PROSODIC FEATURES OF IMPERATIVES IN XHOSA: IMPLICATIONS FOR A TEXT-TO-SPEECH SYSTEM

by

Philippa H. Swart

Thesis presented in fulfillment of the requirements for the degree Master of Arts at the University of Stellenbosch

Study leader: Prof. J.C. Roux

Co-study leader: Prof. E.C. Botha

March 2000

Declaration

I, the undersigned, hereby declare that the work contained in this thesis is my own original work and has not previously in its entirety or part been submitted at any university for a degree.

Summary

This study focuses on the prosodic features of imperatives and the role of prosodics in the development of a text-to-speech (TTS) system for Xhosa, an African tone language. The perception of prosody is manifested in suprasegmental features such as fundamental frequency (pitch), intensity (loudness) and duration (length).

Very little experimental research has been done on the prosodic features of any grammatical structures (moods and tenses) in Xhosa, therefore it has not yet been determined how and to what degree the different prosodic features are combined and utilized in the production and perception of Xhosa speech. One such grammatical structure, for which no explicit descriptive phonetic information exists, is the imperative mood expressing commands.

In this study it was shown how the relationship between duration, pitch and loudness, as manifested in the production and perception of Xhosa imperatives could be determined through acoustic analyses and perceptual experiments. An experimental phonetic approach proved to be essential for the acquisition of substantial and reliable prosodic information.

An extensive acoustic analysis was conducted to acquire prosodic information on the production of imperatives by Xhosa mother tongue speakers. Subsequently, various statistical parameters were calculated on the raw acoustic data (i) to establish patterns of significance and (ii) to represent the large amount of numeric data generated, in a compact manner.

A perceptual experiment was conducted to investigate the perception of imperatives. The prosodic parameters that were extracted from the acoustic analysis were applied to synthesize imperatives in different contexts. A novel approach to Xhosa speech synthesis was adopted. Monotonous verbs were recorded by one speaker and the pitch and duration of these words were then manipulated with the TD-PSOLA technique.

Combining the results of the acoustic analysis and the perceptual experiment made it possible to present a prosodic model for the generation of perceptually acceptable imperatives in a practical Xhosa TTS system.

Prosody generation in a natural language processing (NLP) module and its place within the larger framework of text-to-speech synthesis was discussed. It was shown that existing architectures for TTS synthesis would not be appropriate for Xhosa without some adaptation. Hence, a unique architecture was suggested and its possible application subsequently illustrated. Of particular importance was the development of an alternative algorithm for grapheme-to-phoneme conversion.

Keywords: prosody, speech synthesis, speech perception, acoustic analysis, Xhosa

NOTA BENE: This thesis is accompanied by a compact disc (MS Windows format) that contains a digital version of the thesis document as well as sample sound files of the recorded and synthesized data discussed in Chapters 4 and 5.

Opsomming

Hierdie studie fokus op die prodiese eienskappe van imperatiewe en die rol van prosodie in die ontwikkeling van 'n teks-na-spraak-sisteem vir Xhosa, 'n Afrika-toontaal. Die persepsie van prosodie word gemanifesteer in suprasegmentele eienskappe soos fundamentele frekwensie (toonhoogte), intensiteit (luidheid) en duur (lengte).

Weinig eksperimentele navorsing bestaan ten opsigte van die prosodiese eienskappe van enige grammatikale strukture (modus en tyd) in Xhosa. Hoe en tot watter mate die verskillende prosodiese kenmerke gekombineer en gebruik word in die produksie en persepsie van Xhosa-spraak is nog nie duidelik nie. 'n Grammatikale struktuur waarvoor geen eksplisiete deskriptiewe fonetiese inligting bestaan nie, is die van die imperatiewe modus wat bevele uitdruk.

Hierdie studie wys hoe die verhouding tussen duur, toonhoogte en luidheid, soos gemanifesteer in die produksie en persepsie van Xhosa-imperatiewe bepaal kon word deur akoestiese analises en persepsuele eksperimente. Dit het geblyk dat 'n eksperimenteel-fonetiese benadering noodsaaklik is vir die verkryging van sinvolle en betroubare prosodiese inligting.

'n Uitgebreide akoestiese analise is uitgevoer om prosodiese data omtrent die produksie van imperatiewe deur Xhosa-moedertaalsprekers te bekom. Vervolgens is verskeie statistiese analises op die rou akoestiese data uitgevoer om (i) patrone van beduidenheid te bepaal en om (ii) die groot hoeveelheid numeriese data wat gegenereer is meer kompak voor te stel.

'n Persepsuele eksperiment is uitgevoer met die doel om die persepsie van imperatiewe te ondersoek. Die prosodiese parameters soos uit die akoestiese analise bekom, is toegepas in die sintese van bevele in verskillende kontekste. 'n Nuwe benadering tot Xhosa-spraaksintese is gevolg. Monotone werkwoorde is vir een spreker opgeneem en die toonhoogte en duur van hierdie woorde is met TD-PSOLA tegniek gemanipuleer.

'n Kombinasie van akoestiese en persepsuele resultate is aangewend om 'n prosodiese model te ontwikkel vir die sintese van persepsueel aanvaarbare imperatiewe in 'n praktiese Xhosa teks-na-spraaksintetiseerder.

Prosodie-generering in 'n natuurlike taalprosesering-module en die plek daarvan binne die raamwerk van teks-na-spraaksintese is bespreek. Daar is gewys dat bestaande argitekture vir teks-na-spraaksisteme nie sonder sommige aanpassings toepaslik vir Xhosa sal wees nie. Derhalwe is 'n unieke argitektuur gesuggereer en die moontlike toepassing daarvan geïllustreer. Die ontwikkeling van 'n alternatiewe algoritme vir letter-na-klankomsetting was van besondere belang.

Sleutelwoorde: spraaksintese, spraakpersepsie, akoestiese analise, Xhosa

LET WEL: 'n Kompakte skyf (MS Windows formaat) wat 'n digitale weergawe van hierdie dokument bevat, sowel as voorbeelde van opgeneemde en gesintetiseerde spraakleêrs, is ingesluit by hierdie tesis.

Dedicated to my parents and my fiancé, Jan.

Enthusiasm is the mother of effort, and without it nothing great was ever accomplished.

Ralph Waldo Emmerson

Thank you for sharing my enthusiasm.

Acknowledgements

I hereby express my sincere gratitude to the following persons and institutions without whom the completion of this study would not have been possible:

- Professor Justus Roux, chair of the Department of African Languages, University of Stellenbosch, for his inspirational guidance and encouragement throughout;
- Professor Elizabeth Botha of the Department of Electrical and Electronic Engineering, University of Pretoria, for her guidance as co-study leader;
- My parents, Hermann and Philae Swart, for their ongoing love and support;
- Mr. Jan Louw, for his technical expertise, advice and willingness to assist at all times;
- Messrs M. Dlali, S.Dunga, M. Jadezweni, A. Kesse, M. Makomazi, L. Matiwane, X. Mavela, M. Memela, P. Sibula and Z. Witbooi for their contributions as native speakers of Xhosa, towards establishing a corpus of acoustic data;
- Mr. Phumlani Sibula and pupils of the Khayamandi High School who participated in the perceptual experiment and thereby making the research possible;
- Foundation for Research Development (FRD) for financial assistance;
- The Research Unit for Experimental Phonology (RUEPUS) for financial assistance as well as the use of their facilities;
- The Language Laboratory and the computer centre for the humanities (HUMARGA) of the University of Stellenbosch for the use of their facilities.

SOLI DEO GLORIA

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ASL	Analysis/Synthesis Lab
ASR	automatic speech recognition
C	consonant
CD	closure duration
CSL	Computerized Speech Lab
DSP	digital signal processing
F_0	pitch
FL	falling-low
HL	high-low
LL	low-low
Hz	hertz
IPA	international phonetic alphabet / International Phonetic Association
LPC	linear prediction coefficients
LTS	letter-to-sound (rules)
MRT	mean response time
ms	milliseconds
NLP	natural language processing
OC	objectival concord
R	root
TD	total duration
TD-PSOLA	time domain pitch synchronous overlap and add
TP	total pitch
TTS	text-to-speech (system)
V	vowel
VOT	voice onset time

List of Symbols

$\approx$	approximately equal to
Δ	change
Δd	change in duration [s]
Δp	change in pitch [Hz]
χ^2	chi square value
X_{wf}	data vector for the first context of word *w*, feature *f* being compared
Y_{wf}	data vector for the second context of word *w*, feature *f* being compared
D_{wf}	difference data vector for word *w*, feature *f*
d	duration
d_x	duration of syllable *x* [s]
d_{VOT}	duration of voice onset time
e_i	expected frequency
f	feature *f*
$f( )$	function
E_x	absolute loudness of syllable *x*
e_x	relative loudness of syllable *x*
$\left\|\frac{\bar{\mu}}{\sigma}\right\|$	magnitude of the mean divided by the standard deviation
$\bar{\mu}$	mean
$\frac{\bar{\mu}}{\sigma}$	mean divided by the standard deviation
$\tilde{\mu}$	median
p	pitch
p_x	pitch of syllable *x* [Hz]
F_s	sampling frequency
S	significance value according to the Wilcoxon test
σ	standard deviation
Σ	sum
´	high tone
`	low tone
ˆ	falling tone

Chapter One

Orientation

1.1 Introduction and Motivation of Study

This study focuses on the prosodic features of imperatives in Xhosa and the role of prosodics in the development of a text-to-speech system for an African tone language.

The perception of *prosody* is manifested in suprasegmental features such as *fundamental frequency* (pitch), *intensity* (loudness) and *duration* (length). The speaker uses these features to convey important information additional to that conveyed by the segmental composition of the utterance. As mentioned by House (1992:9) "*the perception of prosody has long been recognized to be important and necessary for the perception of spoken language*". Therefore, it is essential that prosodic information be encoded in the production of natural speech.

Prosodic features such as tone and intonation play a crucial part in contributing to the specific meaning of an utterance. This is true for all languages, but even more so for tone languages such as Thai, Chinese and most African Languages (cf. House, 1990:12; Cruttenden, 1986:8 and Maddieson, 1978). This study concerns Xhosa, an African Language of the Nguni subgroup. Xhosa is spoken, as a mother tongue, by 17.5% of the South African population of 37.9 million people (Roux 1998:351). Xhosa is characterized as a tone language. It follows that segmental information alone may not be sufficient, in order to produce natural, comprehensible Xhosa speech. Changes in tone bring forth changes in meaning. For example, Riordan (1969) shows that the word *ithanga* has three different meanings depending on the tonal pattern of the word. Riordan (1969) presents the following data:

(1.1) (a) *íthàngà*[1] 'a pumpkin'

[1] Note that (´) is used to mark a syllable with a high tone; (`) is used to mark a syllable with a low tone and (^) marks a syllable with a falling tone. These tone marks do not carry any additional meaning such as duration or loudness.

