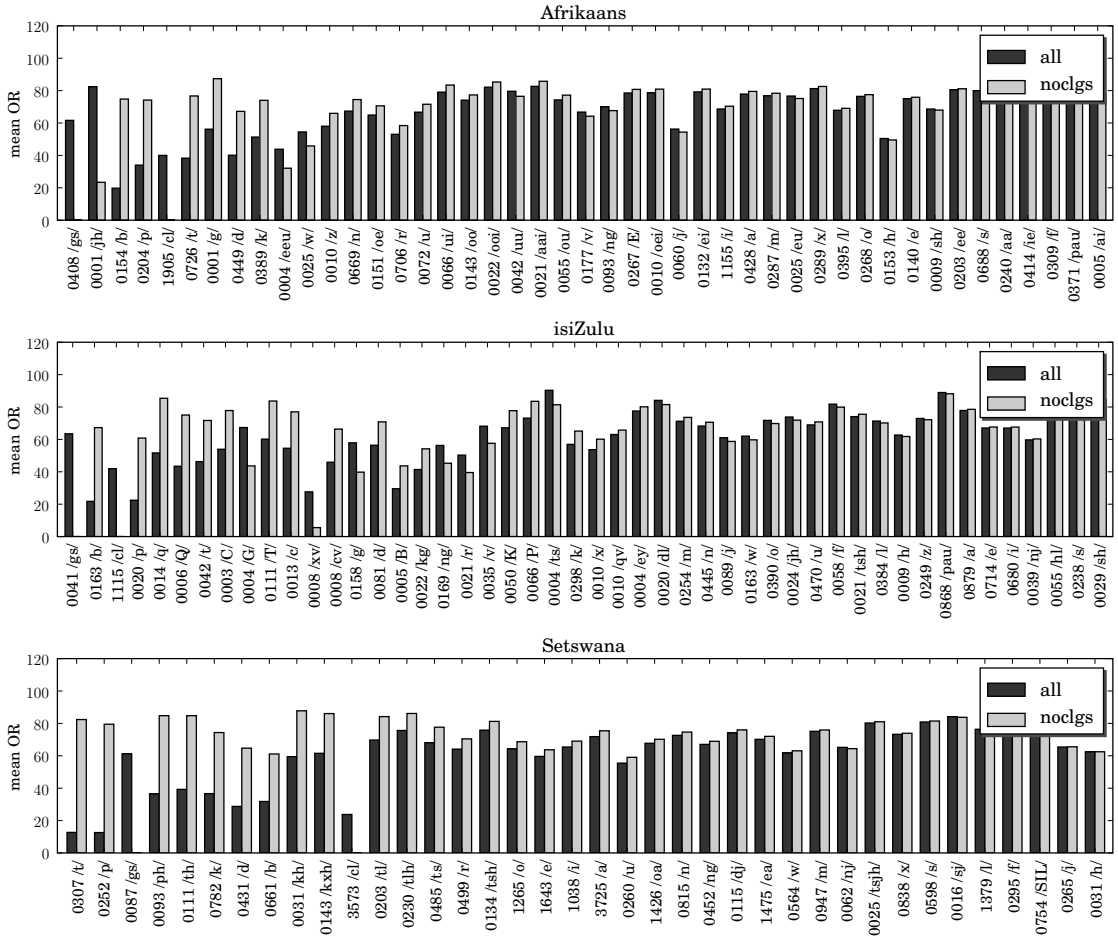

Figure 4.2: *Mean overlap rates achieved by the baseline system when including and excluding closure and glottal stop segments.*

It is clear from these results that independent modeling of plosive closure and burst portions in particular, result in significantly reduced overlap rates for these segments, suggesting that these segments are not as satisfactorily modeled and isolated compared to other phones. Considering that our conventional three state left-to-right HMM topology imposes a minimum phone duration constraint of three times the step size (that is a minimum phone duration of 30ms for our conventional system parameters in Table 3.1), Table 4.1 shows why the results in Figure 4.2 can be expected.

Category	Afrikaans		isiZulu		Setswana	
	< 30ms	total	< 30ms	total	< 30ms	total
plosives	51.12%	1923	37.70%	1016	33.76%	3101
closures & glottal stops	13.71%	2313	31.40%	1156	9.34%	3660
trills	18.56%	706	4.76%	21	7.41%	499
fricatives	2.51%	1636	0.00%	622	0.06%	1778
approximants	5.00%	480	1.45%	691	5.62%	2208
vowels	0.54%	3863	0.06%	3137	0.41%	10832
nasals	0.48%	1049	0.44%	907	0.04%	2276
other	0.00%	371	0.21%	940	0.00%	754
affricates	N.A.	N.A.	0.00%	69	0.33%	902
total	12.33%	12341	8.94%	8559	6.15%	26010

Table 4.1: *Proportions of segments with durations of less than 30ms*

Based on these observations, a logical proposition would be to modify the nature of the HMM models by dividing the three states used into models with fewer states (e.g. a one-state model for closure segments and two-state model for the plosive burst segments), expecting improved alignment results. Although this would certainly allow the system to specify shorter durations for these segments, it is later shown (Section 4.2.3.4, Table 4.8) that alignment accuracy is reduced in this case compared to using three-state HMMs throughout.

4.2.2 FEATURE EXTRACTION

In order to effectively model segments with relatively short durations, the system needs to be capable of output at a higher time resolution. One way to achieve this is by increasing the time resolution of the underlying signal representation. In the following sections the effects of various parameters regarding feature extraction are investigated empirically.

4.2.2.1 INITIAL OBSERVATIONS

Before considering the effects of changing the feature resolutions, some initial experiments comparing feature types were performed. Segmentation results comparing MFCCs and LPCs (Linear Prediction Coefficients) implemented in *HTK* seemed to confirm the relative effectiveness of MFCCs also in this context. We also experimented with the feature vector coefficients. It is customary when building speech recognition models, to extract around 12 MFCCs, include the energy and the delta

and acceleration coefficients of these 13 base features combined into a 39-dimensional vector. Experiments were conducted comparing the base features (13 coefficients) with base features including deltas (26 coefficients) and base features with deltas and delta deltas (39 coefficients). While the mean overlap rates for these three cases were comparable, the 39-dimensional features clearly resulted in the most consistent results (i.e. with significantly lower variance in the overlap rate).

4.2.2.2 EXPERIMENT 1: FEATURE RESOLUTION

When framing and windowing the signal prior to feature vector calculation the step and window sizes effectively determine the time and frequency resolution. Smaller window and step sizes increase the effective time resolution while a smaller window size reduces the accuracy of the spectral estimate.

As a result of the application of a window function to each of the signal frames in order to improve the spectral estimate despite discontinuities at the edges of the window, the step size is usually chosen so that consecutive windows overlap in time. Other parameters depending on the variation of the window and step sizes include the number of frames used to calculate the delta and acceleration coefficients.

In this section we experiment with features extracted at various resolutions in order to determine the effect of these parameters on segmentation accuracy. The same basic procedure is followed as described in Section 3.3.1.1 for a range of window and step sizes (see Table 4.2).

Features	
Type	MFCCs (39 coefficients)
Window function	Hamming
Window size	**1 - 40ms**
Step size	**1 - 19ms**
Models	
Initialisation	Flat start
Topology	3-state left-to-right
State distributions	1 mixture GMM
Context-dependence	Tied-state triphones

Table 4.2: *Summary of parameters used during experiment 1.*

Figure 4.3 depicts mean overlap rates and corresponding standard deviations for each of the corpora at various step and window sizes.

Under the circumstances considered here it is evident that two regions with higher (and more consistent) overlap rate exist for each corpus (especially evident for Afrikaans and Setswana). This is due to the phone type distribution - phonetic segments with different acoustic properties achieving better accuracies when the feature resolution matches, e.g. shorter segments and segments with high frequency energy are sufficiently modeled by short windows with the higher time resolution benefiting segmentation results while voiced segment models such as vowel models benefit from window sizes which can include at least one or two pitch periods, causing an increase in segmentation accuracy despite the lower time resolution.

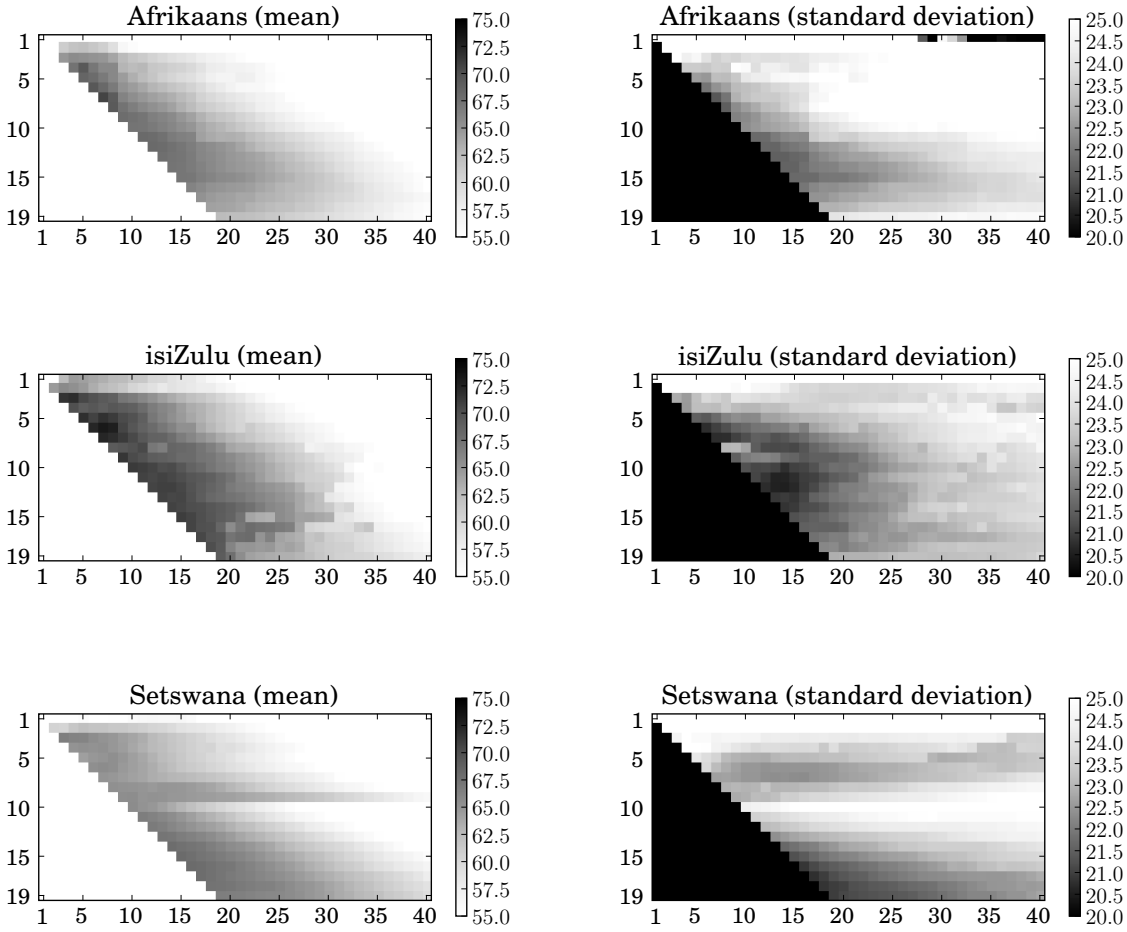

Figure 4.3: *Plots of the mean and standard deviations of the overlap rate for ranges of the window and step size where* $windowsize \geq stepsize$*. Darker points represent higher mean overlap rate as well as lower deviation. Highest overlap rates achieved using flatstart model initialisation are as follows: Afrikaans: 70.75% where* $stepsize$ *= 7ms and* $windowsize$ *= 7ms, isiZulu: 73.21% where* $stepsize$ *= 6ms and* $windowsize$ *= 7ms and Setswana: 67.99% where* $stepsize$ *= 15ms and* $windowsize$ *= 15ms.*

In the Setswana case the results show some bias (as before) for parameters close to the parameters used before manual checking (i.e. 20ms window size with a 10ms step size).

Another observation is that higher overlap rates are more consistently achieved when window and step sizes are chosen so that there is little overlap between consecutive windows, i.e. it seems that choosing the smallest practical window size for a particular step size is beneficial.

4.2.2.3 EXPERIMENT 2: PITCH-SYNCHRONOUS FEATURES

The results obtained in the previous section suggests that the optimal window and step size combination (time-frequency trade-off) is largely dependent on the phone category, especially between categories with distinctly different spectral characteristics. Thus one could presumably improve segmentation results globally (over all phonetic categories) if one could extract features at different rates depending on the properties of the phone. One way of achieving this would be to detect voiced and unvoiced sections of speech and perform fundamental frequency (f_0) analysis of the voiced parts in order to optimally extract features. Features for unvoiced sections can be extracted at a default higher rate with smaller window size under the assumption that these sections mostly contain distinguishing high frequency energy, while features for voiced sections can be extracted to include a multiple of the pitch period at relevant time instants (Figure 4.4 illustrates a pitch-synchronous feature framing scheme).

Such a scheme was implemented using the pitchmark estimation algorithm in the *Praat* software package (Boersma, 2001) to extract pitchmarks for voiced sections (in the 75-300Hz fundamental frequency range) and unvoiced sections were filled with extraction instants at a default step size. Subsequent feature extraction based on pitchmark locations was performed using the *Edinburgh Speech Tools* package (Taylor *et al.*, 1999) extracting 39-dimensional MFCCs comparable to the features calculated by *HTK* used in the previous experiment. Thus for this experiment the *feature extraction* component (see Figure 3.1) is effectively replaced by a pitch-synchronous version where the window and step sizes vary according the output of a prior pitch analysis stage. Table 4.3 summarises the experimental parameters.

Features	
Type	MFCCs (39 coefficients)
Window function	Hamming
Window size	$\mathbf{2 \times stepsize}$
Step size	$\mathbf{f_0^{-1}}$ **or 5ms**
Models	
Initialisation	Flat start
Topology	3-state left-to-right
State distributions	1 mixture GMM
Context-dependence	Tied-state triphones

Table 4.3: *Summary of parameters used during experiment 2.*

The results obtained from this process are compared to the process based on static resolution

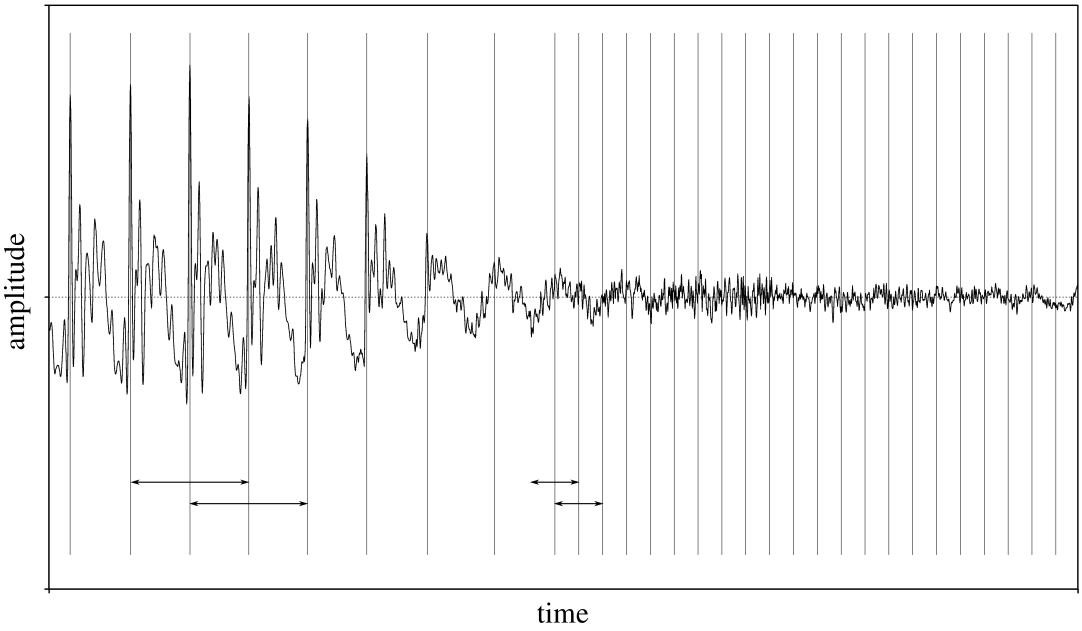

Figure 4.4: *An example of how a speech signal is analysed in order to extract features pitch-synchronously. The vertical lines represent central points around which windows are extracted, at the start of the example, these points are determined by fundamental frequency analysis of the voiced section, while for the unvoiced section towards the end of the signal, extraction points are regularly placed based on a default step size. The horizontal arrows indicate the window size for windows centered at different extraction points*

features at two different operating points (05s_10w - where the step size is 5ms and window size is 10ms, and 10s_20w - where the step size is 10ms and window size is 20ms.) The comparison in terms of the mean overlap rate per broad phone category can be seen in Figure 4.5.

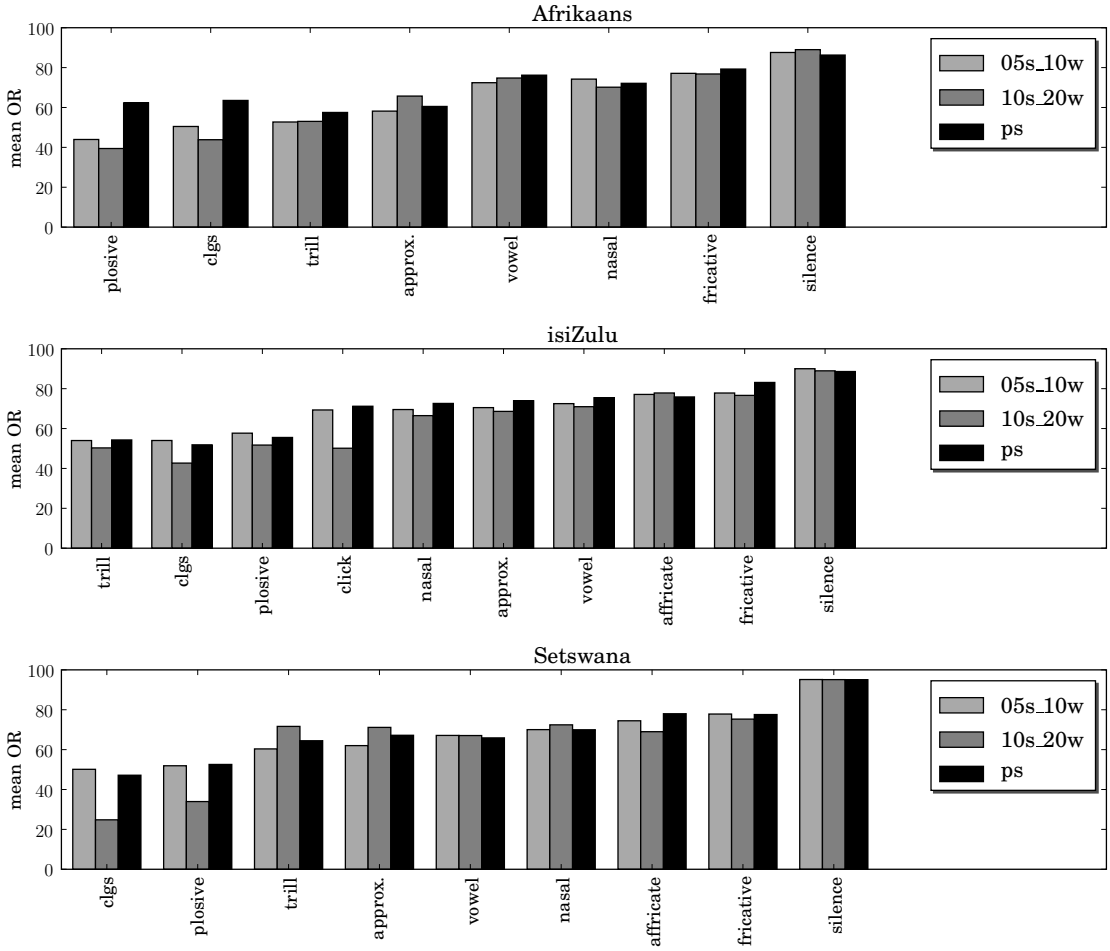

Figure 4.5: *A comparison of the mean overlap rates achieved on each broad phone category by the pitch-synchronous and static resolution features.*

Considering especially the vowel, plosive and closure categories, it is clear that this feature extraction process is successful in facilitating more consistently accurate results across phone categories for the cases of Afrikaans and isiZulu compared to either of the static resolution cases. These results are summarised in Table 4.4. In the case of Afrikaans and isiZulu the alignments based on the pitch-synchronous scheme is clearly more accurate and consistent, while for Setswana the mean accuracy is comparable with the 05s_10w case with a higher degree of deviation. This is probably attributable to the fact that the female voice varies to a higher degree (range) with respect to f_0, possibly leading to more pitchmark errors within the frequency range tested here.

Overall, the results presented here suggest that utilising pitch synchronous feature extraction can indeed lead to an improvement, by modeling segments with diverse acoustic properties well. Because models are trained on vectors of varying window size relying on the location of pitchmarks, model

Language	ps		05s_10w		10s_20w	
	μ	σ	μ	σ	μ	σ
Afrikaans	70.35%	22.35	63.42%	24.12	62.17%	24.17
isiZulu	71.28%	21.38	69.90%	22.25	66.26%	21.80
Setswana	64.18%	24.45	64.41%	22.90	59.52%	25.26

Table 4.4: *Summary of the comparison between the pitch-synchronous and static resolution features.*

state distributions based on these feature vectors are highly dependent on the pitchmark placements and as a result boundary placements are influenced significantly by the pitch analysis. Thus the segmentation results are sensitive to the accuracy of the initial pitch analysis and can lead to a possible increase in gross errors resulting from corresponding errors during the pitchmark estimation process. This trade-off between accuracy and robustness might be desirable when training models from a flat start as in this experiment and the expected number of gross errors can be reduced by carefully considering the fundamental frequency range used during the pitch analysis based on the properties of the voice (e.g. male or female).

Further interesting experiments along these lines might include segmentation based on feature vectors representing a combination of coefficients extracted at different rates (i.e. multi-resolution features).

4.2.3 MODELS

Effectively modeling the underlying segments is undoubtedly another very important consideration which should have an impact on the accuracy of the segmentation. Factors which have potential to influence modelling effectiveness include the model topology, state distribution complexity, distinctness of segments modelled and training procedure (as mentioned in section 4.1.2). In the following sections we present experiments aimed at selecting appropriate parameters and procedures.

4.2.3.1 EXPERIMENT 3: MODEL INITIALISATION

Thus far all experiments have been performed with a generic training procedure based on the flat start initialisation procedure. Such a training procedure is often used. However, if pre-labeled data is available, this training procedure can be enhanced by a more precise model initialisation process. A “bootstrap” data set of this nature should typically be selected with care as the initialisation of models based on this data has a significant impact on the final model parameters (due to the nature of the EM process) which can lead to generalisation problems which is less likely with the flat start approach in the context of embedded re-estimation (described in Section 4.1.3). As a consequence, model initialisation for speaker-independent recognition typically requires a large number of labelled utterances. In the current context, initialising models with extensive amounts of data is not feasible and may not be necessary. Considering the application of the models, generalisation is not required beyond the training data - i.e. models need only accurately describe the acoustic properties and

variability of each phone (or unit modeled) occurring in the single-speaker corpus.

Language	Lang. group	Gender	Utterances	Duration	Phones
Afrikaans	Germanic	Male	17	181 sec.	1714
isiZulu	Nguni	Male	32	259 sec.	1838
Setswana	Sotho	Female	11	140 sec.	1380
Afrikaans	Germanic	Male	115	18 mins.	10382
isiZulu	Nguni	Male	118	16 mins.	6721
Setswana	Sotho	Female	321	45 mins.	24630

Table 4.5: *Properties of the subsets used for bootstrapping and subsequent training and labeling.*

In order to investigate the possibility of bootstrapping the training process, we select minimal subsets of utterances from each of the corpora, with each of the selected subsets containing at least three occurrences of each phone in the specific language (see Table 4.5). These utterances are removed from the main corpus and used during the initialisation process only. This involves using the aligned input transcriptions in order to perform an iterative Viterbi alignment of model states and feature vectors based on a uniform initial segmentation followed by forward-backward re-estimation for each individual phone in isolation. The subsequent embedded re-estimation process and alignment is completed on the remaining utterances as before. Figure 4.6 depicts a block diagram of the process followed in this experiment.

Performing the segmentation procedure with the more sophisticated initialisation on the remaining utterances over a range of feature resolutions results in Figure 4.7.

The addition of a more distinct model initialisation stage to the process evidently increases the overlap rates observed on all corpora as well as improving the consistency of alignments (lower variance). Thus one can conclude that bootstrapping in this context (even with minimal data) can lead to beneficial results when compared to a flat start approach. Comparing the results found here with the experiment based on the flat start training it is also evident that the region with highest overlap rate is now more distinctly concentrated at a higher time resolution, suggesting that the results found in the previous experiments were largely attributable to the training from flat start initialisation leading to better trained models for certain phone types when using complimentary feature parameters.

Language	ps		05s_10w		10s_20w	
	μ	σ	μ	σ	μ	σ
Afrikaans	73.82%	20.68	75.93%	19.81	71.98%	19.65
isiZulu	77.56%	18.75	78.43%	18.86	77.33%	18.18
Setswana	76.75%	20.18	77.74%	18.18	75.62%	16.56

Table 4.6: *A comparison of the results when initialising the training process with a minimal bootstrap data set.*

Table 4.6 summarises the results found when combining phonemic bootstrapping with pitch-synchronous features. It is interesting to note that for this experiment static resolution feature extraction (at high extraction rates) outperforms the variable pitch-synchronous scheme (marginally but