(b) *íthàngá* 'a thigh'

(c) *íthângà* 'a cattle-post'

However, the reliability and authenticity of data such as the above are seriously questioned by Roux (1995a). It is Roux's belief that "*although the impression may implicitly and explicitly be created that, at least at surface level, the tonal patterns of two Nguni Languages (Zulu and Xhosa) have been adequately accounted for, this is probably not the case*" (1995a:21). According to Roux (1995a:21), the tonal data that exists "*reveal questionable methods of data acquisition* (as well as) *inherent inconsistencies*". Roux (1995b:197) objects mainly against the fact that "*linguists working within African languages have up to this day been quite complacent to rely almost exclusively on the impressionistic judgements of a 'trained phonetician' in compiling primary data*".

In the case of *ithanga*, for example, it is evident that prosodic and not segmental information contributes to disambiguation. It seems, however, that the role of tone has been over emphasized in the past. The experimental studies of Roux (1995a; 1995b) have shown that pitch might not be the only prosodic feature contributing to the specific meaning of the word. According to Roux (1995b:200):

> "*it should be quite clear that various other prosodic features may play a role in this disambiguation process. The question which of these features are relevant and which are redundant in the communication process is yet to be answered. This phenomenon once more focuses on the unreliable and incomplete nature of tonal descriptions found in this language* (Xhosa)".

Very little experimental research has been done on the prosodic features of *any* grammatical structures (moods and tenses) in Xhosa, therefore it has not yet been determined how and to what degree the different prosodic features are combined and utilized in the production and perception of Xhosa speech.

One such grammatical structure, for which no explicit descriptive phonetic information exists, is the *imperative mood* expressing commands. The only existing information on imperatives is explicit morphological descriptions (cf. Riordan, 1969; Wentzel et al., 1972). Riordan (1969) supplies implicit information on tone patterns by giving a number

of tone marked examples and Davey (1973:29-32) formulates some phonological rules concerning the tonal structure of imperatives. However, these and other works (Wentzel et al., 1972; Westphal, 1967; Claughton, 1983) are not substantiated by explicit phonetic information concerning pitch, loudness and duration. In order to obtain substantial and reliable prosodic information for imperatives, an experimental phonetic approach is required. At the same time, investigation should focus equally on the production and perception process.

Language and Speech Technology

Engineers and speech scientists have been interested in the topic of spoken language interfaces for computers for more than fifty years. Their goal is to build machines with which humans can converse in the same way as they do with one another. To reach this goal would be the ultimate challenge to our understanding of the production and perception processes involved in human speech communication. Today we are dependant on interactive networks that provide easy access to information and services that we use to do our work and conduct our daily affairs, but only those who are literate and have access to computers can benefit from such networks. The necessity for advances in human language technology is most aptly described by Zue and Cole (1997:1):

> "(these advances) *are needed for the average citizen to communicate with networks using natural communication skills using everyday devices, such as telephones and televisions. Without fundamental advances in user-centered interfaces, a large portion of society will be prevented from participating in the age of information, resulting in further stratification of society and tragic loss in human potential*".

Xhosa speakers do not currently benefit from such technological advances as very little research has been done in this area for Xhosa or any other African language for that matter (De Wet & Botha, 1998). As yet, no spoken language computer interfaces such as Xhosa synthesizers or recognition systems are commercially available.

For Xhosa speaking South Africans to take their rightful place and participate in the information age, research needs to be done to develop these systems (cf. research done on the recognition of Xhosa speech by De Wet, 1999). By its nature this research necessitates

a multi-disciplinary approach incorporating the efforts of linguists, engineers and computer scientists.

Prosody in Language and Speech Technology

It is generally agreed that prosody is one of the most critical aspects of synthesis technology to be improved (SPI Lab, 1999). According to Ostendorf (nd.) the additional information provided by prosody will become increasingly important, as systems move towards less constrained and more natural interaction (see also Dutoit, 1993:32). Speech synthesis systems fare well where intelligibility is concerned (the best systems achieve 99% scores), but the naturalness of these systems is judged in the fair-to-good range which does not match the quality and prosody of natural speech (Kamm et al., 1997:270).

Prosodic modeling is seen as one of the most difficult problems in speech synthesis to date. It is still unknown how prosodic parameters interact in fluent speech. This interplay between the various prosodic parameters is described as one of the hottest research topics in the field of speech synthesis (D'Allessandro & Liénard, 1997:173; Dutoit, 1993:32). A number of prosodic models and transcription formalisms have been developed over the years (summary in Dutoit, 1993:33-34) but prosody has been described as "*the most universal and the most language specific characteristic of speech*" (Dutoit, 1993:33). All human languages have some basic prosodic elements in common, but these elements are combined in different ways for different languages. It follows that one prosodic model for one specific language cannot be generalized and used for another language. Each different language requires a unique prosodic model.

Linguists working on the African languages have a dual role to play in the development of speech technology. Firstly, we may work within the frameworks set by international researchers and build upon the existing knowledge to design a unique prosodic model for each African language. Secondly, we may make new contributions to the field. Adopting this dual role, this study aims to build on and contribute to the field of prosody generation in text-to-speech systems in particular.

Speech synthesis may be performed at different levels, but the most general system is that of text-to-speech synthesis. Text-to-speech (TTS) systems have various applications (ref. paragraph 3.6) that could benefit not only the Xhosa speaking community, but all South Africans in one way or another. One such application is that of the support of manpower deployment (Laver, 1994:4284) where there is a need for messages (commands in particular) to be transmitted securely. This could be applied in areas ranging form the battlefield and the police service to emergency service operations.

Although part of this study addresses the architecture for all components of the natural language processing module (NLP), emphasis is placed on the methods of parameter extraction and prosody generation to be applied in a practical Xhosa TTS system. We are convinced that the imperative structure in Xhosa would provide us with an appropriate platform from which to undertake this investigation. Prosodic information for imperatives does, however, not only have important applications in TTS systems; it may also have implications for automatic speech recognition (ASR). Speech based information retrieval systems are required to recognize commands and this necessitates an understanding of how commands are produced and perceived on suprasegmental level. State-of-the-art systems are, however, not implemented in this way as yet.

1.2 Scope and Aims of Investigation

Whilst experimental phonetics form the basis of this investigation, it also aims to employ a methodological-applied approach rather than being merely descriptive in nature. Broadly stated, the primary aim of this study is to determine and describe the role of prosodic features in the production and perception of imperatives in Xhosa.

More specifically, the primary aims of the study are as follows:

- To determine what acoustic features, at segmental and suprasegmental level, Xhosa mother tongue speakers apply to issue commands (*encoding*).
- To determine what segmental and suprasegmental information subjects need in order to perceive commands (*decoding*).

- To determine the relationship between pitch, duration and loudness and the statistical significance of these features in the production of imperatives in Xhosa.
- To determine the relationship between pitch, duration and loudness and the statistical significance of these features in the perception of imperatives in Xhosa.

Information regarding the production of imperatives will be acquired through the acoustic analysis of speech data. It will be shown how prosodic parameters may be extracted. The prosodic parameters will be used to synthesize commands and these stimuli will in turn be presented in a perceptual experiment to determine how mother tongue listeners perceive commands. Both the acoustic analysis and perceptual experiment will be done on a large scale, which requires the automation of analysis and testing procedures. Throughout the study it will be shown how a variety of procedures can be automated so as to enhance the accuracy and reliability of data and results.

The secondary aim of this study is to investigate the specific implications the acoustic and perceptual information, as well as the unique nature of the Xhosa language, might have for the development of a TTS system and in particular, for prosody generation. More specifically, the secondary aim of this study is therefore:

- To formulate a model at segmental and suprasegmental level, which may be implemented in a Xhosa TTS system in order to generate natural sounding and intelligible commands.

The purpose is to present the essential methodological and theoretical framework for prosody generation, that might also be applied in further studies. Therefore, the study is limited to only a few contexts in which imperatives might occur. This study should be considered as part of a potentially greater research project. The ideal is to build a text-to-speech system that will be able to handle an unlimited vocabulary and produce natural, comprehensible Xhosa speech. This task can only be accomplished if further studies of this nature[2] are conducted in order to obtain prosodic information on all the modes of grammatical usage in Xhosa. Such grammatical structures include the infinitive, indicative mood, participial mood, subjunctive mood, consecutive mood etc.

[2] Refer to the work of Jones et al. (1998a, 1998b) on queclaratives in Xhosa.

1.3 Organization of Study

Since this is an inter-disciplinary study it may contain material that is unfamiliar to readers of a specific academic field. To accommodate both readers from linguistic and engineering backgrounds, more introductory information and detailed descriptions are given at times, than would normally be the case.

In Chapter 2 an introduction to prosody is given. This chapter focuses on the role of prosody in the production and perception of natural speech. The relevant terms regarding prosody such as duration, loudness, pitch, prominence, tone and intonation are introduced.

Chapter 3 deals with the concept of speech synthesis. An introduction to the general field of speech synthesis and text-to-speech systems is given. The different aspects of natural language processing (NLP) are discussed and attention is given to the role of prosody in the synthesis of speech. An architecture for the NLP module of a Xhosa text-to-speech system is proposed and subsequently illustrated. Two prominent methods of synthesis are explained and recommendations for choosing an appropriate synthesis system for Xhosa are presented. Finally, the evaluation of text-to-speech systems as well as their various applications are discussed.

Chapter 4 focuses on the production of imperatives in Xhosa. The morphological structure of imperatives in Xhosa is explained. An acoustic analysis that was conducted on a corpus of imperatives is discussed with reference to the aims of the analysis, the data used and the recording and preparation of the data. The results of the statistical analysis of the acoustic data are also presented.

In Chapter 5 the perceptual experiment that was conducted on imperatives in Xhosa is discussed. This experiment is primarily based on the acoustic analysis conducted as reported in the previous chapter. The chapter is divided into subsections in which the aims of the experiment are explained as well as the method used, the preparation of stimuli, the compilation and presentation of the perception tests and the results acquired. Two examples are also given of how the results may be applied in the prosody generation module of a Xhosa TTS system.

Final conclusions are drawn in Chapter 6 and suggestions for future research are presented.

Chapter Two

Prosody

2.1 Chapter Overview

In this chapter different aspects of prosody and the role it plays in the production and perception of natural speech are discussed. The aim of this chapter is to familiarize the reader with the terms related to prosody, as they will be used in the acoustic and perceptual analyses that follow. It also aims to explain the relationship between prosodics and the segmental features of speech so that it may become clear why prosodic features should indeed be acoustically and perceptually analyzed.

Firstly, a general introduction to prosody is given and this will be followed by a more detailed discussion of the terms duration, loudness, pitch, tone and intonation.

2.2 Introduction to Prosody

In the words of Ostendorf (nd.), "*we as human listeners, bring many sources of information to bear on the problem of interpreting an utterance, including syntax, semantics, our knowledge of the world and conversational context, as well as prosody*".

Prosody concerns suprasegmental characteristics of speech through which the speaker conveys information about the structure of the message and the locus of important parts of the message. Through these suprasegmental characteristics the speaker also conveys information about his attitude towards the ongoing discourse and about his emotional state (Dirksen et al., 1995). Prosodic phrase structure and prominence patterns give clues as to how to parse a word string, which element is in focus, whether a point is in question, and whether there has been a change in topic. As such, prosodics provide the link between the acoustic realization and the linguistic interpretation of a word (Ostendorf, nd.).

As an acoustic structure, prosody extends over several segments or words and can not be localized to a specific sound segment, or change the segmental identity of speech segments

(Price, 1997:47). Price (1997:47) goes on to explain that "*prosody consists of a phonological aspect (characterized by discrete, abstract units) and a phonetic aspect (characterized by continuously varying acoustic correlates)*" (see also Cruttenden, 1986:2).

Prosodic characteristics comprise of the melody of the speech, word and phrase boundaries, (word) stress, (sentence) accent, rhythm, tempo, and changes in the speaking rate (Van Bezooijen & Van Heuven, 1997d). Johns-Lewis (1986:xix) and Cruttenden (1986:2) believe that the three physical parameters most commonly given as being *prosodic* are **duration** (perceived as length), **intensity** (perceived as loudness) and **fundamental frequency** (perceived as pitch). These three acoustic correlates of prosody will now be explained in more detail, as they form the basis of analysis in this study.

2.3 Duration

According to Cruttenden (1986:2), "*length concerns the relative durations of a number of successive syllables or the duration of a given syllable in one environment*". The duration of particular syllables can be measured in order to judge whether varying degrees of accent involve varying degrees of lengthening. Before this can be done, however, some decisions regarding syllable boundaries should be made. These decisions can to some extent be arbitrary (Cruttenden, 1986:2). There are many different influences on the absolute duration of a segment or syllable. One such influence is the 'innate' length of sounds. The point and manner of articulation of the segment itself, as well as the preceding and following segmental sounds may condition the innate length of a sound[3]. Suprasegmental factors such as the fact that the last syllable before a pause is often lengthened, also influence the measurement of absolute duration (see Cruttenden, 1986:2 and Lehiste, 1970:53). When syllable boundaries are to be determined, it is essential to apply the method consistently. Spectrograms are useful for this purpose, since they provide clear information regarding the place and manner of articulation of the sound. Speech analysis

[3] As explained in Section 4.3.2.1, the data analyzed in this study had a /C+ala/ structure so as to minimize the effect of innate length on the comparison results.

software also provide the means to *tag*[4] and annotate the speech signal at these boundaries. This method may also assist in the consistent measurement of duration.

2.4 Loudness

The acoustic correlate loudness concerns "*changes of loudness within one syllable or the relative loudness of a number of successive syllables*" (Cruttenden, 1986:2). Cruttenden (1986:3) explains that "*loudness as perceived by the listener is related to the breath-force which a speaker uses*". The concepts of amplitude, sound pressure, power, energy, and intensity are all related and, according to Lehiste (1970:113), all of these concepts are involved in the description of the acoustic correlates of respiratory effort.

As a subjective property of a sound, loudness is most directly related to intensity (Lehiste, 1970:113). An absolute measurement of intensity is given in decibels (dB). As Cruttenden (1986:3) points out, however, "*the relationship of absolute intensity to perceived loudness is by no means linear (a sound has to be much more than doubled in absolute intensity before it will be heard as twice as loud) and moreover the relationship is different at different frequencies*".

Lehiste (1970:114) also shows that perceived loudness depends upon the fundamental frequency, the spectral characteristics and the duration of the sound.

2.5 Pitch

The term pitch refers to the "*varying height of the pitch of the voice over one syllable or over a number of successive syllables*" (Cruttenden, 1986:2). Physiologically, pitch is directly related to the rate of vibration of the vocal cords within the larynx. Variation in the rate of vibration is primarily produced by the length and tension of the vocal cords, controlled by the intrinsic muscles of the larynx, and secondarily, by the air pressure below the larynx (Cruttenden, 1986:3). Rate of vibration of the vocal cords is reflected in the acoustic measurement of *fundamental frequency* (F_0). This term refers to the number of repetitions of the regular waveform within one second. Such a regular waveform is

[4] Section 4.3.2.3 presents the method of phoneme boundary determination and tagging applied in this study.

typically produced when the vocal cords vibrate for voicing. In other words, the number of times that the vocal cords completely open and close in one second is directly related to the frequency of the waveform (Cruttenden, 1986:3-4).

House (1990:9) found that, of the three acoustic correlates of prosody, fundamental frequency is generally recognized as supplying listeners with the greatest amount of prosodic information on many different levels simultaneously. House (1990:10) explains that the perception of F_0 movement is referred to as *tonal movement.* The auditory system transforms tonal movements in speech first into relevant pitch movements and then transforms this pitch movement into relevant linguistic or paralinguistic categories. Such categories include stress, focus and emphasis, word accents and intonation (House 1990:9).

2.6 Prominence: Tone and Intonation

Now that the terms length, loudness and pitch have been defined, an explanation concerning the importance of these three acoustic correlates of prosody will be given. Firstly, the relationship between their measurable attributes and their linguistic function is often complex. Secondly, these three features contribute in varying degrees to give some syllables *prominence* when compared to other syllables (Cruttenden, 1986:7).

Crystal (1997: 331) defines the term prominence as follows: "*A term used in auditory phonetics to refer to the degree to which a sound of syllable stands out from others in its environment.*" In certain instances syllables are given prominence in order to convey specific lexical information. In other words, moving the prominence from one syllable to the other, without changing the segmental composition, produces a change in meaning. When pitch plays the most important role in providing this syllabic prominence, this language in question can be called a *tone* language. Cruttenden (1986:8) explains tone as "*a feature of the lexicon, being described in terms of prescribed pitches for syllables or sequences of pitches for morphemes or words*".

But prominence is not only a feature of words; it is also a feature of connected speech where prominence is on the one hand given to certain words within a sentence and on the other hand given to certain sentences within a larger unit of connected speech. These sequences or patterns of prominent and non-prominent syllables that form the framework

of connected speech, produce a particular rhythmical effect which can be otherwise referred to as *intonation* (Cruttenden, 1986:7). According to Cruttenden (1986:7,8) "*intonation concerns which syllables are prominent, how they are made prominent, and to what extent they are made prominent; it also concerns how the movement from one prominent syllable to the next is accomplished*".

2.7 Chapter Summary

Prosodic features as suprasegmental characteristics of speech were discussed in this chapter. It was explained how acoustic correlates of prosody such as duration, loudness, pitch, as well as tone and intonation are utilized to give structure and meaning to a message.

The following facts regarding the prosodic features were established:

- The measurement of duration of particular syllables can be useful in determining whether varying degrees of accent involve varying degrees of lengthening.
- Loudness is the subjective property of a sound that is most directly related to intensity.
- Pitch is directly related to the rate of vibration of the vocal cords and this is reflected in the acoustic measurement of *fundamental frequency.*
- Tone is a feature of the lexicon; being described in terms of prescribed pitches for syllables or sequences of pitches for morphemes or words.
- In tone languages pitch plays the most important role in providing syllabic prominence within a sentence.
- Intonation concerns the sequences of prominent and non-prominent syllables that form the framework of connected speech.

Chapter Three

Xhosa Text-to-Speech Synthesis

3.1 Chapter Overview

This chapter deals with the concept of speech synthesis. An introduction to the general field of speech synthesis, as well as a specific kind of speech synthesizer namely a text-to-speech (TTS) synthesizer is given. This is followed by a description of a proposed architecture for a Xhosa text-to-speech synthesizer. The emphasis is on the different aspects of natural language processing and how existing architectures may be adapted and then applied to process Xhosa text. Language processing rules for Xhosa are discussed and a new set of grapheme-to-phoneme rules is proposed. The proposed architecture for natural language processing is subsequently exemplified. Attention is given to the role of prosody in the synthesis of speech. A more detailed application for imperatives in Xhosa is then dealt with in Chapter 5 after thorough acoustic and perceptual analyses have been done.

The theory of digital signal processing is discussed in short and two prominent methods of low-level synthesis are explained. Recommendations for choosing an appropriate synthesizer for a Xhosa TTS system are presented. The last two sections in this chapter provide some information regarding the evaluation of TTS systems and suggest various applications.

3.2 Introduction

Speech Synthesis

Speech synthesis is concerned with the generation of speech by a speech output system. A speech output system is some artefact, whether a dedicated machine or a computer program, that produces signals that are intended to be functionally equivalent to, or simulate the speech produced by humans (Bloothooft et al., 1995; Van Bezooijen & Van Heuven, 1997a). The general field of speech synthesis can be divided into two areas, namely *high-level synthesis* and *low-level synthesis.* High-level synthesis deals with the

linguistic and prosodic processing of the input. For this reason the high-level parameters are more dependent on the specific language in question and less dependent on the implementation of the synthesizer itself. A low-level synthesizer, on the other hand, physically converts the high-level parameters into sound.

The majority of speech output systems are driven by text input. These systems are called *text-to-speech systems.*

Text-to-Speech Systems

Text-to-speech (TTS) synthesis is defined by Dutoit (1993:18) as "*the automatic production of speech, through a grapheme-to-phoneme transcription of the sentences to utter*". In other words, a TTS system uses textual information as input and converts it into speech output. Besides the vast amount of digital text available, the input text can also be directly introduced in the computer by an operator or scanned and submitted to an Optical Character Recognition system. A TTS synthesizer should be able to read any text aloud and therefore TTS synthesis is also known as *unlimited vocabulary speech synthesis.*

To accomplish the text processing and speech production tasks, TTS systems generally have two main components: the Text module and the Speech module, or rather, the

- Natural Language Processing module (NLP) and
- Digital Signal Processing module (DSP).

The NLP module contains the algorithms that process the input text. Language processing involves the production of a narrow phonetic transcription of the text as well as providing the parameters concerning prosody. The output of the NLP module is essential for the pronounciation of the input text.

The NLP module is divided into the following three smaller modules:

- Text analysis
- Automatic phonetization
- Prosody generation

For languages such as English or French the text analysis module is composed of four sub-modules namely:

- Pre-processing
- Morphological analysis
- Contextual analysis and
- Prosodic parser (Dutoit, 1997a; Van Bezooijen & Van Heuven, 1997b).

After the text has been successfully analysed, it is phonetically transcribed in the *automatic phonetization* module. The *prosody generation* module is the last module within the NLP module and this is where the suprasegmental features that contribute to the naturalness of the speech to be synthesized are generated.

Once the textual input has been thoroughly processed, the speech signal is synthesized in the DSP module. Two prominent types of speech synthesizers are currently in use, namely:

- Rule based synthesizers and
- Concatenative synthesizers

The general description of a TTS system given above is based on architectures currently applied for languages other than Xhosa (cf. Van Bezooijen & Van Heuven, 1997; Laureate, 1996; Bloothooft et al., 1995; Lange, 1993; Bell Labs, 1997; Hertz, 1999; Allen, 1991:742; Dutoit, 1993:24; Sproat, 1997:176). While many different approaches to text processing and speech synthesis for different languages exist, it has not yet been established which approach would be appropriate for the African languages. In Section 3.3 some suggestions will be made for the natural language processing of Xhosa text. The proposed architecture is based on those that have been applied successfully for other languages, but certain language specific factors are also taken into account. Rule based and concatenative synthesizers are discussed in the DSP section and recommendations regarding a possible method of synthesis for Xhosa are also given.