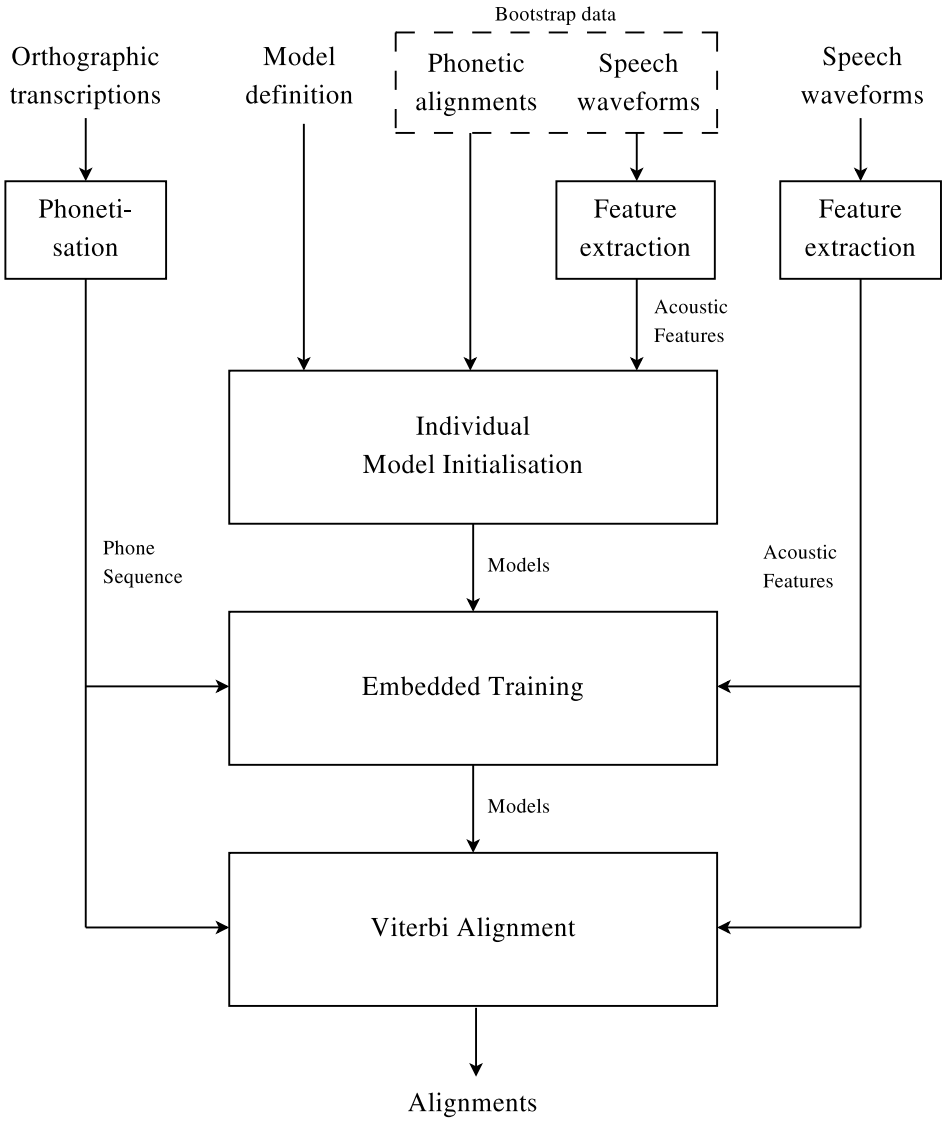

Figure 4.6: *A bootstrapped HMM-based alignment system.*

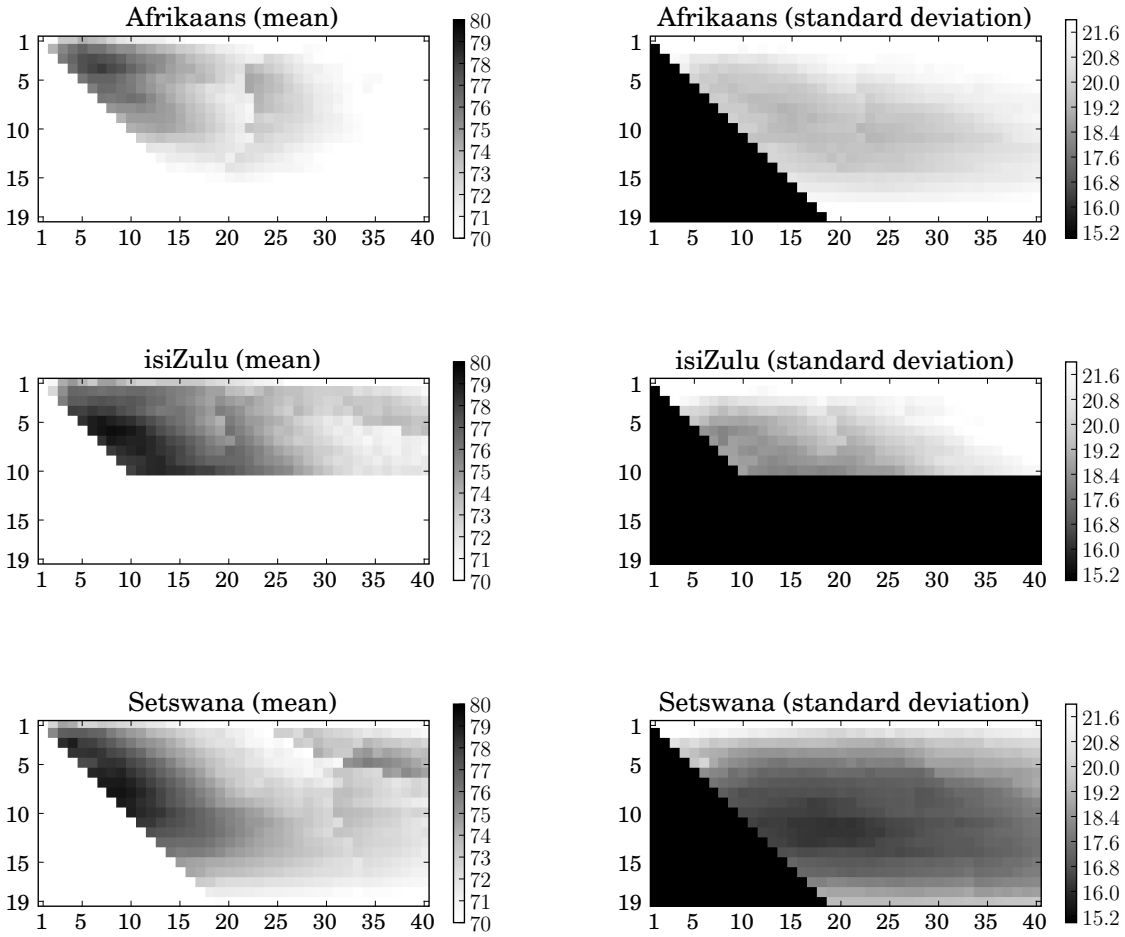

Figure 4.7: *Plots of the mean and standard deviations of the overlap rate for ranges of the window and step size where* $windowsize \geq stepsize$*, for models bootstrapped with phonemically transcribed data. Darker points represent higher mean overlap rate as well as lower deviation. In the case of isiZulu, the experiment could only be run for step sizes up to 10ms due to difficulties initialising infrequent short segments. Highest overlap rates achieved using minimal data for model initialisation are as follows: Afrikaans: 78.14% where* $stepsize$ *= 4ms and* $windowsize$ *= 7ms, isiZulu: 79.54% where* $stepsize$ *= 6ms and* $windowsize$ *= 8ms and Setswana: 79.30% where* $stepsize$ *= 8ms and* $windowsize$ *= 9ms.*

also with less variance in the overlap rate). It could be that the degree of error in the pitchmark estimation is significant when comparing the results to the well trained models based on static resolution features.

4.2.3.2 EXPERIMENT 4: CROSS-LANGUAGE MODEL INITIALISATION

In view of the positive results obtained in the previous experiment compared to the flat start initialisation and considering that even a very small amount of bootstrap data benefited the alignment accuracy significantly, we consider the possibility of employing resources in another language to bootstrap the training process. In this section we experiment with this notion by proposing that each phone model in the corpus be initialised according to the broad phonetic category it belongs to (assuming that the acoustic properties of each phone is largely similar within these categories). Thus when training models for segmentation in a new language without any available resources, the idea is that all phones belonging to a specific broad category from a labeled corpus in a different language are used to initialise each phone model in the local language belonging to the same broad category.

For this experiment the same procedure as in the previous section (4.2.3.1) is followed, except that the native language bootstrap data is replaced by the DARPA TIMIT (Garofolo *et al.*, 1993) testing corpus (this consists of male and female speakers of American English from eight different regions or dialects). Each phone is initialised by following the mapping strategy described above, i.e. each phone belonging to a specific broad category starts with a model identical to all other phones in the same category, based on the acoustic properties of all phones in the same category in the bootstrap set (exact categorisations used for this experiment can be found in Appendix B). The only form of normalisation of the acoustic features used during this experiment is Cepstral Mean Normalisation (CMN), which was indeed performed during all the experiments thus far.

The results are presented in Figure 4.8. The mean overlap rates achieved during this experiment is clearly very similar to the results from the previous experiment, both in terms of the absolute range of the overlap rate as well as the trend in higher overlap rate regions.

An initialisation strategy such as proposed and demonstrated here would allow one to perform segmentation of prototypical corpora in new languages without any human interaction as in this case no pre-aligned utterances are needed in the new language and due to the simplicity of the mapping process this can be done automatically using the information about the properties of phonetic segments that are generally acquired during the process of building spoken language systems such as TTS. The resulting accuracy of alignments is also higher than what can be achieved through the simple flat start and pitch-synchronous approaches demonstrated in the previous sections.

4.2.3.3 EXPERIMENT 5: CONTEXT DEPENDENCE AND STATE DISTRIBUTIONS

For building a speaker-independent phone recogniser, it is standard practice to train context-dependent phone models (either bi- or triphones) in order to compensate for the distinct acoustic properties exhibited by phones in different contexts. Due to the fact that this generally leads to a significant

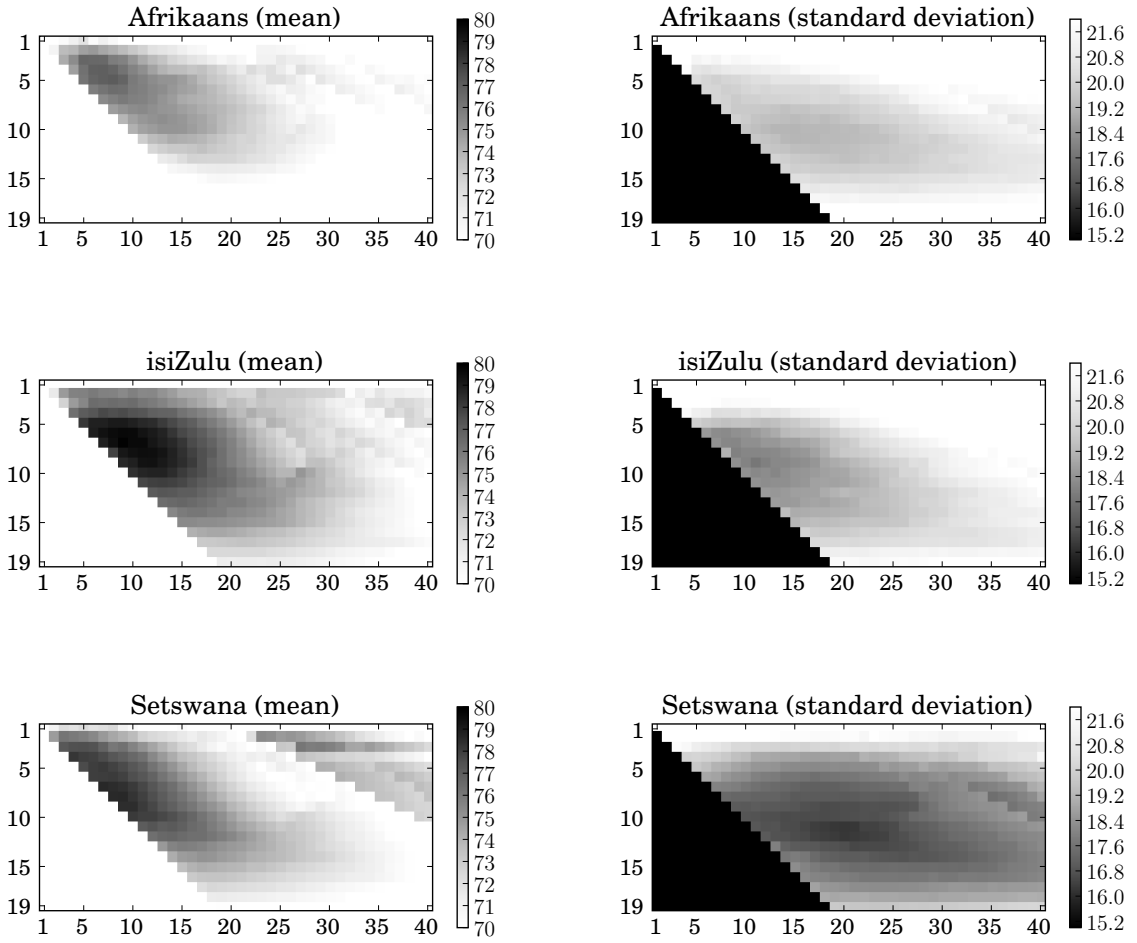

Figure 4.8: *Plots of the mean and standard deviations of the overlap rate for ranges of the window and step size where* $windowsize \geq stepsize$ *with cross-language initialisation. Darker points represent higher mean overlap rate as well as lower deviation. Highest overlap rates achieved using mapped data for model initialisation are as follows: Afrikaans: 77.08% where* $stepsize$ *= 5ms and* $windowsize$ *= 8ms, isiZulu: 79.90% where* $stepsize$ *= 7ms and* $windowsize$ *= 9ms and Setswana: 78.69% where* $stepsize$ *= 8ms and* $windowsize$ *= 9ms.*

increase in the number of models required to model all distinct triphones (or biphones) occurring in a particular language, the technique of decision tree-based state-tying is usually employed to cluster and share state distributions between such models in order to enable sufficient training in practice (Young *et al.*, 1994, 2005). This approach generally leads to higher phone recognition rates. However for the segmentation task, some researchers have suggested that the use of context-dependent models results in less accurate alignments due to the additional modeling of the phonetic context (Ljolje *et al.*, 1996). Other experiments have shown that in some cases context-dependent models do outperform context-independent models at segmentation (Toledano *et al.*, 2003).

Similarly it is customary when developing a phone recogniser to consider increasing the state distribution complexity by increasing the number of Gaussian mixture components. This is usually done systematically until empirical recognition results on a predefined test set indicate optimal modeling.

Features	
Type	MFCCs (39 coefficients)
Window function	Hamming
Window size	10ms
Step size	5ms
Models	
Initialisation	Native bootstrap
Topology	3-state left-to-right
State distributions	**1 - 16 component GMM**
Context-dependence	**Tied-state triphones / monophones**

Table 4.7: *Summary of parameters used during experiment 5.*

For this experiment the system depicted in Figure 4.6 is employed with the initialisation process based on the minimal bootstrap data selected (Table 4.5) in order to determine the effects on alignment performance when increasing the number of Gaussian mixture components for both triphone and monophone models. Triphone models are trained by duplicating initial monophone models, re-estimation and subsequent decision tree-based state-tying based only on questions considering the monophone context (i.e. no sophisticated phonetically motivated rules are introduced). Table 4.7 summarises the experimental parameters for this experiment.

The results obtained are depicted in Figure 4.9. Triphone-based segments result in consistently higher overlap rates with the manual reference segments. There is also no gain in overlap rates by increasing the number of Gaussian mixture components per state, in fact a linearly decreasing overlap rate is observed. This might be attributable to the fact that single-speaker models do not benefit from more complex models or that the training data does not allow sufficient modeling of more complex structures in the state output distributions.

4.2.3.4 EXPERIMENT 6: MODEL TOPOLOGY

Choosing a suitable model topology is a critical consideration in any HMM-based system in that the HMM topology is essentially an assumption about the temporal structure of the data. If a particu-

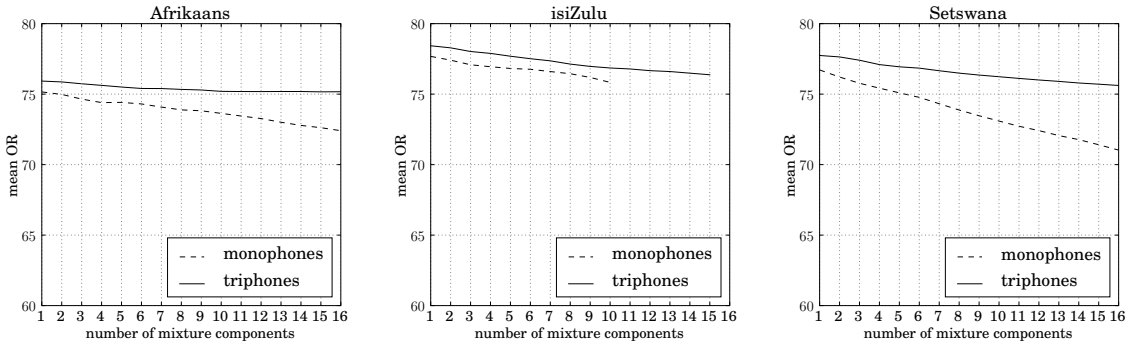

Figure 4.9: *Mean overlap rates obtained when varying the number of Gaussian mixtures per state for both triphones and monophones.*

lar topology is chosen which is not largely representative of the temporal structure of the segments to be modeled, then this could lead to reduced modeling accuracy and as a consequence reduced performance.