A Text-to-Speech System for Xhosa

No TTS system currently exists for Xhosa and the appropriate approach to natural language processing and digital signal processing has not yet been established. The

morphology of Xhosa differs greatly from that of other (European) languages. Therefore, it is expected that some alternative approach will be taken regarding certain aspects within the NLP module. However, thorough research in this field is necessary in order to determine which aspects of NLP should be adapted and to what extent.

Although prosody generation is the focus of this study, Section 3.3 below will demonstrate the processes involved in order to reach this stage. A number of language specific factors that may influence Xhosa text processing will be introduced. The NLP architecture proposed is merely suggestive in nature and the fact that a practical system may eventually function differently is acknowledged.

3.3 Natural Language Processing for Xhosa

The NLP module does text processing so as to enable the system to pronounce the input text naturally. This module utilizes rules and lexicons in combination to produce a narrow phonetic transcription of the text. Xhosa has a relatively uncomplicated orthographic system, i.e. according to Roux (1989:74), "*a system free of conventions relating to the use of orthographic symbols*". The language is very 'phonetic' and therefore allows for a transcription to be made solely by employing so-called grapheme-to-phoneme rules. The system should then be able to pronounce Xhosa text through a simple process of text pre-processing and grapheme-to-phoneme conversion. However, prosody generation for Xhosa will require a deeper knowledge of the language albeit not as deep as the knowledge required for language translation, information retrieval etc.

Subsequently, the NLP module may consist of a series of sub-modules (or processes). Each sub-module enriches its input by adding specific auxiliary information. This way processing occurs sequentially and in manageable chunks, instead of 'simultaneously' in one global module. Such a modular approach makes it possible to experiment with different implementations of sub-modules without redeveloping the whole system. Language processing is done on segmental level in the text analysis and automatic phonetization modules and on suprasegmental level in the prosody generation module. The NLP module for Xhosa as demonstrated in Figure 3.1 will be described in detail in the subsequent sections.

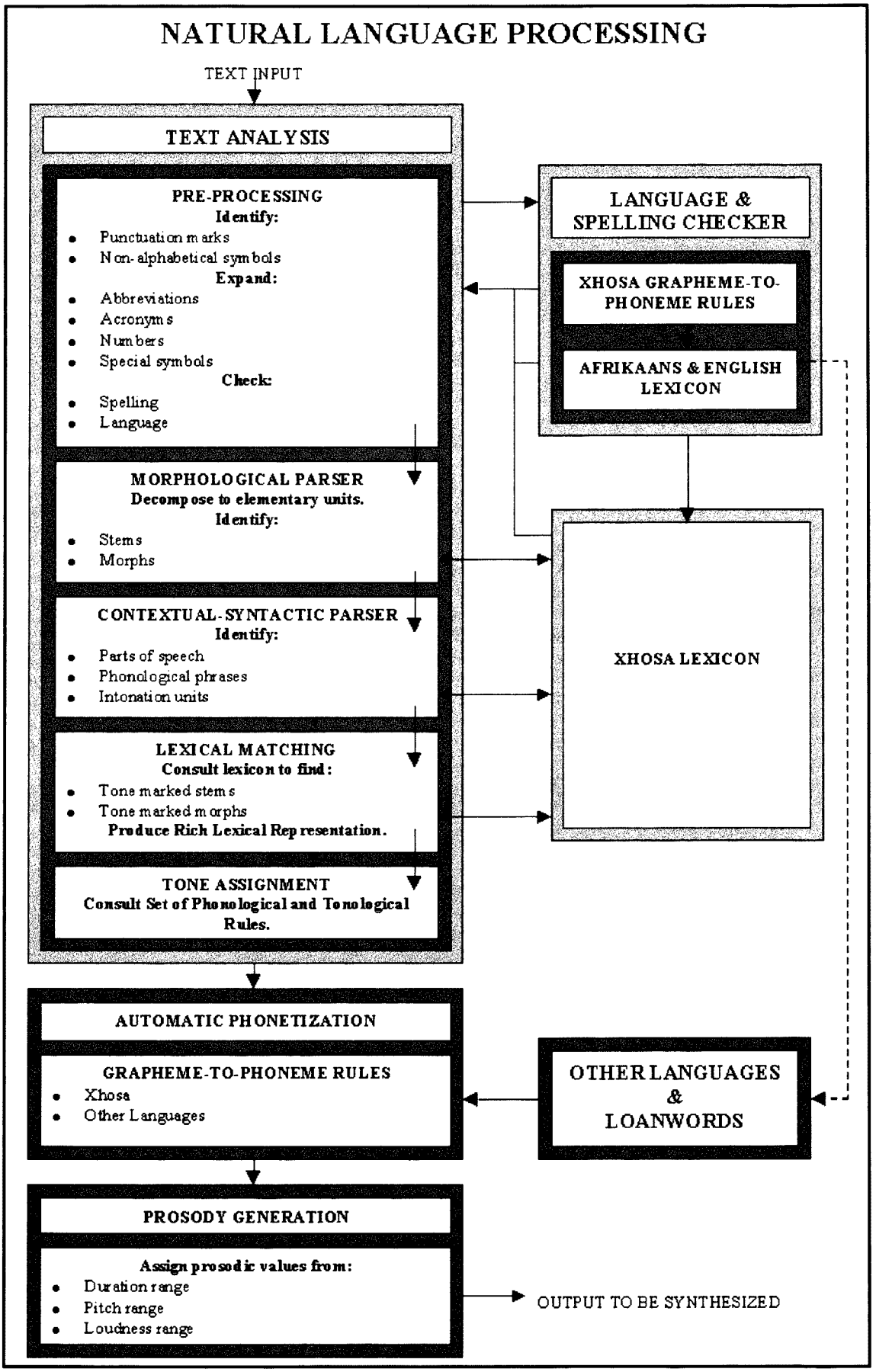


Figure 3.1 Natural Language Processing for Xhosa.

3.3.1 The Text Analysis Module

As explained in the introduction, the NLP module of TTS systems for European languages such as English or French is usually composed of four text analysis sub-modules namely the pre-processing, morphological analysis, contextual analysis and prosodic parser sub-modules (Dutoit, 1997a; Van Bezooijen & Van Heuven, 1997c). However, to suit the morphologic and syntactic structure of Xhosa, the text input may be processed using the following five sub-modules:

- Pre-processing sub-module
- Morphological analysis sub-module
- Contextual-syntactic parser
- Lexical matching sub-module
- Tone assignment sub-module

The Xhosa text is processed by each sub-module in the order given above. The structure of the respective sub-modules are discussed next.

The Pre-Processing Sub-Module

The *pre-processing* sub-module organizes the input sentences into manageable lists of words. In this process of *text normalization,* punctuation marks and other non-alphabetic textual symbols are identified and abbreviations, acronyms, numbers, special symbols, etc. are expanded to full texts (i.e. orthographic strings) where needed.

These processes would be sufficient for normalizing pure Xhosa texts, but modern Xhosa would present a problem since it contains many loanwords. In Xhosa, loanwords are most likely borrowed from Afrikaans and English. These words are phonologized so that they may conform to the /CV/ structure of Xhosa. An example of a phonologized loanword taken from Afrikaans would be:

(3.1) *itafile* 'table'
/VCVCVCV/

However, not all loan words are fully phonologized, as the next example of a word borrowed from English shows:

(3.2) *ijam* 'jam'
/VCVC/

In many instances English and Afrikaans words are used alongside Xhosa, for example, counting is usually done in English. Additionally, spelling mistakes might also appear in the text. These problems could be dealt with in a separate sub-module that operates in conjunction with the pre-processing sub-module. This sub-module would function as a language and spelling checker[5].

In the language and spelling checker sub-module, the CV structure of Xhosa may be exploited in the identification of Xhosa words. A set of Xhosa phonological (grapheme-to-phoneme) rules (cf. Section 3.3.2) may be applied to check the phonological structure of each word. Words of which the phonological structure does not concur with that of Xhosa should be identified. If necessary, a Xhosa lexicon[6] may be consulted to check for spelling mistakes. If spelling mistakes are identified, these words should be marked as such and corrected. Words of which the phonological structure does not concur with that of Xhosa can then be checked in a relatively small lexicon, which consists of Afrikaans and English words and loanwords[7] that are frequently used by Xhosa speakers. Once matches for these words are found in the Afrikaans/English lexicon, these words should be marked so that they can be taken up again in the pre-processing sub-module for further analysis. English and Afrikaans words will not be morphologically parsed or analysed in the lexical matching and tone assignment sub-modules, and could therefore theoretically be carried down directly to the automatic phonetization module. However, these words do occupy a specific position in the context of the sentence and is part of the syntactical structure. Therefore, it should also be taken up again in the pre-processing sub-module together with all other words processed in the language and spelling checker.

---

[5] Software for an efficient Xhosa spelling checker has not as yet been developed.
[6] The exact nature of the lexical inscriptions is yet to be determined, but some suggestions are presented later in this section.
[7] The loanwords included in this lexicon would be those that have been partially phonologized.

The Morphological Analysis Sub-Module

For a language such as English, the task of the *morphological analysis* sub-module is to decompose inflected, derived, and compound words into their elementary graphemic units and to propose all possible part of speech categories for each word individually (Dutoit, 1997a). However, the morphological structure of Xhosa is very different from that of the European languages and therefore the morphological decomposition process for Xhosa involves much more than just inflected, derived and compound words as they appear in English.

English, for example, follows a disjunctive writing style where a word is a segmental unit and a sentence is a combination of more than one of these independent segmental units. On the other hand, Xhosa is a morphologically complex language (also known as a *polysynthetic* language). It uses a conjunctive writing style where one segmental string might represent a full sentence with reference to a subject, a verb and an object. One segmental string (the sentence) may consist of clusters of morphemes. These morphemes can appear as affixes, prefixes and suffixes connected to a stem and they may have inflectional or derivational features. Inflectional morphemes may, for example, express tense, mood, negativity, aspect and agreement. Examples of derivational morphemes are the causative, passive, applicative and reciprocal (Du Plessis, 1978).

The noun in Xhosa has an identifiable morphology, which serves as a governing factor in realising agreement morphology on other categories such as the verb and nominal modifiers. Each noun falls into one of 16 to 20 separate groups or classes and each noun takes the class prefix of the noun class to which it belongs. The subject noun is linked to the verb, adjectives etc. through subjectival agreement and in the same way the object noun can also be linked to the verb and other parts of speech. This system of linking with the noun is a characteristic feature of Xhosa and related African languages (Wentzel et al., 1972.)