Determining an appropriate model topology is difficult in practice. By modeling a data sequence with models of different topologies one may uncover temporal structures of different nature. If one possesses sufficient amounts of phonetically labeled data, various strategies such as the state splitting algorithm discussed in (Murrell and Tapamo, 2008) could be implemented in order to infer a likely topology. However, in a scenario where training is largely unsupervised, providing a suitable topology is critical. Thus the selection of a basic topology is often decided by prior knowledge of the temporal properties of the units to be modeled. Although various different topologies have been used to model phones, a popular topology is the so-called left-to-right (see Figure 4.1) topology where all states have to be traversed and transitions only allow repeating a single state or skipping one state forward. Such left-to-right models allow a sensible uniform segmentation initialisation strategy such as implemented in *HTK*, which simplifies the process considerably.

In these experiments this basic topological design is adopted so that experimenting with the topology reduces to considering the number of states per model. Thus far we have employed a standard three-state model topology as the basis for all segment types. In Section 4.2.1 the aim of explicitly labeling and thus modeling of closure and burst sections of plosives and similar phones was proposed. If one considers that this involves modeling two shorter segments which are traditionally modeled by a single HMM comprising three states, then it makes intuitive sense to attempt to modify HMM topologies for those segments specifically, possibly by decreasing the number of states per model. Assuming that closure segments have acoustic properties with a simpler temporal structure than burst segments, we perform an experiment based on experimental parameters identical to what was considered in Section 4.2.3.1, with the exception that all closures are modeled with single-state HMMs and all plosive burst segments are modeled by two-states HMMs.

In Table 4.8, a comparison of the alignment results (higher and lower time resolution features) can be found between three-state models used throughout and the shorter models for closure and burst

Language	particular models 05s_10w		uniform models 05s_10w		particular models 10s_20w		uniform models 10s_20w	
	μ	σ	μ	σ	μ	σ	μ	σ
Afrikaans	71.52%	22.60	75.93%	19.81	71.68%	20.91	71.98%	19.65
isiZulu	76.73%	21.87	78.42%	18.86	75.20%	21.55	77.33%	18.18
Setswana	74.35%	22.60	77.74%	18.18	73.17%	20.86	75.62%	16.56

Table 4.8: *The overlap rates achieved when using one-state and two-state models for closure and burst portions of plosive phones respectively compared to simply using three-state models throughout all segment types.*

segments proposed. Three-state models used throughout clearly results in consistently more accurate alignments. This is possibly due to one and two-state models not efficiently modeling even these short segments.

Consequently we investigate the effect of varying the number of states per model on alignment accuracy, once again using a single model topology for all segment types. The experimental parameters are summarised in Table 4.9.

Features	
Type	MFCCs (39 coefficients)
Window function	Hamming
Window size	10ms
Step size	5ms
Models	
Initialisation	Native bootstrap
Topology	**[1 - 8]-state left-to-right**
State distributions	1 mixture GMM
Context-dependence	Tied-state triphones

Table 4.9: *Summary of parameters used during experiment 6.*

The results in Table 4.10 indicate that more states seem to improve the accuracy and consistency of results, with a peak in the overlap rate measured for models consisting of around five states. Increasing the number of states much beyond this leads to a gradual decrease in the overlap rate which can once again be associated with the corresponding higher minimum duration imposed by using more states.

Further experimentation with models containing fewer states for short closure and plosive segments did not yield results conclusively better than segmentation with 5-state HMMs throughout.

4.3 DISCUSSION

In this chapter the requirements of segmentation were considered (Section 4.1) in order to refine a baseline HMM-based phone recogniser configuration towards achieving more accurate alignment results in the given context. Firstly, the explicit modeling of important sub-phonemic segments was included and it was shown that the inclusion of such segments requires refining the system to allow higher time resolution alignments.

Number of states	Afrikaans		isiZulu		Setswana	
	μ	σ	μ	σ	μ	σ
1	65.54%	24.20	71.00%	22.16	68.36%	22.26
2	74.41%	21.13	73.63%	18.68	72.92%	19.22
3	75.93%	19.81	78.43%	18.86	77.74%	18.18
4	76.89%	19.63	80.25%	17.80	79.55%	17.56
5	77.15%	19.15	80.40%	17.34	79.34%	17.03
6	76.83%	19.03	-	-	79.59%	16.64
7	75.40%	19.52	-	-	78.87%	16.66
8	74.52%	19.84	-	-	78.77%	16.57

Table 4.10: *Overlap rate statistics on the three corpora for increasing number of states per model.*

A systematic set of experiments was performed in an attempt to understand how the various processes and parameters involved during the application of HMMs for speaker-specific segmentation in the context affect segmentation performance. During these experiments, the overlap rate was used as the primary measure of success.

Experimenting with a range of feature resolutions with a flat start based model initialisation approach proved problematic, with different feature resolutions leading to improved performance for different types of phonemes based on the acoustic properties of segments. This problem was overcome by developing a pitch-synchronous feature extraction scheme which achieved a relative improvement in alignment accuracy.

Further experiments showed that a large gain in alignment accuracy could be achieved by initialising individual phone models more appropriately and that this could be practically realised by a minimal set of bootstrap data or by using an existing high quality corpus of non-native language data via a simple broad phone category mapping.

Experimenting with model complexity showed that the use of context-dependent models in conjunction with a decision tree-based state tying procedure consistently resulted in more accurate results. In this context, a single component GMM state emission probability density function performed best. Lastly the accuracy of alignments tended to increase with an increasing number of HMM states. Experimenting with the model topology for specific segment types (e.g. trying to model shorter segments with fewer states) did not yield a substantial improvement compared to a single model topology used uniformly over all segment types.

The improvement in accuracy obtained in this chapter over the baseline system from Chapter 3 in comparison with the results obtained when comparing inter-transcriber agreement in Section 2.2.1 can be seen in Figure 4.10 and Table 4.11 in terms of the boundary accuracy and overlap rate measures respectively. These results suggest that the system accuracy has reached the level where further improvement in alignments will be difficult to reliably measure against the reference alignments.

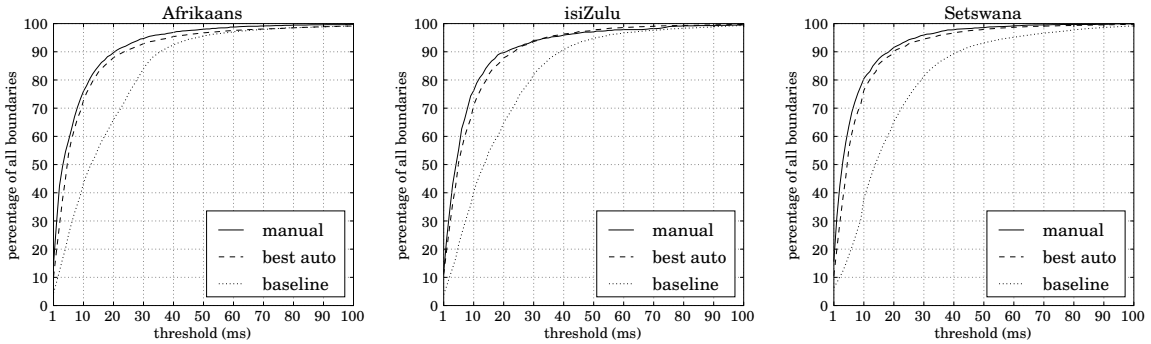

Figure 4.10: *A comparison of the boundary accuracy curves obtainable by the baseline and refined system in relation to manual agreement.*

Method	Afrikaans		isiZulu		Setswana	
	μ	σ	μ	σ	μ	σ
baseline	62.17%	24.18	66.26%	21.80	59.78%	25.08
best auto	77.15%	19.15	80.40%	17.34	79.34%	17.03
manual agreement	79.41%	18.90	81.16%	17.82	82.18%	16.54

Table 4.11: *A comparison of the overlap rates obtainable by the baseline and refined system in relation to manual agreement.*

4.4 CONCLUSIONS

The effectiveness of the training procedure is a crucial factor when training a speaker-specific set of HMMs for phonetic segmentation. When combining an effective training procedure with parameters that are suitable given the properties of the segments to be modeled, the results obtainable are comparable with what can be expected from an expensive and time consuming manual alignment process and it is plausible that the automatic alignments obtained here are more suited to building systems than alignments resulting from a manual procedure involving a number of inexperienced individuals.

In conclusion, a few suggestions are made based on the observations made in this chapter:

- A flat start model initialisation procedure is less accurate than initialisation with manual alignments. However when flat start initialisation is used, employing a pitch-synchronous feature extraction scheme can significantly improve alignment results.
- The minimum segment length imposed by the model topology and feature resolution combination is a concern and should in practice be limited to approximately 25ms or less.
- Decreasing the step size beyond 5ms does not seem to improve alignment quality (this in fact causes a significant increase in the observed overlap rate variance). The reliable calculation of delta and acceleration coefficients for small increments might be a factor affecting these results.
- When bootstrapping models as suggested in this chapter, the use of 5-state left-to-right tied-state triphone models with single component GMM state emission densities based on 39-

dimensional MFCCs extracted every 5ms with a window size of 10ms seem to be a reasonable choice in the given context.

CHAPTER FIVE

EXPLICIT PHONETIC BOUNDARY PLACEMENTS

Defining exact boundaries between phonetic segments in speech is difficult, especially in those contexts where co-articulation between neighbouring phones renders boundary definition somewhat ambiguous. Nevertheless, for the purposes of spoken language research and system development, a pragmatic approach is necessary in order to define such boundaries as accurately and consistently as possible. Research into the development of corpus-based text-to-speech systems has suggested that consistency (in addition to accuracy) of boundary placements is an important factor when considering the eventual quality of these systems (Clark *et al.*, 2007; Makashay *et al.*, 2000).

Most early development of speech corpora involved manual effort by language or phonetics experts with a significant amount of experience in identifying phonetic segments from visual and auditory information. This reliance on expert human involvement has endured, despite advances in speech recognition and machine learning techniques applied to automating this task. As much is evident when one considers that high quality corpora are still manually checked by such individuals (Pitt *et al.*, 2005). In large resource collection efforts the development of annotated corpora has typically been realised by the collaboration of a large number of trained individuals. The collaboration of multiple individuals is essential in order to complete the sizable task of manually verifying the quality of phonetic alignments within acceptable time-frames, and to have reliable methods of quality assurance. Due to the ambiguities which exist at phonetic transitions, it is common to define protocols for the placement of phonetic boundaries based on broad phonetic class categories in order to ensure the consistency of the end result across different individuals (Cole *et al.*, 1998; Pitt *et al.*, 2005).

The application of Hidden Markov Models (HMMs) to phonetic segmentation can be likened to the first stage of the expert procedure (described in Section 2.3) and in cases where such models are sufficiently trained, this leads to boundary placements which for the most part are fairly close to the "ideal" locations. This is especially the case when manually segmented data exists with which to bootstrap the process involved in training HMMs (as was shown in the previous chapter). Nevertheless, a

large amount of research has been done on further reducing the discrepancies between HMM based and manually obtained boundaries (i.e. "boundary refinement") (Toledano *et al.*, 2003; Sethy and Narayanan, 2002; Kim and Conkie, 2002). This has been justified by the observation that manually segmented and refined automated methods usually result in better quality synthesis when compared to baseline methods (Adell *et al.*, 2005; Saito, 1998). Although such an automated process is generally termed "boundary refinement", we argue from a pragmatic point of view (where we accept some inherent uncertainty in the process identifying phone boundaries and are interested in segmentation results for the purposes of building spoken language systems) that this process is conceptually more accurately seen as the automatic application of a boundary placement protocol. Considered in this light, the majority of research focusing on reducing the discrepancies between manual and automated alignments are essentially trying to automatically apply the same alignment protocol involved during the manual segmentation process. Research of this nature has generally followed two approaches, focusing on two aspects of the problem:

1. Implicit learning of the conventions/protocol through the application of statistical machine learning techniques relying on samples of manually segmented data, focusing on various statistical models (e.g. GMMs / SVMs and HMMs) and data fusion techniques based on the same features used in general speech recognition systems (e.g. MFCCs).

2. Determining properties/features of the speech data which can be used to produce phone boundary candidates that coincide with manually placed boundaries.

In the given context (limited data, with the goal of building TTS systems), both the applicability and feasibility of the above techniques and findings need to be carefully considered. Firstly, considering the findings in (Makashay *et al.*, 2000) that more consistent alignments (even if considered inherently less accurate) may result in better quality synthesis, combined with the accuracies obtained in the previous chapter, which by virtue of the procedure are obtained in a self-consistent way, it is prudent to consider the further refinement of boundary placements (or indeed the protocol governing such placements) as a process that should be tailored to the particular target application. Such a process should be applied consistently and the outcome should be measured in the context of the application (e.g. perceptual experiments in the case of TTS). Thus, although carefully manually aligned segments generally result in better quality concatenative TTS systems for example which suggests that the underlying boundary placement conventions/protocol is aligned to some degree with requirements of this application, ideally one would like to develop a refinement framework that can be tuned to the target application, i.e. where one can experiment with the refinement protocol.

With this goal, relying on the implicit learning of "new" protocols based on manual interpretations thereof is an inflexible approach considering that large amounts of labeled data are required for this approach to be successful. The identification of boundary candidates identifying suitable phone boundaries is a more promising avenue of research toward a flexible refinement framework; however the motivation for selecting certain criteria on which to base boundary identifying candidates should

ideally be motivated by the target application. This in contrast with boundary candidates based on interesting analyses of the speech signal where the effect of such properties on the target application quality might be difficult to predict (such as (Almpanidis and Kotropoulos, 2008)). Furthermore, the feasibility of machine learning techniques (such as described in point 1 above) is limited in the context of developing corpora toward building systems for languages where resources and expertise are scarce. This is the case for two primary reasons:

1. Corpora are designed minimally in order to minimise effort in text selection (it is difficult to find reliable electronic texts for these languages) and expertise required during recording and annotation. This results in corpora where some phonetic contexts simply do not have sufficient coverage in order to train adequate acoustic models based on general features.

2. No manually checked corpora pre-exist in most of the languages of the developing world, because of a lack of skilled persons to perform such tasks. Corpora which are hand checked are small and have mostly been produced by persons with limited background and training (see chapter 2).

For the purposes of developing relevantly annotated corpora with the goal of building high quality spoken language systems in the current context, we investigate in this chapter, the possibility of automated extraction and application of acoustic cues to identifying phonetic transitions selected by considering features used by human transcribers during such a refinement process. The explicit identification of reliable acoustic cues of this nature would present the following opportunities toward the development of an automated refinement stage:

- Boundary candidates obtained in this way can serve as an independent point of reference for judging the quality of alignments (whether automatically or manually obtained).

- These boundary candidates can be integrated into an automated procedure in order to refine boundary placements or improve the quality of training acoustic models, taking into consideration a specific protocol with the end goal of the segmented corpus in mind.

In the following sections we present an initial analysis of the effectiveness of various cues for detecting phonetic events in different contexts, in order to determine the feasibility and potential impact of applying this information.

5.1 ACOUSTIC FEATURES

Typical protocols incorporate practical guidelines for the identification of phonetic boundaries based on acoustic cues exhibited by various features that can be extracted or calculated and displayed. This includes the signal energy, estimated fundamental frequency, periodicity (voicing), extracted formant contours, spectral characteristics and waveform shape. Instructions on boundary placement range from complex and highly conditional (e.g. when transcribing approximants, some suggest observing

the formants, F3 and F4 for "energy reduction") to relatively simple and clearly defined (e.g. place a phonetic boundary "just prior to the burst of energy" when transcribing a stop consonant). Considering this and initial experiments on how reliably one can estimate or extract all of these features, we have concentrated on the following features for the automatic identification of segmentation cues:

- Signal intensity,
- Fundamental frequency (f0),
- Signal envelope, and
- Cepstral distances.

Due to difficulties in reliably determining the number of formants present as well as the exact contours, we have chosen to rely on the use of a "cepstral distance" measure (defined in Section 5.2.3.5) which we hope will identify changes in the formants and general spectral changes with sufficient accuracy.

5.2 EXPERIMENTAL SETUP

We again employed the *Praat* and *HTK* software packages to aid in extracting features from the three corpora described in Table 2.1.

5.2.1 BROAD PHONETIC CLASSES

The most practical and relevant view of phonetic transition contexts for this study is based on broad phonetic categories. All segment labels in the above-mentioned corpora are thus mapped to one of the following labels in accordance with IPA definitions: *affricate*, *approximant*, *click*, *fricative*, *nasal*, *pause*, *stop*, *trill* and *vowel*.

The *pause* label is used both with reference to long pauses (typically only occurring at the beginning and end of utterances) and short segments associated with little signal energy such as glottal stops and closures.

5.2.2 GENERATING BOUNDARY CANDIDATES

In general, boundary candidates are established by firstly calculating or estimating contours for the particular feature and either using this contour directly where applicable or deriving a subsequent contour representing the slope by means of numerical differentiation. After obtaining the appropriate representation, we employ a simple peak detection algorithm in order to generate boundary candidates at specific time instants. We briefly present these methods below.

5.2.2.1 NUMERICAL DIFFERENTIATION

In order to obtain a relatively smooth contour suitable for subsequent peak detection to be effective, we firstly calculate the difference between each sample of the original contour x to obtain a new sequence of differences x_d defined for time instants in-between the original time instants. An odd number N of "difference samples" are framed resulting in a frame x_{df} for each time instant. From this the gradient is determined by first windowing the frame with a simple exponential window function:

$$w[n] = 2^{-|n-\frac{N-1}{2}|}, \tag{5.1}$$

obtaining a frame with weighted differences x_{dfw}:

$$x_{dfw}[n] = x_{df}[n]w[n], \tag{5.2}$$

and calculating the slope at t (the time instant at the center of the frame) by averaging the weighted differences in each frame:

$$x'[t] = \frac{1}{N}\sum_{n=1}^{N} x_{dfw}[n]. \tag{5.3}$$

5.2.2.2 PEAK DETECTION

For detecting local extrema that are of interest during candidate identification, we frame the relevant contour, obtaining an odd number of samples that constitute each frame and simply flag the time instant of the central sample within the frame if it is a global extremum within the frame.

5.2.3 ACOUSTIC CUES

Taking into account the observations in Section 5.1 we experimented with extracting features and identifying candidates automatically. We now briefly describe the particular cues investigated.

5.2.3.1 INTENSITY DYNAMICS

It was observed that many phonetic transitions coincide with changes in the signal intensity and initial experiments indicated that the slope of the intensity contour peaked near potential boundaries. We thus determine intensity values at 5ms intervals and subsequently obtain the derivative and flag the local minima and maxima of the resulting contour (we distinguish between candidates at minima and maxima).

5.2.3.2 WAVEFORM ENVELOPE

Between neighbouring voiced regions such as vowels and nasals, "dips" in the waveform can indicate a phonetic transition. By obtaining the waveform envelope and flagging local minima, such events

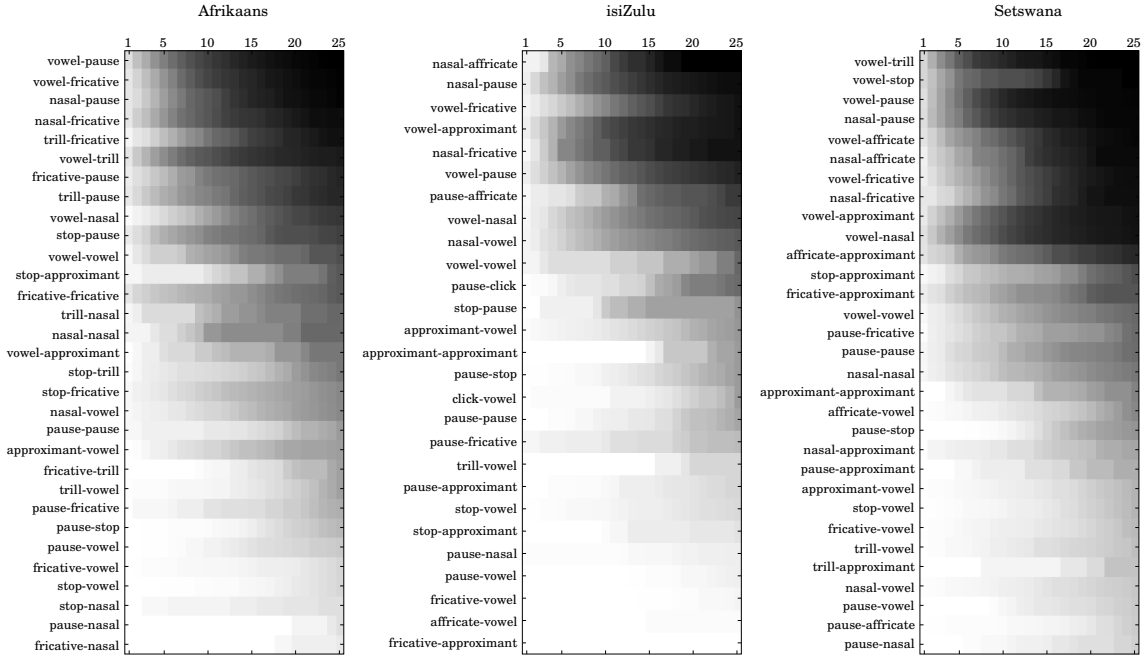

Figure 5.1: *Detection rates: for each phonetic transition context we obtain detection rates for a range of time thresholds (in milliseconds), darker areas represent higher detection rates; this figure represents rates when using the intensity gradient minima cue for each of the languages.*

can be detected. The use of the intensity contour directly was considered, but in cases such as just mentioned, the envelope provides a more pronounced cue.

5.2.3.3 VOICING

By means of a pitch analysis in the frequency range 75Hz to 600Hz, one obtains regions that have a strong periodic component which can be identified as voiced regions. By distinguishing between periodic and aperiodic regions one can place boundary candidates between neighbouring regions in the hope of detecting transitions between voiced and unvoiced segments.

5.2.3.4 FUNDAMENTAL FREQUENCY DYNAMICS

It has been noted that there exists structure within the f0 contour which can be used to identify phonemic events (Saito, 1998). We attempt to detect these events by employing the *Praat* pitch detection algorithm (Boersma, 1993) in the 75Hz to 600Hz range and analysing the slope of the resulting contour.

5.2.3.5 CEPSTRAL DISTANCE

As a measure of spectral difference, which is often used directly via observing the spectrogram or more specifically the changes in formants in order to identify boundary locations manually, we calculated 12 mel frequency cepstral coefficients in 20ms windows with a 2ms time shift. Using this

observation sequence we consider windows of N observations, calculate the average of the first $N-1$ observations and simply calculate the Euclidean distance between the last observation and the average calculated in order to obtain a contour representing a measure of difference between each observation and the prior $N-1$ observations. This contour exhibits peaks at points where the spectral properties change radically.

5.2.4 EVALUATION METRIC

Because boundary candidates will not coincide exactly with reference boundary locations, we consider a reference boundary location to be *detected* when a candidate boundary is located within a certain time threshold of the reference (following a strategy similarly defined in (Estevan *et al.*, 2007)). Subsequently we define an *unambiguous detection* where only detections with at most one candidate within the defined window around the reference are considered. This discredits detections where false alarms are present. For a specific phonetic transition context we can thus define the *unambiguous detection rate* as the ratio between the number of unambiguous detections and the number of occurrences for each context.

5.3 RESULTS

By analysing the detection rates for various cues and phonetic contexts over a range of time thresholds, it is possible to obtain a detailed picture of the success of each cue based on phonetic context (see Figure 5.1 for an example). To investigate the detection rates for individual phonetic contexts, we have to evaluate a range of time thresholds instead of one common threshold (such as 20ms, which is often used), because of the relative durations of phones (e.g. stop phones often have average durations of less than 20ms).

In the subsequent sections we present quantitative results obtained when applying the cues described in the previous sections.

5.3.1 TRANSITION DETECTION: COVERAGE

To measure the utility of each cue, the number of detections as a percentage of the total number of transitions is determined. This is done by firstly distinguishing contexts which are deemed successfully detected in general (it was decided that any transition context with detection rates in excess of 70% would be considered), after which detections are summed for these contexts. The results of this process are presented in Table 5.1.

By using the same notion of successfully detected context, it is also interesting to note the combined transition coverage by the complete set of cues. Figure 5.2 shows the cumulative coverage when the total occurrences for successfully detected phonetic contexts by each cue are added in turn.

Cue	Afrikaans	isiZulu	Setswana
Intensity gradient maxima	39.8%	49.1%	38.1%
Intensity gradient minima	36.4%	28.9%	37.4%
Cepstral difference	32.3%	53.5%	35.2%
Waveform envelope minima	36.9%	33.0%	52.8%
Voicing	4.4%	5.8%	37.5%
F0 gradient extrema	3.6%	10.0%	17.9%

Table 5.1: *Cue significance: the percentages reflect the fraction of all phonetic transitions which are successfully detected by each of the listed cues; only transition contexts for which at least 70% detection is achieved are included in these counts.*

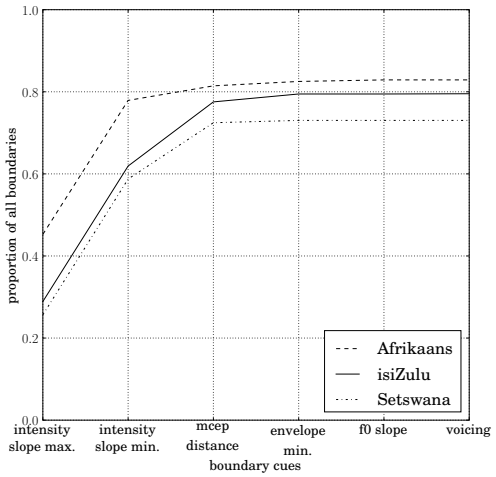

Figure 5.2: *Coverage: the graphs represent the fraction of all phonetic transitions when the number of occurrences of successfully detected transition contexts are accumulated for each language.*

5.3.2 PROBLEMATIC CONTEXTS

By differencing the set of contexts that are successfully detected with the complete set, the set of contexts which are least successfully detected is obtained (listed in Table 5.2). The sets obtained are not surprising considering most of the contexts listed are generally found to be relatively ambiguous (e.g. approximant-vowel transitions) and difficult to distinguish even by manual transcribers (see Figure 2.2). Some of the contexts listed here are also relatively short in duration which suggests that the candidate generation methods used might not be well suited to these conditions.

The relative importance of refining the transition contexts investigated in this chapter is considered in Table 5.3 where the RMSE between the best set of alignments from the previous chapter (see Table 4.11) and the reference alignments are shown for each individual context.

5.4 CONCLUSION

In this chapter we demonstrated the possibility of generating phonetic boundary candidates based on specific acoustic cues that were extracted for three different languages. We showed that it is possible

Afrikaans:
stop-fricative, stop-trill, stop-pause, vowel-nasal, trill-approximant, fricative-pause, approximant-vowel, trill-stop, nasal-nasal, vowel-approximant, fricative-fricative
isiZulu:
pause-affricate, stop-approximant, approximant-pause, affricate-pause, approximant-vowel, stop-vowel, vowel-vowel
Setswana:
pause-affricate, stop-approximant, trill-pause, trill-approximant, approximant-pause, nasal-trill, trill-trill, stop-stop, affricate-pause, approximant-vowel, affricate-affricate, stop-vowel, fricative-nasal, vowel-vowel, pause-trill, fricative-fricative

Table 5.2: *Problematic transition contexts: the contexts listed here were not successfully detected by any of the cues investigated.*

to detect actual boundary positions to a large degree (especially in contexts where the specific cue is relevant from the perspective of speech production).

Although each cue had specific contexts where it outperformed others, the most significant cues were based on the intensity contour and cepstral distance. The fundamental frequency proved to be less successful than expected (based on (Saito, 1998)), but this can probably be attributed to the nature of the reference TTS corpora where the tone is kept more constant than in purely natural speech. Another interesting observation is that the voicing cue worked reasonably well for the female voice but poorly for the male voices, based on these results one should probably carefully consider the exact pitch range of the specific voice before attempting to use this cue.

The problematic contexts remaining seem to be either acoustically ambiguous (e.g. approximant-vowel boundaries cannot be easily distinguished by spectral properties or by observing the waveform. This is also evident from the manual agreement statistics given in Figure 2.2.) or present cases where our method of candidate generation fails. Segments with very short durations can cause the peak detection method or averaging process set up for the average case to miss detections and particularly the cepstral distance measure proposed would also be more effective for longer segments. Future work in detecting the remaining transitions might involve more sophisticated candidate generation or the application of more appropriate features (formant contours might prove successful).