The verb in Xhosa also has a complex derivational structure in that it takes verbal derivational morphemes (such as the applicative, causative, reciprocal etc.) that influence the argument structure of the sentence.

Concatenating prefixes and suffixes to the stem entails morpho-phonological syllable structure processes. The phonological /CV/ or /CwV/ structure of Xhosa is maintained in that certain phonemes are added, deleted or changed when the morphemes and stems in words are concatenated. Phonological processes regarding syllable structure in Xhosa are, for example, vowel coalescence, initial vowel deletion (V_1 del.) and semivocalisation. Thus, when a morphological analysis of the sentence is done, these processes should be taken into account.

The nominal modifier *bonke* in the following sentence exemplifies a word where a V_1 del. process occurred:

(3.3) aba + ntu ba + onke

$V \rightarrow \phi$ / C_ + V/

Abantu bonke 'All the people'

The semivocalisation process can be exemplified as follows:

(3.4) um + ntu + ana

/ (C) u + a / → (C)wa

umntwana 'child'

Sentences may be decomposed to their elementary graphemic units (their morphs) with the use of morpho-phonological rules and by querying a lexicon. Theoretical, descriptive analyses and rules for the morphology, syntax and phonology of Xhosa exist (cf. Du Plessis:1978 and Du Plessis & Visser:1992) but it has not yet been automated. The lexicon referred to, is a database or categorized list of all the stems and morphs of the language. It should also contain information on all the levels of language analysis (phonology, morphology, syntax, semantics, tone etc.).

Sanfilippo (1997:105) recommends that lexical acquisition be done automatically from large text corpora since the manual creation of lexical resources is expensive, time-consuming and error prone. With regard to morphological decomposition, Theron (1999) proposes the automatic construction of rule sets for the morphological and phonological levels of language analysis using a two-level computational morphological framework.

This method was applied successfully for Xhosa noun locatives (Theron:1999), but otherwise these rules and lexicons are yet to be developed for the African languages. Because computational linguists aim to design these systems to produce accurate results in the shortest possible time, it should be noted that the type of rules that are derived with methods such as the above, may not coincide with 'traditional' morphological rules that currently exist. The following example illustrates the basic morphological decomposition of a sentence using rules and/or querying the Xhosa lexicon. The assumption is made that the output will eventually be accomplished with the use of automatically derived rules, but since we do not yet know the nature of these rules, the more 'traditional' rules are applied in this example:

(3.5) *Ndibabiza bonke abantu* 'I am calling all the people'

/ndi + ba + biz + a ba + onke aba + ntu/

The Contextual-Syntactic Parser

Once the basic segmental units are identified the *contextual-syntactic parser* queries the Xhosa lexicon and assigns parts of speech categories and semantic properties to each unit. Each unit is considered in its context and the list of its possible part of speech categories is reduced to a very restricted number of highly probable hypotheses, given its semantic properties and the corresponding possible parts of speech of neighbouring units (Dutoit, 1997a). A basic prosodic structure is also derived through the identification of phonological phrases and intonation units. Assigning parts of speech can be exemplified as follows:

(3.6) *Ndibabiza bonke abantu* 'I am calling all the people'

/ndi + ba + biz + a ba + onke aba + ntu/

/ndi/ = subject agreement prefix
/ba/ = object agreement prefix
/biz/ = verb stem
/a/ = verb suffix (present aspect)

/ba/ = object agreement affix
/onke/ = quantifier stem
/aba/ = noun prefix (indicating class and number)
/ntu/ = noun stem (object)

The Lexical Matching Sub-Module

In the Xhosa lexicon there are some words that are distinguished only by their tonal pattern. Because tone is a lexically distinctive feature, the units that make up the lexicon should be tone marked.

The process of *lexical matching* involves the extraction of the particular stems and morphs as they have been identified in the morphological and syntactical analyses, from the different lists in the lexicon. These tone marked stems and morphs are then concatenated through the application of phonological rules so as to produce a *rich lexical representation.* Now the sentences have the correct phonological structure, but each segment still has the tonal feature as it has been assigned in its most basic form.

The Tone Assignment Sub-Module

Although the rich lexical representation now includes tonal features the *tone assignment* process is not complete before tonological rules have been applied. A set of tonological rules should be utilized in order to 'normalize' the tonal pattern of the sentences.

3.3.2 The Automatic Phonetization Module

Dutoit (1997a; 1993:27) also calls this the "*letter-to-sound*" (LTS) module and explains that it automatically produces a phonetic transcription of the incoming text. This can be accomplished by using a set of letter-to-sound (or *grapheme-to-phoneme)* rules that simply map sequences of graphemes into sequences of phonemes. In a language such as English, most words appear in genuine speech with several phonetic transcriptions. In this case the pronunciation module of a TTS system for English contains a pronunciation dictionary that records words whose pronunciation could not be predicted on the basis of general rules (Dutoit, 1997a; Sproat, 1997:178; Van Bezooijen & Van Heuven, 1997b). However,

Xhosa is a very phonetic language in that its phonetic representation is very similar to its orthographic representation. Such pronunciation problems as are encountered in English do not occur in Xhosa. Therefore Xhosa sentences can be transcribed phonetically, on the basis of rules alone. English, Afrikaans and loanwords that were identified in the spelling checker should also be transcribed using either grapheme-to-phoneme rules or the English/Afrikaans lexicon.

Grapheme-to-Phoneme Rules for Xhosa

A set of conversion rules for Xhosa already exists (cf. Roux, 1989:78) and the success rate of this conversion system was found to be 99.5%. These rules were developed for a stand-alone grapheme-to-phoneme (G-P) conversion system that would primarily be used for linguistic purposes. The rule construction was done in such a way that the system is compelled to consider both orthographic characters and phonetic characters in order to execute the rules and transcribe the text accurately. In certain instances it is also necessary for the system to return to and correct previously transcribed sections where rules were inappropriately executed. These factors add to the computational complexity of the algorithm and may consequently impede the processing speed of the system. Since the existing system was primarily used for linguistic purposes, real-time processing was considered to be of lesser importance. However, should a grapheme-to-phoneme converter form part of a TTS system, the goal would be to achieve accurate conversion at minimum cost (Roux, 1989:75). This was in fact the motivation for developing a set of rules, based on those of Roux (1989), that would be more suitable to be embedded in an automatic phonetization module of a Xhosa TTS system.

In Table 3.1 a new set of grapheme-to-phoneme rules that can be applied to phonetically transcribe pre-processed Xhosa text is presented. Each rule defines how a particular orthographic input string maps to a corresponding phonetic output string. The convention that is used (as shown in 3.7 below), is that the orthographic template pattern to be matched, is written on the left hand side and the phonetic output string on the right hand side, with an arrow between the two strings.

(3.7) `orthographic template pattern → phonetic output string`

In the transcription process, the rules are employed as follows: the orthographic input text is scanned from left to right for the longest template pattern that can be matched, as defined in the set of rules. The rules are ordered alphabetically, but in a way that longer patterns are evaluated before shorter patterns in order to avoid the exclusion of certain rules from the search. Consider the following example where the input text is:

(3.8) *iingcango* 'doors'

The first template pattern to be matched would be 'ii' and not 'i' (both patterns can be matched, but 'ii' is the longer pattern of the two).

If a particular template pattern in the orthographic text is matched, no more patterns are searched for, and the phonetic output corresponding to that pattern is generated and saved. The processor always advances to the rest of the untranscribed input. For 3.8 the phonetic output [i:] would be generated corresponding to the pattern 'ii'. Now the processor advances to the untranscribed input and finds a match for the pattern 'ngc'. This pattern is transcribed as: [ɲ͡ɟ].

The processor uses a look-ahead approach so that it mostly scans forward. The most accurate rule is always applied the first time, meaning that it is never necessary to return to a previously transcribed string in order to change it to a more accurate phonetic representation.

Provision is made for defining generic sets such as a vowel and a consonant set. In the template pattern generic sets may occur in combination with orthographic text, in which case, vowels are represented with a *V* and consonants with a *C*. For example, the rule below states that if a /g/ occurs between two vowels, it should be transcribed as [g̬]:

(3.9) $VgV \rightarrow$ [g̬]

The vowel and consonant sets for Xhosa represent all the orthographic symbols that may appear in texts, contrary to those of Roux (1989), that represent Xhosa phonemes in the

form of phonetic symbols. These sets are defined as: V = /a,e,i,o,u/ and C = /b,c,d,f,g,h,j,k,l,m,n,p,q,r,s,t,v,w,x,y,z/ respectively.

The notation $s = C^n$, where $1 \leq n \leq 4$, means that the sub-string s may contain between one and four occurrences of the characters in set C. For example, the strings /t, th, tyh, ntyh/ would be possible instances of C^n.

The matching of context sensitive template patterns such as in (3.9) sometimes requires the processor to scan backwards. In these cases the processor scans one orthographic character to the left without changing that character. For example, consider the following input string:

(3.10) *ugaba* [uga̰ɓa] 'stalk'

The first pattern to be matched, transcribed and saved would be 'u'. When advancing to the untranscribed orthographic string, the processor considers /g/. To determine whether the '*V*g*V*' template pattern can be matched, the processor scans one character to the left (i.e. in the backward direction) and identifies the orthographic character /u/ as a vowel. Now the processor scans one character to the right (i.e. in the forward direction) and identifies the orthographic character /a/ also as a vowel. The '*V*g*V*' pattern is subsequently matched and only the orthographic character /g/ is transcribed. The orthographic character /a/ is not transcribed in the execution of this rule, it is only transcribed by a subsequent rule i.e. a → [a].

The proposed algorithm maintains the high level of accuracy without sacrificing processing speed. It is also less complex and can easily be extended to other African languages. These properties make the algorithm more suitable for implementation in an automatic phonetization module.

Table 3.1 lists the grapheme-to-phoneme rules for Xhosa.

Table 3.1 Grapheme-to-phoneme rules for Xhosa.

C = b,c,d,f,g,h,j,k,l,m,n,p,q,r,s,t,v,w,x,y,z			
V = a,e,i,o,u			
aa → [aː]	gr → [ɣ]	ngc → [ᵑǃ̬]	p → [p']
a → [a]	*V*gw → [ɡʷ]	ngq → [ᵑǀ̬]	qh → [ǃʰ]
bh → [b̥]	gx → [ᵑǁ̬]	ngx → [ᵑǁ̬]	q → [ǃ]
b → [ɓ]	*V*g*V* → [ɡ̥]	nkc → [ŋ ǀ]	rh → [x]
ch → [ǀʰ]	g → [g]	nkh → [ŋkʰ]	r → [r]
c → [ǀ]	hl → [ɬ]	nkq → [ŋ ǃ]	sh → [ʃ]
dl → [ɮ]	*V*h*V*→[ɦ]	nkx → [ŋ ǁ]	s → [s]
dy → [ʄ̥]	h → [h]	nk*V* → [ŋk']	tsh → [tʃ']
*V*d*V* → [d̥]	imf → [iᵑɸf']	nty → [ɲc']	tyh → [cʰ]
d → [d]	imv → [iᵑɸv]	nyh → [ɲʰ]	th → [tʰ]
emfi → [eᵑɸf'i]	ii → [iː]	nc → [ᶯ]	tl → [tɬ']
emfu → [eᵑɸf'u]	i → [i]	ng → [ᵑg]	ts → [ts']
emvi → [eᵑɸvi]	*V*j*V*→ [d̥ʒ̥]	nj → [ɲdʒ]	ty → [c']
emvu → [eᵑɸvu]	j → [dʒ]	nq →[ᶯǃ]	t → [t']
emf → [ɛᵑɸf']	kh → [kʰ]	nx → [ᶯǁ]	umb → [um̥b]
emv → [ɛᵑɸv]	kr → [kx']	ny → [ɲ]	um*C* → [um̥]
e*C*^{n}i → [e] 1 ≤ *n* ≤ 4	k → [k']	nz → [ndz]	uu → [uː]
e*C*^{n}u → [e] 1 ≤ *n* ≤ 4	l → [l]	n → [n]	u → [u]
ee → [ɛː]	mb → [mb]	o*C*^{n}i → [o] 1 ≤ n ≤ 4	v → [v]
e → [ɛ]	m → [m]	o*C*^{n}u → [o] 1 ≤ *n* ≤ 4	w → [w]
f → [f]	ntyh → [ɲcʰ]	oo → [ɔː]	xh →[ǁʰ]
gc → [ᵑǀ̬]	ndl → [ndɮ]	o → [ɔ]	x →[ǁ]
gq → [ᵑǃ̬]	ndy → [nʄ]	ph → [pʰ]	y → [j]
			z → [z]

Pronunciation Problems

Pronunciation problems with heterophonic homographs (i.e. words that are pronounced differently even though they have the same spelling) are encountered in English. These words do not exist in Xhosa, but minimal pairs (such as the word *ithanga* mentioned in Chapter 1) where tone is the distinctive feature, do occur. The underlying tones for these words are stored in the lexicon, as well as their semantic attributes. In addition the lexicon stores information regarding the possible subjects and complements that verbs may take. With the availability of this pre-determined information, these pronunciation problems could be eliminated.

New words and proper names may be problematic since they cannot all be stored in the dictionary. For English, this can be overcome by deriving these names from others via morphological processes (Dutoit, 1997a; Sproat, 1997:178). Guidelines for the pronunciation of proper names are derived from graphemic analogies that exist between these names and other words in the language.

As for Xhosa, proper names are often derived from nouns and verbs. A name may even constitute a sentence on its own. Within a larger sentence, the Xhosa proper name is usually preceded by a prefix /u-/ and the name itself is written with a capital letter. This structure can easily be recognised by the system and the pronunciation of the name would be the same as that of the noun or verb of which it has been derived. However, the tones assigned to proper names are not necessarily the same as those assigned to the original noun or verb. In the case of proper names for females the prefix /u/ is sometimes followed by the prefix /no/ which is written with a capital letter. Different forms of proper names are exemplified in (3.11) below.

(3.11)	*Themba*	male name meaning 'hope'
	Ndibona uThemba.	'I see Themba.'
	Nondumisa	Female name meaning 'praise'
	Ndibona uNondumisa.	'I see Nondumisa.'
	Monwabisi	male name meaning 'one who brings happiness'
	Ntombikayise	female name meaning 'her father's daughter'

Dutoit (1997a) shows that TTS systems make use of one of two automatic phonetization modules. Either a *dictionary-based strategy* is followed, where a maximum of phonological knowledge is stored in a lexicon and where the morphemes that cannot be found in the lexicon are transcribed by rule, or a *rule-based transcription system* is adopted, where only those words that are pronounced in such a particular way that they constitute a rule on their own, are stored in an exceptions dictionary. It has previously been established that grapheme-to-phoneme conversion for Xhosa can be done most successfully using only rules, but it should be noted that the pronunciation problem of tone will prevail if a tone marked lexicon is implemented.

3.3.3 Prosody Generation Module

As indicated in Chapter 2, the speaker uses prosodics to structure his message and to convey information that cannot be derived from text alone. Such information include segmentation cues in the form of phrase boundaries, indication of word stress, sentence accent, rhythm, tempo and changes in speaking rate and intonational meaning. Using prosodic features such as pitch, duration and intensity, the speaker also expresses his attitude and emotion. In the case of a TTS system, the prosody of an utterance reflects the communicative intentions of the writer of the input text. Van Bezooijen & Van Heuven (1997a) explains that "*the reconstruction of the writer's intentions is an implicit part of the linguistic interface*" and that "*all errors in the linguistic interface may detract from the quality of the output speech*".

Prosodics play an essential part in producing natural sounding synthetic speech (Sagisaka, 1997:168). Without natural sounding prosodics the synthetic speech produced will sound monotonous and robot-like and will probably not be accepted by the user (Witten, 1986:81-82; Dutoit, 1997a).

According to Van Bezooijen and Van Heuven (1997a) "*conventional spelling provides a reasonable indication of what sounds and words have to be output, but typically under-represents prosodic properties of the message*". It has been shown within the text analysis and automatic phonetization modules described above, how lexicons in combination with phonological and tonological rules are to be used to extract the necessary information regarding these output sounds and words. The output given by the automatic phonetization module would thus, in many respects, be correct, but 'lifeless'. To complete the NLP module described above a processing module is required that assigns prosodic properties that are strongly based on the tonal characteristics of the language, but also provides meaning and naturalness on a higher level.

Assigning prosodic properties to an utterance is a complicated task (Dutoit, 1993:32). Firstly, Dutoit (1993:33) explains that unlike segmental features, suprasegmental units cannot be given binary values (voiced/unvoiced, nasalized/non-nasalized etc.). These features form a continuum and defining the elementary prosodic units strongly depends on how the utterance will be perceived. Secondly, prosodic features do not correspond

directly to pitch and duration values and they should always be considered relative to all the other linguistic features that constitute the utterance (Dutoit, 1993:33).

To generate prosody in a TTS system, a flow of prosodic parameters should be computed that will control the synthesizer. By computing the pitch, duration and loudness of each speech segment, an intonational and duration model should be designed and this should be used to apply the prosodic structure to the utterance (D'Allessandro & Liénard, 1997:172).

The following chapters form the basis for the design of such a prosodic model for Xhosa imperatives. Intonational, durational and loudness rules to be applied in the prosody generation module are derived from acoustic and perceptual analyses. A practical example of how prosodic parameters can be calculated for imperatives in a Xhosa TTS system is given in Chapter 5.

3.3.4 Examples of Natural Language Processing for Two Xhosa Sentences

The processes employed and subsequent output of the NLP module for a Xhosa TTS system will now be illustrated for the two sentences shown in Table 3.2. Only the text analysis, language and spelling checker and automatic phonetization modules are exemplified in this section, while an application of the prosody generation module (for imperatives in particular) follows later in Section 5.10.

Table 3.2 Example sentences.

	Xhosa	**English Translation**
Sentence 1	Lala!	Sleep!
Sentence 2	Bhalani page 3!	Write page three!

As noted earlier, the NLP module is viewed as a series of smaller processes. Processing occurs sequentially instead of 'simultaneously' and each process enriches its input by adding specific auxiliary information. However, output conventions need to be defined and followed consistently.

In our case the following conventions were defined:

1. A *command* provides information and is defined as follows:

   `command = information`

   The *command* part signifies the parameter involved (e.g. part of speech) while the *information* part contains the value of the parameter, if applicable. Examples are:

   (3.12) `intransitive verb stem =` lal-
   `language = xhosa`

   In certain cases the information part is absent or redundant and it is then omitted. Only the command part is retained.

2. A *rule* is a mapping or transformation of an input string to an output string. It is denoted by an arrow with the *template pattern* to be matched written on the left-hand side of the arrow and the output on the right hand side.

   `template pattern → output`

   For example:

   (3.13) `FFL →` LHL
   bh → ḅ

3. The commands and rules are enclosed in angle brackets. Any number of commands and rules may be grouped together within a pair of brackets and are separated by means of commas.

   `<command = information, pattern → change, command = information>`

4. Commands and rules are applicable to all orthographic text that follow the closing angle bracket up to the opening angle bracket of the next set of commands and rules.

   `<command = information>` word(s) `<command = information>` word(s)

The usage of these conventions will become clear in the examples presented below. A flowchart of the NLP module for each example sentence is presented in Figures 3.2 and 3.3.

Each sub-module in the processing chain adds information specific to that sub-module. All information is propagated from the first sub-module down to the last sub-module. However, in the examples presented below, the propagation of information is not always explicitly indicated.

Example 1: *Lala!*

Text analysis

The sentence is pre-processed in this NLP sub-module:

1. The exclamation mark is identified and the system notes that this is a command sentence. This is indicated in the flowchart by ! .
2. Upper case letters are converted to lower case.
3. The sentence is checked in the language and spell-checking sub-module. No spelling mistakes are identified and the sound structure concurs with that of the Xhosa language and, therefore, it is taken up in the text analysis sub-module again for further analysis.

The morphological parser applies rules and/or queries the Xhosa lexicon to decompose the sentence:

4. The sentence is decomposed into stems and morphs: /lal-/ and /-a/.

The following information is extracted in the contextual-syntactic parser:

5. /lal-/ is identified as an intransitive verb stem.
6. /-a/ is identified as a verb suffix.

Tone values are assigned to each syllable in the lexical matching sub-module yielding a rich lexical representation:

7. /lal-/ is marked with a high tone according to the lexical entry *lal-* → *lál-* .
8. In the same manner /-a/ is marked with a low tone: *-a* → *-à.*

In the tone assignment sub-module tonological rules are applied to the rich lexical representation of the sentence so that the tone marking for the sentence can be corrected.

9. The tonological rule *HL→HL* is applied and subsequently the tonal marking HL does not change.

Automatic phonetization

In the automatic phonetization module the set of Xhosa grapheme-to-phoneme rules (ref. Table 3.1) is queried. The following rules are applied for this example.

10. l →[l]

11. a →[a]

Finally, the output of the automatic phonetization sub-module is presented in the form of a narrow phonetic transcription of the sentence: [lálà] ! . The transcription is enclosed in square brackets by IPA-convention and the sentence punctuation is placed outside the brackets so as not to be mistaken for a phonetic symbol. In this case the punctuation is an exclamation mark, indicating that it is a command sentence. The output of the automatic phonetization module forms the input information for the prosody generation module.