The identification of boundary candidates presented here will allow us to improve the quality of the alignment process automatically. This can be done by defining a protocol similar to protocols designed to allow consistency between multiple human transcribers and using this directly or integrating candidates into training procedures in order to refine models with respect to precise boundary placements. Another useful application would be to flag potentially misaligned boundaries during

Afrikaans		isiZulu		Setswana	
transition context	RMSE	transition context	RMSE	transition context	RMSE
trill_stop	88	**approximant_pause**	106	**trill_pause**	108
pause_pause	77	vowel_stop	42	**fricative_fricative**	59
vowel_approximant	64	stop_pause	38	nasal_nasal	48
stop_pause	59	**vowel_vowel**	36	**trill_trill**	44
approximant_vowel	56	trill_vowel	32	**vowel_vowel**	42
trill_approximant	44	vowel_pause	26	**pause_trill**	42
pause_trill	41	vowel_approximant	22	pause_pause	38
trill_trill	39	vowel_trill	22	pause_fricative	37
nasal_nasal	37	nasal_fricative	20	pause_approximant	32
vowel_vowel	36	nasal_vowel	20	**stop_stop**	31
nasal_pause	33	**approximant_vowel**	18	**approximant_pause**	30
trill_pause	31	nasal_pause	18	nasal_approximant	29
pause_nasal	31	affricate_approximant	17	stop_pause	29
vowel_pause	31	pause_pause	16	nasal_pause	26
fricative_pause	26	vowel_nasal	15	**affricate_affricate**	26
fricative_fricative	26	fricative_trill	15	fricative_pause	23
stop_nasal	23	fricative_approximant	15	affricate_approximant	22
vowel_trill	22	pause_vowel	14	trill_vowel	21
stop_approximant	21	**stop_approximant**	14	**trill_approximant**	19
pause_fricative	20	vowel_fricative	14	approximant_approximant	18
fricative_trill	19	approximant_fricative	12	vowel_stop	18
fricative_nasal	19	**stop_vowel**	12	pause_vowel	17
vowel_fricative	17	click_vowel	12	vowel_approximant	16
trill_fricative	17	trill_stop	11	nasal_fricative	16
stop_fricative	16	nasal_affricate	11	vowel_nasal	16
fricative_vowel	16	pause_fricative	11	fricative_approximant	15
pause_vowel	16	nasal_approximant	11	vowel_pause	15
vowel_stop	15	vowel_affricate	11	**approximant_vowel**	15
stop_trill	13	fricative_vowel	10	pause_nasal	14
vowel_nasal	13	fricative_pause	10	**stop_approximant**	14
pause_approximant	13	pause_approximant	10	nasal_stop	13
pause_stop	13	pause_click	10	vowel_affricate	13
trill_vowel	13	stop_fricative	10	nasal_vowel	12
trill_nasal	12	pause_nasal	10	affricate_vowel	11
nasal_vowel	10	**pause_affricate**	9	**pause_affricate**	11
fricative_stop	10	click_approximant	8	vowel_fricative	10
nasal_fricative	10	affricate_vowel	8	fricative_vowel	10
fricative_approximant	8	pause_stop	7	nasal_affricate	9
stop_vowel	8	nasal_stop	7	vowel_trill	9
nasal_trill	8	nasal_click	7	pause_stop	9
nasal_approximant	8	stop_trill	6	**stop_vowel**	8
stop_stop	6	approximant_approximant	6	**affricate_pause**	7
nasal_stop	5	**affricate_pause**	4	**nasal_trill**	4
				fricative_nasal	1

Table 5.3: *RMSE (ms) between the best HMM-based system and manual refinements for each transition context. Contexts which are not successfully detected (Table 5.2) are shown in boldface.*

quality control of manually or automatically segmented corpora.

An important observation is that boundary refinement based on these candidates can be done automatically and with the target use in mind. This presents opportunity for further research questions relating to text-to-speech synthesis quality when relying on certain acoustic cues to define boundaries. Important acoustic properties relating to speech parametrisation used for speech synthesis should also be explored, e.g. when employing the Harmonics Plus Noise Model (Stylianou, 2001), the maximum voiced frequency contour might prove relevant when performing segmentation.

CHAPTER SIX

CONCLUSIONS AND FUTURE WORK

In this work the problem of accurate automatic phonetic speech segmentation was investigated in the context of developing annotated speaker-specific corpora specifically for application in TTS systems in scarce resourced environments. The following section summarises work presented along with insights gained during this study. This is followed by a section discussing automated corpus development based on the work presented here and lastly a brief proposal regarding future work. Discussions on the feasibility of some boundary refinement techniques investigated during this work can be found in Appendix A where preliminary results are presented.

6.1 AUTOMATIC SPEECH SEGMENTATION WITH LIMITED DATA

Based on a literature review on the topic of automatic phonetic segmentation, a number of research questions relevant in the context of segmentation in a resource scarce environment were identified (Section 1.3). Towards answering these questions, a high level framework (Figure 2.3) and suitable measures of success considering both segment isolation and phonetic boundary placement (Section 2.1) were adopted.

In order to investigate segmentation in the given context within the framework and using the measures adopted, three single-speaker speech corpora in different languages (and typical of the scenario of interest) were employed as reference data. These reference sets were evaluated and the following conclusions were made:

- The level of agreement between transcribers involved in labeling these corpora is somewhat lower than what is typically found in other similar studies, and
- Certain phone types consistently result in lower overlap rates between manual transcribers.

In Chapter 3 the establishment of a baseline segmentation system was investigated by implementing two of the predominant approaches encountered in the literature. These two systems were compared based on empirical results by each system measured against the reference data sets. The insights gained through these experiments follows:

- HMM-based segmentation is more convenient, robust and accurate than DTW when aligning data in new languages for all practical scenarios.
- When aligning data HMMs perform relatively robustly even with very few utterances or phone occurrences, however extensive sparsity of phone occurrences over a number of phone categories should be avoided.
- Neither of the baseline systems implemented achieve accuracies comparable with manual alignments.

Based on these results, improvements to the alignment accuracy within a conventional HMM-based framework (see Figure 3.1) were investigated. Our research focused on the investigation of the *feature extraction*, *model initialisation*, *model description* and *training* components of this framework. The observations made in this chapter resulted in the following conclusions:

- The use of a flat start model initialisation procedure is unlikely to result in acceptable alignment results over all broad phone categories. In general, it results in good alignments for longer segments such as vowels and fricatives and less precise results for shorter segments such as plosives and closures. This problem can be overcome to a degree by using a pitch-synchronous feature extraction scheme.
- In practice, when training and applying HMMs on the same speaker specific set of data for segmentation, it is possible to achieve a significant improvement in alignment accuracy by initialising individual models with a minimal set of labeled bootstrap utterances.
- The initialisation of broad categories of models yields a similar improvement in alignment accuracy over the flat start procedure and can be realised by using a large multi-speaker corpus in a different language (such as the TIMIT corpus) via a relatively simple phone category-based mapping scheme.
- Alignment accuracy in a system such as investigated here can be seen as a trade-off between the system output resolution and modeling precision. In Chapter 4, system parameters resulting in increased alignment accuracy were established.
- The refinements introduced in Chapter 4 result in a system capable of accurate and robust alignments (including the explicit alignment of very short segments such as closures) comparable with what has been achieved by manual transcribers in this context.

Table 6.1 gives an overview of the progress made towards accurate and robust segmentation by comparing the results achieved at various stages of the investigation. As can be seen here, the alignment results achieved after the work presented in Chapter 4 are of consistency comparable to the inter-transcriber results and are at this stage possibly even more appropriate when constructing a concatenative TTS system than the reference alignments, considering the conclusions by (Makashay *et al.*, 2000).

Language	Boundary comparisons				OR	
	< 5ms	< 10ms	< 20ms	RMSE	μ	σ
DTW baseline(Section 3.3.1.2)						
Afrikaans	12.05%	25.91%	47.87%	127.32ms	54.37%	26.69
isiZulu	14.74%	30.57%	53.58%	91.73ms	62.92%	26.09
Setswana	12.15%	25.08%	48.66%	185.76ms	55.27%	26.91
HMM baseline (Section 3.3.1.1)						
Afrikaans	20.22%	39.25%	63.81%	26.90ms	62.17%	24.18
isiZulu	16.71%	35.56%	62.14%	27.25ms	66.26%	21.80
Setswana	14.83%	30.55%	62.29%	28.66ms	59.78%	25.08
HMM with high time-resolution (Section 4.2.2.2)						
Afrikaans	21.07%	38.77%	71.52%	32.57ms	63.42%	24.12
isiZulu	27.07%	45.67%	69.37%	27.27ms	69.90%	22.25
Setswana	19.87%	40.38%	71.15%	26.02ms	64.41%	22.90
HMM with pitch-synchronous features (Section 4.2.2.3)						
Afrikaans	34.82%	58.79%	79.51%	29.93ms	70.35%	22.35
isiZulu	29.02%	50.93%	75.22%	27.44ms	71.28%	21.38
Setswana	24.53%	47.62%	70.42%	31.34ms	64.18%	24.45
HMM with minimally bootstrapped models (Section 4.2.3.1)						
Afrikaans	43.19%	66.47%	85.41%	23.69ms	75.93%	19.81
isiZulu	39.25%	62.35%	84.95%	19.37ms	78.43%	18.86
Setswana	42.30%	68.94%	87.70%	18.26ms	77.74%	18.18
HMM with 5-state bootstrapped models (Section 4.2.3.4)						
Afrikaans	45.13%	69.01%	86.91%	21.83ms	77.15%	19.15
isiZulu	42.57%	66.24%	86.51%	17.65ms	80.40%	17.34
Setswana	45.84%	71.52%	88.81%	16.50ms	79.34%	17.03
Manual agreement (Section 2.2.1)						
Afrikaans	54.58%	73.35%	88.84%	16.40ms	79.41%	18.90
isiZulu	49.33%	74.35%	89.49%	17.62ms	81.16%	17.82
Setswana	58.05%	77.85%	90.64%	17.36ms	82.18%	16.54

Table 6.1: *Summary of the progress made in improving alignment results.*

In Chapter 5 we discussed the problem of local refinement with respect to alignment for the purpose of system development and consequently investigated the feasibility of automatically extracting acoustic cues identifying phone transitions. Here the following conclusions were made:

- It is possible to a identify a large proportion (around 80%) of phone boundary types by automatically extracting and applying acoustic features similar to features used by manual transcribers.

- Features most effective in predicting phone boundaries involve the intensity gradient and a measure of the spectral variation at each speech frame compared to a number of preceding frames.

- Some phone transitions are not well discovered using the acoustic cues investigated. However, this could also be due to inconsistent placements for these categories in the reference sets (compare Table 5.2 and Figure 2.2).

6.2 AUTOMATED TTS CORPUS DEVELOPMENT

Corpus-based approaches to the development of TTS systems rely on the recording and annotation of speech corpora in order to enable the creation of appropriate speech signal representations which can be used to synthesise new signals representing unique input utterances. Given such annotated corpora, various techniques exist in order to extract and represent acoustic units necessary for synthesis (e.g. the HMM-based approach used in the HTS system (Tokuda *et al.*, 2000)). However in the case of a concatenative unit-selection approach, synthesis units (e.g. diphones or halfphones) are constructed by extracting the acoustic signal directly from the corpus using the phonetic boundary alignments considered in this work. Such units are then generally used directly with little or no additional signal processing applied in order to adapt the acoustic properties to suit particular contexts during synthesis, rather relying on the selection of the most appropriate unit from the corpus, based on cost functions aimed at determining its acoustic suitability. This has the advantage of resulting in a speech signal with natural sounding quality (due to the absence of signal distortion associated with extensive signal processing) when appropriate units exist within the corpus. Unfortunately acceptably covering acoustic unit variation present in expressive speech in this way requires a large number of recordings read from well designed text. Thus in most practical scenarios (even for corpora an order of magnitude larger than those considered here), variations in the speech recordings are carefully controlled in order to limit and reduce unnecessary variation. This is an even more painstaking and important process when relying on very small corpora such as presented in Section 2.2. However if one can successfully manage this process then it is possible to develop systems with a high level of intelligibility and even naturalness using the unit-selection approach (Louw *et al.*, 2006). Factors which most notably affect synthesis quality include the following:

- Alignment accuracy: Ensuring consistency of alignments and limiting the number of gross errors existing in the corpus is of greatest concern once a relatively accurate alignment process is established.

- Acoustic suitability: The variation in the recorded voice needs to be limited and this specifically involves the following two properties:

    1. Speech rate: The rate at which the voice talent speaks during recording sessions determines the relative durations of acoustic segments and should be kept as constant as

possible so that the range within which segment durations vary is not excessive.

2. Pitch variation: Similarly to the speech rate, it is preferable to limit the range within which the voice pitch varies in order to increase the likelihood of segments being acceptable in different contexts.

In the next two sections we briefly discuss how these two factors can be considered during corpus development while minimising manual effort.

6.2.1 ALIGNMENT ACCURACY

In Chapter 4 we concluded that alignments resulting from an appropriately applied HMM-based segmentation system are of sufficient quality to construct an acceptably intelligible concatenative TTS system without extensive manual verification. At this level of accuracy, the intelligibility of resulting TTS systems are only severely compromised (with regards to corpus deficiencies) due to the presence of gross alignment errors or missing units. Fortunately, due to the relative robustness of the alignment process (demonstrated throughout this study and specifically in Section 3.4.2), the possible existence of such gross errors are only a practical concern when mismatches exist between the predicted and realised pronunciations during the recording session. The recording process is usually carefully executed and thus mismatches are generally avoided, with the possible exception of slight pronunciation variations due to commonly accepted or dialectal pronunciation variants. Automatic methods can compensate for these problems relatively easily during the alignment process via a re-alignment process where the HMM models are used at some stage during the training process to recognise possible variations (in a constrained fashion), if the possible variations can be predicted prior to the alignment process. An example of this approach is described in (Clark *et al.*, 2006) where vowel reductions and the insertion of short pauses are automatically managed.

Due to the large amount of variation in duration and acoustic variability of segments of different types, it is difficult in practice to identify even gross misalignments through duration or acoustic score outliers. However this information can to some degree be successfully employed within specific phone categories (e.g. plosives): specifically, using durational qualities within these categories such as the fact that plosives are generally relatively short in duration while vowels are generally longer.