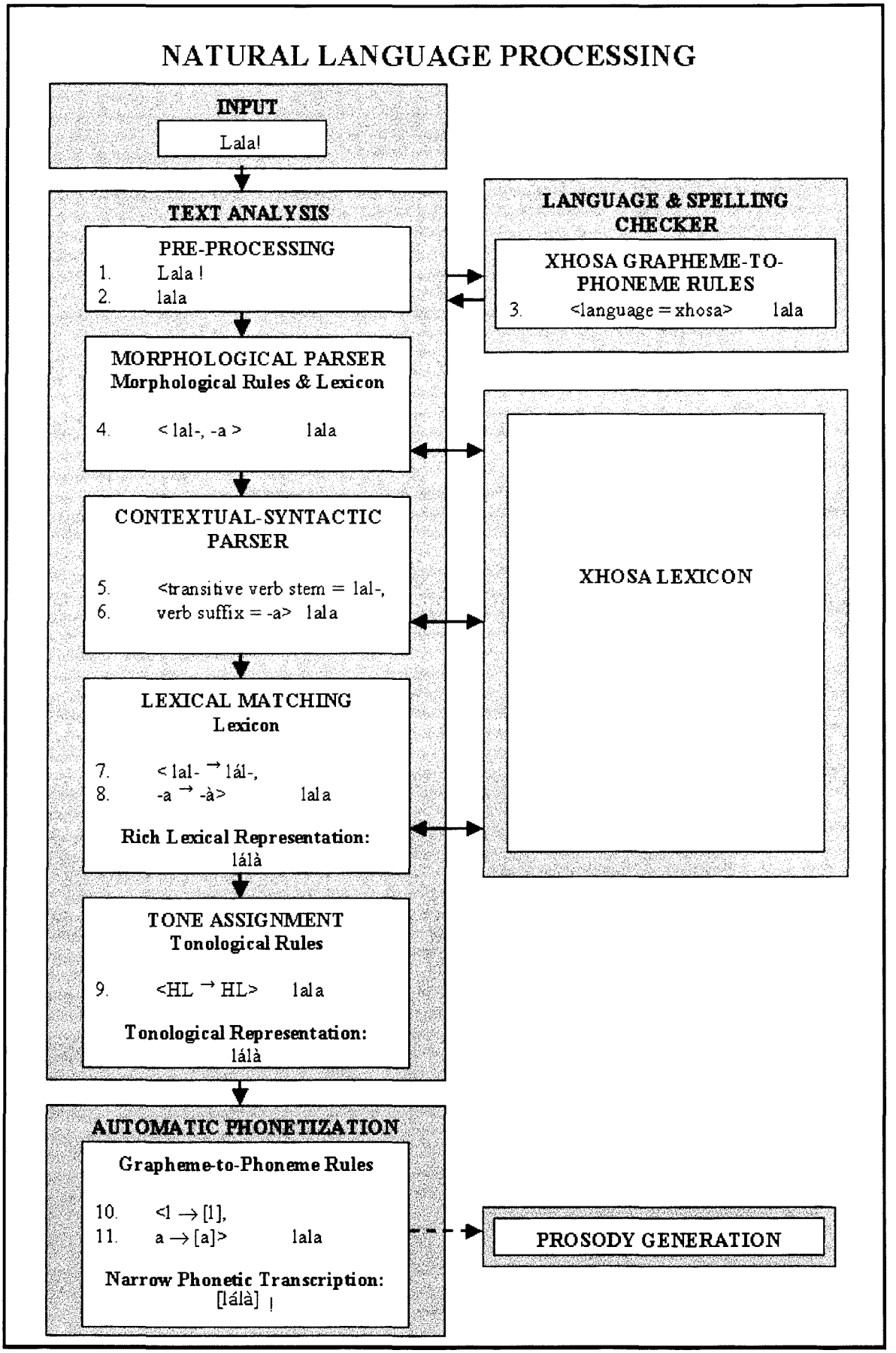


Figure 3.2 The NLP module for sentence 1: Lala!

Example 2: *Bhalani* page 3!

Text analysis

The sentence is pre-processed in this NLP sub-module:

1. The exclamation mark is identified and the system notes that this is a command sentence. This is indicated in the flowchart by ! .
2. Upper case letters are converted to lower case.
3. The character '3' is identified as an ordinary number and expanded to <*language* = *english*> *three*. Note that the system has been designed to convert numbers to their English equivalents.
4. The sentence is checked in the language and spell-checking sub-module. No spelling mistakes are identified for /bhalani/. The system identifies /page/ as a foreign, non-Xhosa word and marks it as <*language* = *-xhosa*>.
5. The word /page/ is located in the Afrikaans and English lexicon and is identified as an English word and marked as such <*language* = *english*>. This study is not concerned with the generation of the necessary speech synthesis parameters for Afrikaans and English words. For the rest of the example we will assume that these parameters are assigned values from an Afrikaans and English lexicon.
6. The processing path returns to the text analysis module where further processing on the sentence continues.

The morphological parser applies rules and/or queries the Xhosa lexicon to decompose the word *bhalani*. Note that the two English words are not morphologically parsed here.

7. The word is decomposed into stems and morphs: /bhal-/, /-a-/ and /-ni/.

The following information is extracted in the contextual-syntactic parser:

8. /bhal-/ is identified as a transitive verb stem.
9. /-a-/ is identified as a verb suffix.
10. /-ni/ is identified as a plural suffix.
11. /page/ is identified as an object.
12. /three/ is identified as an adjective.

Tone values are assigned to each syllable in the lexical matching sub-module yielding a rich lexical representation:

13. /bhal-/ is marked with a falling tone according to the dictionary entry *bhal-* → *bhâl-*

14. /-a-/ is marked with a low tone: *-a-* → *-à-*
15. /-i/ is marked with a low tone: *-i* → *-ì*

The rich lexical representation of the word *bhalani* is /bhâlànì/.

In the tone assignment sub-module tonological rules are applied to the rich lexical representation of the sentence so that the tone marking for the sentence can be corrected.

16. The tonological rule *FLL→LHL* is applied and subsequently the word is marked as /bhàlánì/.

Automatic phonetization

The following grapheme-to-phoneme rules taken from Table 3.1 are applied for this example in the automatic phonetization module:

17. bh →[b̥]
18. a → [a]
19. l → [l]
20. n → [n]
21. i → [i]

Please note that the grapheme-to-phoneme rules only apply to Xhosa text.

Finally, the output of the automatic phonetization sub-module is presented in the form of a narrow phonetic transcription of the sentence: [b̥àlánì peydž θri:] ! . The output of the automatic phonetization module forms the input information for the prosody generation module.

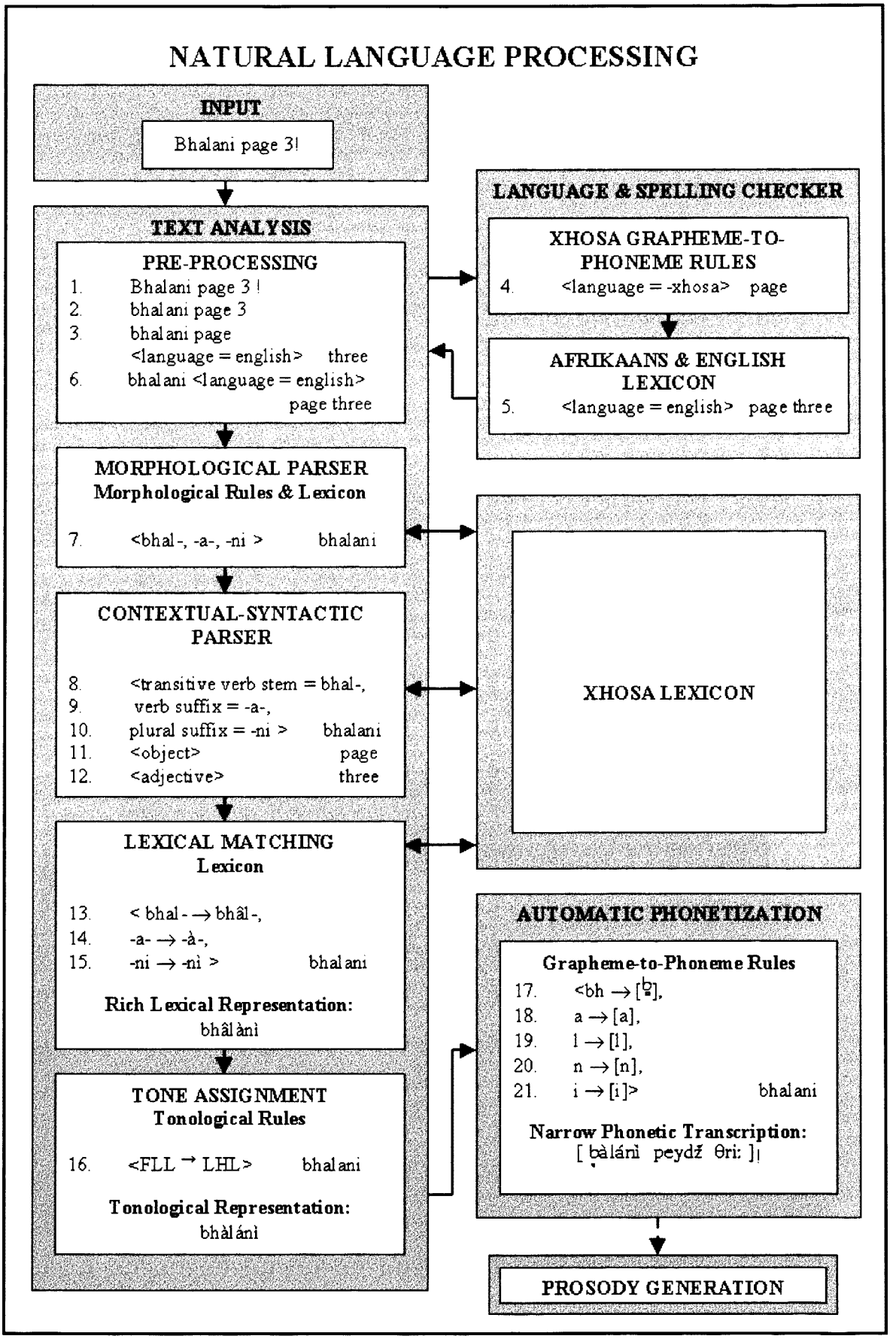


Figure 3.3 The NLP module for sentence 2: Bhalani page 3!

The next step in the NLP module would be to assign prosodic properties to the previously analysed text. The output of the automatic phonetization sub-module, as well as all the information that has been propagated from the text analysis sub-module forms the input for the prosody generation module. To demonstrate the function of a prosodic model for imperatives in Xhosa, it must first be determined how the parameters should be calculated. The following two chapters are devoted to this process. After essential acoustic and perceptual information has been acquired, an application of the results are presented in Section 5.10. From the language specific explanation and application of the NLP module presented in this section, we return to the theory of speech synthesis in general. The following section presents some relevant concepts concerning digital signal processing.

3.4 Digital Signal Processing

Returning to the theory of speech synthesis in general, the digital signal processing (DSP) module of a TTS system will now be discussed.