While the HMM-based alignment process achieves robust results even with limited training data it was concluded in Chapter 3 that the sparsity of phone occurrences is a significant factor when considering alignment accuracy. For the case of isiZulu where a number of plosives and clicks can also be aspirated in some contexts a marginal improvement in the results from a flat start can be obtained when aspirated and non-aspirated phones are grouped together. Thus, the complexity of the phone set should be evaluated in the interest of alignment accuracy. Such simplifications do however become less effective when bootstrapping is performed as is done in Chapter 4.

6.2.2 ACOUSTIC SUITABILITY

As mentioned above, a large amount of manual effort is required in order to ensure the relative consistency of recordings with regards to acoustic properties. This can be a difficult and time consuming process involving the analysis of recordings using visual representations and other aids. However, once a speech signal has been phonetically annotated, factors pertaining to acoustic suitability can be measured to some degree. This form of automated analysis and quality control is usually done as a pre-processing stage before constructing the unit catalogue through the removal of durational and other acoustic outliers from the set of segments in the corpus. The results in this work suggest that it is unnecessary to postpone such a process until the completion of the recording process, due to the possibility of accurate alignment results despite little training data. Thus one could design the recording process in such a way that, once a subset of utterances have been recorded, these recordings are immediately aligned and automatically analysed in order to determine which utterances need to be considered for re-recording before continuing the process. When developing corpora where the utterances are carefully and minimally designed for unit coverage, it is often a significant problem when even a small number of utterances or segments are unusable. This makes an analysis stage as part of the recording process a potentially valuable option in the context. Based on the results obtained here such a procedure could be implemented for recording subsets of around 100-150 utterances (approximately 40-60 minutes of audio). These subsets can be used to automatically analyse the properties of the recordings, including speech rate, tone variation and changes in the quality of the voice by performing phonetic alignment and comparing duration, fundamental frequency and acoustic scores per segment. In this way potential factors impacting synthesis quality such as excessive speech rate and tone variation can be identified efficiently during the recording process.

6.3 CONCLUSION AND FUTURE WORK

During this work an automated, accurate, consistent and robust phonetic speech segmentation system suitable for the rapid development of small prototype TTS corpora in new languages with minimal ideal resources was developed. In addition to this, initial work was done on further refinement of alignments and practical suggestions made towards an efficient corpus development strategy.

Future work should involve:

- Implementing and evaluating the corpus development strategy suggested in Section 6.2.
- Developing a refinement and/or quality assurance process based on the findings in Chapter 5.
- Applying the techniques developed here to build TTS systems and further evaluating these techniques in the context of such systems via perceptual experiments.

These developments will make it possible to further automate the development of TTS systems. Even in the absence of such automation, however, the results of the current research demonstrate how

intelligible concatenative TTS systems can be developed rapidly and efficiently for under-resourced languages.

APPENDIX A

LOCAL REFINEMENT TECHNIQUES

A.1 FEASIBILITY AND EFFECTIVENESS OF LOCAL REFINEMENT TECHNIQUES

During this work two approaches towards local refinement were implemented:

1. A new local refinement procedure motivated by the success of the "cepstral distance" measure proposed in Section 5.2.3.5 whereby the boundary between segments is refined by calculating the average vector for each segment and comparing the Euclidean distance between frames extracted at a high rate around the current boundary placement to each of the two average vectors, and

2. Explicitly modeling the acoustic properties of a phonetic transition through the use of GMMs describing the acoustic vectors around manual placements as described in (Wang *et al.*, 2004).

Since it would be difficult to demonstrate improvements in alignments through comparison with the reference alignments based on the output of the most successful system at this point, we describe the application of these techniques and briefly present preliminary results based on applying these techniques on alignments by the most successful procedure based on flat start training (i.e. the pitch-synchronous system described in Section 4.2.2.3).

A.1.1 EUCLIDEAN DISTANCE LOCAL REFINEMENT

This method aims to mimic the manual refinement process where a person locally re-adjusts phone boundaries by considering the adjacent segments and determining the point where the signal properties more closely resemble the following segment compared to the properties of the previous segment.

Since relatively few feature vectors (of high dimensionality) are available on which to base this decision, this is implemented by simply calculating the mean vector for sets of vectors considered to be part of a particular segment and comparing vectors around the provisional boundary placement to each of the two average vectors by means of the Euclidean distance. If a clear point is found where a transition in these relative distances occur then the provisional boundary is moved to this point.

For this experiment, refinement features were extracted with 10ms window size every 1ms, only using 13-dimensional MFCCs (i.e. only base features without delta or acceleration coefficients). In Table A.1 we compare the results when comparing the refined alignments with the base and best HMM-based alignments (refer to Table 6.1).

Language	Base			Re-aligned			Best HMM		
	μ OR	σ OR	RMSE	μ OR	σ OR	RMSE	μ OR	σ OR	RMSE
Afrikaans	70.35%	22.35	29.93ms	71.62%	23.07	29.82ms	77.15%	19.15	21.83ms
isiZulu	71.28%	21.38	27.44ms	74.55%	22.78	24.75ms	80.40%	17.34	17.65ms
Setswana	64.18%	24.45	31.34ms	67.96%	27.89	31.20ms	79.34%	17.03	16.50ms

Table A.1: *Summary of the alignment results obtained by the Euclidean distance local refinement method compared to unrefined and best HMM-based alignments.*

Using this method it is clearly possible to get an improvement in alignment agreement with the manual transcriptions and Figure A.1 shows that in all cases fine placement accuracy is increased with little if any increase in larger discrepancies. The well trained HMM-based procedure however, remains the most accurate alternative.

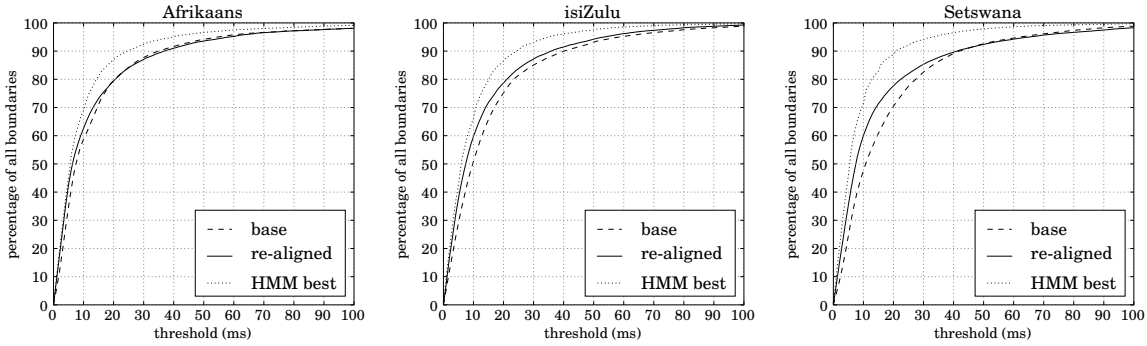

Figure A.1: *A comparison of the boundary accuracy curves obtainable by the base and refined system in relation to manual agreement compared to the best HMM-based results.*

A.1.2 REFINEMENT BY BOUNDARY MODEL

The number of unique phonetic transitions is potentially n^2 where n is the number of unique phones. For this reason and the fact that the signal characteristics at such transitions are variable in nature, modeling transitions explicitly is usually data intensive. In (Wang *et al.*, 2004) a method of representing such transitions by concatenating non-overlapping frames of features into a super-vector is described. Such super-vectors are then used to estimate GMMs describing phonetic transitions after

transitions are clustered into similar categories using a Classification and Regression Tree (the use of GMMs and HMMs to model phonetic transitions was also described in (Sethy and Narayanan, 2002)).

In the context described here it is unlikely that models will be sufficiently well estimated from the limited data available (even when pooling such transitions into similar categories). We thus experimented with boundary models extracted from the TIMIT corpus, constructing models for transitions between broad classes (applying the same mapping scheme as successfully employed in Section 4.2.3.2), in order to be able to apply such models to our native data.

When constructing models of this nature there are a number of parameters that need to be considered, including (a) the number of frames describing the boundary, (b) window and step sizes, (c) complexity of the mixture model and (d) the nature of the features. For our initial investigation we experimented with these parameters and subsequently features similar in nature to those suggested in (Wang *et al.*, 2004) but with fewer coefficients (13-dimensional MFCCs without any dynamic features): Thus, 5 frames with window size of 20ms, resulting in a 65-dimensional super-vector representing a segment of 110ms around each boundary placement. GMMs with the number of mixture components ranging from 1 to 30 were estimated (using an iterative mixture splitting technique to increase mixtures) based on these features. Existing boundaries are then re-aligned by extracting similar super-vectors within a window around existing boundary placements and choosing the time instant where this feature results in the highest likelihood given the pre-trained model. As in the previous section we applied this technique to boundary placements based on the output from the process described in Section 4.2.2.3. The results are presented in Table A.2.

Language	Base			GMM Re-alignment		
	μ OR	σ OR	RMSE	μ OR	σ OR	RMSE
Afrikaans	70.35%	22.35	29.93ms	62.23%	22.50	37.07ms
isiZulu	71.28%	21.38	27.44ms	67.14%	21.68	32.61ms
Setswana	64.18%	24.45	31.34ms	66.13%	21.59	42.23ms

Table A.2: *Summary of the alignment results obtained by the GMM boundary model-based refinement approach using models trained on the TIMIT corpus, compared to unrefined alignments.*

These results are not encouraging, indicating that re-alignments based on these boundary models result in more variable boundary placements. This could be due to a number of factors including ineffective modeling of transitions pooled into broad categories or mismatches between the acoustic properties exhibited in the training data and target data (e.g. speech rate differences to which this technique would be sensitive). One can conclude from these initial results that the approach followed here is unsuccessful and further analysis would be required to determine reasons for this.

APPENDIX B

PHONETIC DESCRIPTIONS

In this appendix complete phonetic descriptions are provided for each language phone set used in this study. These descriptions are presented in the following section in tables with the following format:

- Column 1: A description of the articulation.
- Column 2: The ASCII-based (American Standard Code for Information Interchange) symbols used in this study.
- Column 3: The corresponding IPA symbols.
- Column 4: Mappings to the “radio” phone set used by American English voices in the Festival Speech Synthesis System. These mappings were developed (for the experiment in Section 3.3.1.2) to be perceptually as close as possible to the the original sounds. In some cases this required practical compromises (e.g. the isiZulu /ŋ/ phone had to be mapped to /n/ due to missing diphone units in the English synthesiser for the contexts required in isiZulu).
- Column 5: Broad category mappings used in Section 4.2.3.2. These mappings were also based on perceptual similarities with the requirement that all defined categories exist in the phone set used by the TIMIT corpus and are named accordingly. This results in categories such as “plosives” including the click phones from isiZulu and “approximant” containing the trills occurring in especially Afrikaans and Setswana.

B.1 PHONETIC DEFINITIONS AND MAPPINGS

B.1.1 AFRIKAANS

Description	ASCII	IPA	Mapping in Section 3.3.1.2	Categorisation used in Section 4.2.3.2
Vowels				
low central vowel	a	a	ah	short vowel
low back vowel with duration	aa	ɑː	aa	long vowel
diphthong	aai	ɑːi	oy	diphthong
diphthong	ai	ai	ay	diphthong
mid-low front vowel	E	ɛ	eh	short vowel
mid-low front vowel	e	æ	ae	short vowel
mid-high front vowel with duration	ee	eː	ih	long vowel
diphthong	eeu	eːu	uw	diphthong
diphthong	ei	əi	ey	diphthong
rounded mid-high front vowel with duration	eu	øː	ih	long vowel
central vowel	i	ə	ax	short vowel
high front vowel	ie	i	iy	short vowel
rounded mid-low back vowel	o	ɔ	aw	short vowel
rounded high back vowel	oe	u	uh	short vowel
diphthong	oei	ui	oy	diphthong
rounded mid-high back vowel with duration	oo	oː	uw	long vowel
diphthong	ooi	ɔi	oy	diphthong
diphthong	ou	əu	ow	diphthong
rounded mid-low front vowel	u	œ	ax	short vowel
diphthong	ui	œy	ay	diphthong
rounded high front vowel	uu	y	iy	short vowel

Description	ASCII	IPA	Mapping in Section 3.3.1.2	Categorisation used in Section 4.2.3.2
Consonants				
voiced bilabial plosive	b	b	b	voiced plosive
voiced alveolar plosive	d	d	d	voiced plosive
voiceless labiodental fricative	f	f	f	voiceless fricative
voiced velar plosive	g	g	g	voiced plosive
voiceless glottal fricative	h	h	hh	voiceless fricative
palatal approximant	j	j	y	approximant
voiced post-alveolar fricative	jh	ʒ	zh	voiced fricative
voiceless velar plosive	k	k	k	voiceless plosive
alveolar lateral approximant	l	l	l	approximant
bilabial nasal	m	m	m	nasal
alveolar nasal	n	n	n	nasal
velar nasal	ng	ŋ	ng	nasal
voiceless bilabial plosive	p	p	p	voiceless plosive
alveolar trill	r	r	r	approximant
voiceless alveolar fricative	s	s	s	voiceless fricative
voiceless post-alveolar fricative	sh	ʃ	sh	voiceless fricative
voiceless alveolar plosive	t	t	t	voiceless plosive
voiced labiodental fricative	v	v	v	voiced fricative
voiced labio-velar approximant	w	w	w	approximant
voiceless velar fricative	x	x	f	voiceless fricative
voiced alveolar fricative	z	z	z	voiced fricative
Other				
closure	cl			closure/glottal stop
glottal stop	gs			closure/glottal stop
silence	pau		pau	silence

B.1.2 ISIZULU

Description	ASCII	IPA	Mapping in Section 3.3.1.2	Categorisation used in Section 4.2.3.2
Vowels				
low central vowel	a	a	ah	short vowel
mid-low front vowel	e	ɛ	eh	short vowel
diphthong	ey	əi	ey	diphthong
high front vowel	i	i	iy	short vowel
rounded mid-low back vowel	o	ɔ	aw	short vowel
rounded high back vowel	u	u	uh	short vowel
Consonants				
aspirated voiced bilabial plosive	B	bʰ	b	voiced plosive
voiced bilabial plosive	b	b	b	voiced plosive
aspirated dental click	C	|ʰ	t	voiceless plosive
dental click	c	|	t	voiceless plosive
voiced dental click	cv	|͡	t	voiced plosive
voiced alveolar plosive	d	d	d	voiced plosive
voiced alveolar lateral fricative	dl	ɮ	l	voiced affricate
voiceless labiodental fricative	f	f	f	voiceless fricative
aspirated voiced velar plosive	G	gʰ	g	voiced plosive
voiced velar plosive	g	g	g	voiced plosive
voiceless glottal fricative	h	h	hh	voiceless fricative
voiceless alveolar lateral fricative	hl	ɬ	l	approximant
palatal approximant	j	j	y	approximant
voiced post-alveolar affricate	jh	dʒ	jh	voiced affricate
aspirated voiceless velar plosive	K	kʰ	k	voiceless plosive
voiceless velar plosive	k	k	k	voiceless plosive
voiced velar implosive	kg	ɠ	k	voiced plosive

Description	ASCII	IPA	Mapping in Section 3.3.1.2	Categorisation used in Section 4.2.3.2
Consonants (continued)				
alveolar lateral approximant	l	l	l	approximant
bilabial nasal	m	m	m	nasal
alveolar nasal	n	n	n	nasal
velar nasal	ng	ŋ	n	nasal
palatal nasal	nj	ɲ	n	nasal
aspirated voiceless bilabial plosive	P	p^h	p	voiceless plosive
voiceless bilabial plosive	p	p	p	voiceless plosive
aspirated (post)alveolar click	Q	$!^h$	k	voiceless plosive
(post)alveolar click	q	!	k	voiceless plosive
voiced (post)alveolar click	qv	! ͡	k	voiced plosive
alveolar trill	r	r	r	approximant
voiceless alveolar fricative	s	s	s	voiceless fricative
voiceless post-alveolar fricative	sh	ʃ	sh	voiceless fricative
voiceless alveolar plosive	T	t^h	t	voiceless plosive
voiceless alveolar plosive	t	t	t	voiceless plosive
ejective alveolar affricate	ts	ts’	ch	voiceless affricate
ejective post-alveolar affricate	tsh	tʃ’	ch	voiceless affricate
voiced labiodental fricative	v	v	v	voiced fricative
voiced labio-velar approximant	w	w	w	approximant
aspirated alveolar lateral click	X	$\|\|^h$	k	voiceless plosive
alveolar lateral click	x	‖	k	voiceless plosive
voiced alveolar lateral click	xv	‖ ͡	k	voiced plosive
voiced alveolar fricative	z	z	z	voiced fricative
Other				
closure	cl			closure/glottal stop
glottal stop	gs			closure/glottal stop
silence	pau		pau	silence

B.1.3 SETSWANA

Description	ASCII	IPA	Mapping in Section 3.3.1.2	Categorisation used in Section 4.2.3.2
Vowels				
low central vowel	a	a	ah	short vowel
mid-high front vowel	e	e	eh	short vowel
mid-low front vowel	ea	ɛ	ea	short vowel
high front vowel	i	i	iy	short vowel
rounded mid-high back vowel	o	o	ao	short vowel
rounded mid-low back vowel	oa	ɔ	aw	short vowel
rounded high back vowel	u	u	uh	short vowel
Consonants				
voiced bilabial plosive	b	b	b	voiced plosive
voiced alveolar plosive	d	d	d	voiced plosive
voiced post-alveolar affricate	dj	dʒ	jh	voiced affricate
voiceless labiodental fricative	f	f	f	voiceless fricative
voiceless glottal fricative	h	h	hh	voiceless fricative
palatal approximant	j	j	y	approximant
voiceless velar plosive	k	k	k	voiceless plosive
aspirated voiceless velar plosive	kh	k^h	k	voiceless plosive
alveolar lateral approximant	l	l	l	approximant
bilabial nasal	m	m	m	nasal
alveolar nasal	n	n	n	nasal
velar nasal	ng	ŋ	ng	nasal
palatal nasal	nj	ɲ	n	nasal

Description	ASCII	IPA	Mapping in Section 3.3.1.2	Categorisation used in Section 4.2.3.2
Consonants (continued)				
voiceless bilabial plosive	p	p	p	voiceless plosive
aspirated voiceless bilabial plosive	ph	p^h	p	voiceless plosive
alveolar trill	r	r	r	approximant
voiceless alveolar fricative	s	s	s	voiceless fricative
voiceless post-alveolar fricative	sj	ʃ	sh	voiceless fricative
voiceless alveolar plosive	t	t	t	voiceless plosive
aspirated voiceless alveolar plosive	th	t^h	t	voiceless plosive
ejective voiceless alveolar lateral plosive	tl	tl’	t	voiceless plosive
aspirated voiceless alveolar lateral plosive	tlh	tl^h	t	voiceless plosive
ejective alveolar affricate	ts	ts’	s	voiceless affricate
aspirated alveolar affricate	tsh	ts^h	s	voiceless affricate
ejective post-alveolar affricate	tsj	tʃ’	ch	voiceless affricate
aspirated post-alveolar affricate	tsjh	$tʃ^h$	ch	voiceless affricate
voiced labio-velar approximant	w	w	w	approximant
voiceless velar fricative	x	x	f	voiceless fricative
Other				
closure	cl			closure/glottal stop
glottal stop	gs			closure/glottal stop
silence	SIL		pau	silence

REFERENCES

Adell, J., Bonafonte, A., Gómez, L.A.H. and Castro, M.J. (2005). Comparative study of Automatic Phone Segmentation methods for TTS. In: *Proceedings of ICASSP*, vol. 1, pp. 309–312. Philadelphia, Pennsylvania, USA.

Almpanidis, G. and Kotropoulos, C. (2008). Phonemic segmentation using the generalised Gamma distribution and small sample Bayesian information criterion. *Speech Communication*, vol. 50, no. 1, pp. 38–55.

Andre-Obrecht, R. (1988 January). A New Statistical Approach for the Automatic Segmentation of Continuous Speech Signals. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 1, pp. 29–40.

Barnard, E. and Davel, M. (2006 November). Automatic error detection in alignments for speech synthesis. In: *Proceedings of PRASA*, pp. 53–56. Parys, South Africa.

Black, A.W. and Lenzo, K.A. (2007). *Building Synthetic Voices*.
Available at: `http://www.festvox.org/bsv/`

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In: *Proceedings of the Insitute of Phonetic Sciences*, vol. 17, pp. 97–110. Amsterdam, The Netherlands.

Boersma, P. (2001). *Praat, a system for doing phonetics by computer*. Amsterdam: Glott International.

Clark, R.A.J., Richmond, K. and King, S. (2007). Multisyn: Open-domain unit selection for the Festival speech synthesis system. *Speech Communication*, vol. 49, no. 4, pp. 317–330. ISSN 0167-6393.

Clark, R.A.J., Richmond, K., Strom, V. and King, S. (2006 September). Multisyn voice for the Blizzard Challenge 2006. In: *Proceedings of the Blizzard Challenge Workshop (Interspeech Satellite)*. Pittsburgh, Pennsylvania, USA.

Cole, R., Noel, M. and Noel, V. (1998 December). The CSLU Speaker Recognition Corpus. In: *Proceedings of ICSLP*, pp. 3167–3170. Sydney, Australia.

Estevan, Y.P., Wan, V. and Scharenborg, O. (2007 April). Finding maximum margin segments in speech. In: *Proceedings of ICASSP*, vol. 4, pp. 937–940. Honolulu, Hawai'i, USA.

Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S. and Dahlgren, N.L. (1993). *Darpa Timit: Acoustic-phonetic Continuous Speech Corpus CD-ROM*. US Dept. of Commerce, National Institute of Standards and Technology.

Golipour, L. and O'Shaughnessey, D. (2007 August). A New Approach for Phoneme Segmentation of Speech Signals. In: *Proceedings of INTERSPEECH*, pp. 1933–1936. Antwerp, Belgium.

Gouws, E., Wolvaardt, K., Kleynhans, N. and Barnard, E. (2004 November). Appropriate baseline values for HMM-based speech recognition. In: *Proceedings of PRASA*, pp. 169–172. Grabouw, South Africa.

Jarifi, S., Pastor, D. and Rosec, O. (2008). A fusion approach for automatic speech segmentation of large corpora with application to speech synthesis. *Speech Communication*, vol. 50, no. 1, pp. 67–80.

Kawai, H. and Toda, T. (2004). An evaluation of automatic phone segmentation for concatenative speech synthesis. In: *Proceedings of ICASSP*, vol. 1, pp. 677–680. Montréal, Québec, Canada.

Kim, Y. and Conkie, A. (2002 September). Automatic Segmentation Combining an HMM-Based Approach and Spectral Boundary Correction. In: *Proceedings of ICSLP*, pp. 145–148. Denver, Colorado, USA.

Kominek, J., Bennett, C. and Black, A. (2003 September). Evaluating and Correcting Phoneme Segmentation for Unit Selection Synthesis. In: *Proceedings of EUROSPEECH*, pp. 313–316. Geneva, Switzerland.

Kominek, J. and Black, A.W. (2004 October). A Family-of-Models Approach to HMM-based Segmentation for Unit Selection Speech Synthesis. In: *Proceedings of INTERSPEECH*, pp. 1385–1388. Jeju Island, Korea.

Kuo, J.W. and Wang, H.M. (2006 September). Minimum boundary error training for automatic phonetic segmentation. In: *Proceedings of INTERSPEECH*. Pittsburgh, Pennsylvania, USA.

Ladefoged, P. (1990 September). The Revised International Phonetic Alphabet. *Language*, vol. 66, no. 3, pp. 550–552. ISSN 00978507.

Laureys, T., Demuynck, K., Duchateau, J. and Wambacq, P. (2002). An improved algorithm for the automatic segmentation of speech corpora. In: *Proceedings of the International Conference on Language Resources and Evaluation*, vol. 5, pp. 1564–1567.

Ljolje, A., Hirschberg, J. and van Santen, J.P.H. (1996). Automatic speech segmentation for concatenative inventory selection. *Progress in Speech Synthesis: with 158 illustrations*, p. 305.

Lo, H.Y. and Wang, H.M. (2007 April). Phonetic boundary refinement using support vector machine. In: *Proceedings of ICASSP*, pp. 933–936. Honolulu, Hawai'i, USA.

Louw, J.A., Davel, M. and Barnard, E. (2006). A general-purpose IsiZulu speech synthesizer. *South African journal of African languages*, vol. 2, pp. 1–9.

Makashay, M., Wightman, C., Syrdal, A. and Conkie, A. (2000 October). Perceptual evaluation of automatic segmentation in text-to-speech synthesis. In: *Proceedings of ICSLP*, vol. 2, pp. 431–434. Beijing, China.

Malfrère, F., Deroo, O., Dutoit, T. and Ris, C. (2003). Phonetic alignment: speech synthesis-based vs. viterbi-based. *Speech Communication*, vol. 40, no. 4, pp. 503–515.

Malfrère, F. and Dutoit, T. (1997). High-quality Speech Synthesis for Phonetic Segmentation. In: *Proceedings of EUROSPEECH*, pp. 2631–2634. Rhodes, Greece.

Murrell, B. and Tapamo, J.R. (2008). Heuristics for State Splitting in Hidden Markov Models. In: *Proceedings of PRASA*, pp. 3–8. Cape Town, South Africa.

Ney, H. (1984). The Use of a One-Stage Dynamic Programming Algorithm for Connected Word Recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 263–271.

Park, S., Shin, J. and Kim, N. (2006 September). Automatic speech segmentation with multiple statistical models. In: *Proceedings of INTERSPEECH*, pp. 2066–2069. Pittsburgh, Pennsylvania, USA.

Paulo, S. and Oliveira, L. (2004). *Advances in Natural Language Processing*. Springer Berlin / Heidelberg.

Pitt, M.A., Johnson, K., Hume, E., Kiesling, S. and Raymond, W. (2005). The Buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability. *Speech Communication*, vol. 45, no. 1, pp. 89–95.

Saito, T. (1998 December). On the use of F0 features in automatic segmentation for speech synthesis. In: *Proceedings of ICSLP*, vol. 7, pp. 2839–2842. Sydney, Australia.

Sethy, A. and Narayanan, S. (2002 September). Refined Speech Segmentation for Concatenative Speech Synthesis. In: *Proceedings of ICSLP*, pp. 149–152. Denver, Colorado, USA.

Sharma, M. and Mammone, R. (1996 October). "Blind" speech segmentation: automatic segmentation of speech without linguistic knowledge. In: *Proceedings of ICSLP*, vol. 2, pp. 1237–1240. Philadelphia, Pennsylvania, USA.

Stylianou, Y. (2001). Applying the harmonic plus noise model in concatenative speech synthesis. *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 1, pp. 21–29.

Taylor, P., Black, A.W. and Caley, R. (1998). The Architecture of the Festival Speech Synthesis System. In: *Proceedings of the Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis*, pp. 147–151. ISCA, Jenolan Caves, Blue Mountains, NSW, Australia.

Taylor, P., Caley, R., Black, A.W. and King, S. (1999). Edinburgh speech tools library. *System Documentation Edition*, vol. 1.
Available at: `http://www.cstr.ed.ac.uk/projects/speech_tools/`

Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T. and Kitamura, T. (2000). Speech parameter generation algorithms for HMM-based speech synthesis. In: *Proceedings of ICASSP*, vol. 3, pp. 1315–1318. Istanbul, Turkey.

Toledano, D., Gómez, L.H. and Grande, L. (2003). Automatic Phonetic Segmentation. *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 617–625.

Toledano, D.T., Rodríguez, M. and Escalada, J. (1998 November). Trying to mimic human segmentation of speech using HMM and fuzzy logic post-correction rules. In: *Proceedings of the Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis*, pp. 207–212. ISCA, Jenolan Caves, Blue Mountains, NSW, Australia.

Šarić, Z.M. and Turajlić, S.R. (1995). A new approach to speech segmentation based on the maximum likelihood. *Circuits, Systems, and Signal Processing*, vol. 14, no. 5, pp. 615–632.

Wang, L., Zhao, Y., Chu, M., Zhou, J. and Cao, Z. (2004 May). Refining segmental boundaries for TTS database using fine contextual-dependent boundary models. In: *Proceedings of ICASSP*, vol. 1, pp. 641–644. Montréal, Québec, Canada.

Weide, R.L. (1998). *The CMU pronunciation dictionary, release 0.6*. Carnegie Mellon University.
Available at: `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Veltchev, V. and Woodland, P. (2005). *The HTK Book (for HTK Version 3.3)*. Cambridge University Engineering Department.
Available at: `http://htk.eng.cam.ac.uk/`

Young, S.J., Odell, J.J. and Woodland, P.C. (1994 March). Tree-based state tying for high accuracy acoustic modelling. In: *Proceedings of the Workshop on Human Language Technology*, pp. 307–312. Association for Computational Linguistics, Plainsboro, New Jersey, USA. ISBN 1-55860-357-3.