Speech synthesis concerns the signals that simulate human speech. In the DSP module the speech signal is generated by transforming the above mentioned high-level parameters into sound. In the same fashion in which the human dynamically controls the articulatory muscles and the vibratory frequency of the vocal folds when producing a speech signal, the digital signal generator must take articulatory constraints into account so that the output signal matches the input requirements (Dutoit, 1993:39). D'Alessandro and Liénard (1997:173) mention three types of speech signal generators (low-level synthesizers) that can achieve this task, namely:

- Articulatory synthesizers
- Formant synthesizers
- Concatenative synthesizers

The approach followed in articulatory and formant synthesizers, where the acoustic parameter values for the utterance are generated entirely by algorithmic means, is typically referred to as the *speech synthesis by rule* approach or *rule-based synthesis.* On the other hand, concatenative synthesis, where speech fragments (such as syllables or parts of syllables) are joined to produce the intended utterance, is known as *corpus-based synthesis*

(D'Alessandro & Liénard, 1997:174; Van Bezooijen & Van Heuven, 1997a; Hertz, 1999; Lange, 1993; Dutoit, 1997a).

Rule-based Synthesizers

Articulatory synthesizers employ physical models based on the physiology of speech production and the physics of sound generation in the vocal apparatus (D'Alessandro & Liénard, 1997:173; Fallside, 1994:4264). Accurately modelling the vocal tract and the patterns of airflow in it, requires the solution of complicated equations (Fallside, 1994:4265). This is a very slow process and, according to Fallside (1994:4265), it is therefore not well suited for the real-time requirements of speech synthesis.

Formant synthesis is a descriptive acoustic-phonetic approach to synthesis. D'Alessandro and Liénard (1997:173) explain that in this case the generation of speech "*is not performed by solving equations of physics in the vocal apparatus, but by modelling the main acoustic features of the speech signal*". An example of such an acoustic model is the source-filter model where the source of excitation is analogous to the excitation of the human vocal tract (e.g. by air passing through the glottis during voiced speech and causing the vocal cords to vibrate). Furthermore, the filter is analogous to the acoustic filter formed by the human vocal tract with its articulators such as the tongue and lips (Fallside, 1994:4264). As Dutoit (1993:42) explains: "*...formant synthesis is mainly concerned with the dynamic evolution of parameters related to formant and anti-formants frequencies and bandwidths together with glottal waveforms.*" This too proves to be a difficult and time-consuming approach that eventually impedes the production of real-time synthesized speech (Dutoit, 1993:42; Herz, 1999).

Many rule-based synthesizers such as the Klatt synthesizer (Klatt, 1980); MITalk (Allen et al., 1987); the JSRU synthesizer (Holmes et al., 1964) for English and the multilingual INFOVOX system (Carlson et al., 1982) have been widely integrated into TTS systems (Dutoit, 1993:42).

Corpus-based Synthesizers

The phenomenological models applied in corpus-based synthesizers intentionally refrain from referring to the human production mechanism (Dutoit, 1997a). In the case of

concatenative synthesis, segments of pre-stored natural speech are used as building blocks to construct an arbitrary utterance (Klatt, 1987:758; Fallside, 1994:4270). These speech segments are taken form a database of recorded speech (called the *Speech Corpus*) which is built to reflect the major phonological features of a language (D'Alessandro & Liénard, 1997:173). The speech recorded might either be pronounced naturally or it may be pronounced monotonously, so that the same pitch level is maintained and segment durations are pronounced equally throughout the utterance (cf. 5.5.1).

According to Dutoit (1993:44) a series of preliminary stages has to be fulfilled before the synthesizer can produce its first utterance. These stages of database preparation are summarized below:

1. Segments such as diphones, half-syllables and triphones are chosen as speech units since they involve most of the transitions and co-articulations while requiring an affordable amount of memory.
2. A list of words that correspond to the list of segments is compiled in such a way that each segment appears at least once.
3. This corpus of words is digitally recorded and stored and the elected segments are identified.
4. A segmental database centralizes the results in the form of the segment names, durations etc.
5. Segments are analyzed in a speech analyzer.
6. The output of the speech analyzer is segments that are stored in some parametric form in a *Parametric Segments Database*. Dutoit (1993:45) explains that the segments to be concatenated are usually extracted from different words, meaning that they occur in different phonetic contexts. This could result in amplitude mismatches, but these mismatches can be smoothed or *equalized*.
7. In the equalization process amplitude spectra are imposed upon the related endings of segments (Dutoit, 1993:45).
8. Speech is coded and stored in the *Synthesis Segments Database*.

The synthesis process begins with the deduction of a sequence of segments from the phonetic input of the synthesizer. In this *Segments List Generation* block, the Speech Segments Database is queried for global information on the units that it contains. In the

Prosody Matching block, prosodic events are assigned to the individual segments and the Synthesis Segments Database is queried for the actual parameters of the elementary sounds to be used (Dutoit, 1993:46).

In the *Segments Concatenation* block, these segments are then matched to one another by *smoothing* any discontinuities that might appear at their boundaries (Dutoit, 1993:46-47; Klatt, 1987:758). This smoothing process involves the modification of the fundamental frequency and duration of the segments so that they are more suitable for their new environment. A technique that can be used for this purpose and that has been found to be well suited for TTS synthesis is the Pitch-Synchronous OverLap-Add (PSOLA) algorithms (Fallside, 1994:4270) and in particular, Time-Domain Pitch-Synchronous OverLap-Add (TD-PSOLA) (Moulines & Charpentier, 1990; Dutoit & Leich, 1993 and Dutoit, 1993). According to Dutoit (1997a) the TD-PSOLA technique exhibits "*a very high speech quality (the best currently available) combined with a very low computational cost...*" which makes it a favourable synthesis technique to use.

After all the appropriate parameters have been assigned to the segments chosen form the database, and these segments have been matched and concatenated, the signal can be synthesized. Ideally the output would be intelligible and natural sounding speech.

The Digital Signal Processing module of a concatenative synthesizer that was shortly described above, is represented in Figure 3.4 (a flow chart taken from Dutoit, 1993:43).

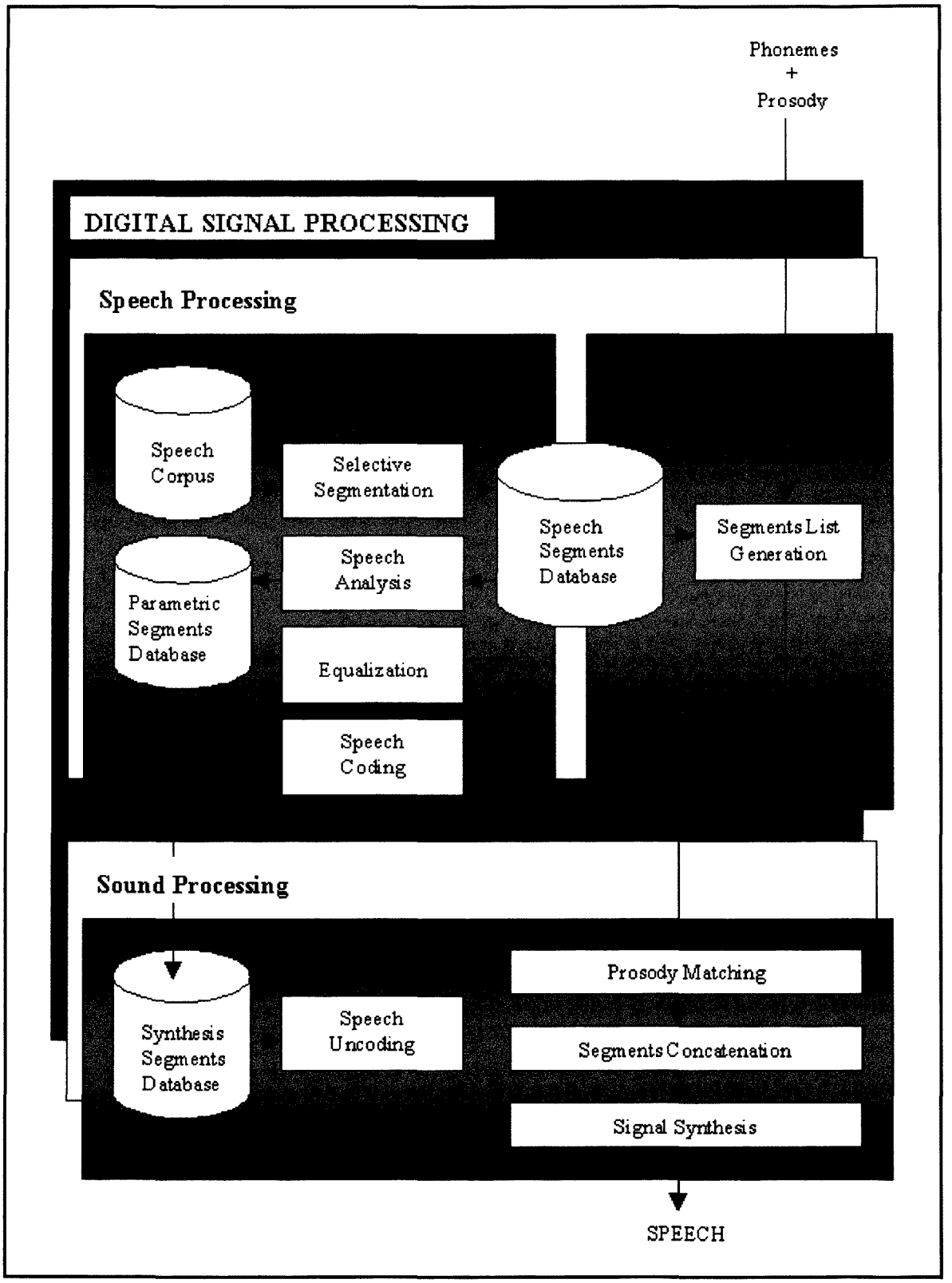


Figure 3.4 A general Concatenation Synthesizer TTS system. (cf. Dutoit, 1993:43).

Unlike the low-level synthesizer, the high-level synthesizer can function language independently. Choosing an appropriate low-level synthesizer does not depend on language to be synthesized, but rather on the purpose of the system.

Phoneticians and phonologists mostly favour rule-based synthesizers (Dutoit, 1993:40), as these systems can be used to study the characteristics of natural speech such as the physiology of speech production and the physics of sound generation. Consequently these systems provide them with the opportunity to test their phonetic and phonological theories and models.

Conversely, corpus-based synthesizers aim to generate high quality and natural sounding speech in real time. Concatenative synthesis is also a relatively cheap synthesis method and it is therefore more appropriate to use for service oriented purposes such as telebanking, e-mail retrieval, database query systems etc.

There is no doubt that service oriented speech technology should play a major role in the development of a multilingual country such as South Africa. When we consider the vast range of possible applications for a Xhosa TTS system (ref. Section 3.6), it should become clear that the development and application of such technology is long overdue. The advantages that concatenative synthesis hold, will contribute immensely in realising the goal of building a TTS system that can be integrated and applied as widely and as soon as possible. On these grounds concatenation is recommended as an appropriate synthesis method for a Xhosa TTS system.

Table A.1 in appendix A provides a summary of some speech synthesis systems. More information and demonstrations of these systems can be found on the Internet.

3.5 The Evaluation of Speech Synthesis Systems

The evaluation of spoken language systems constitutes a research area in itself, since the technology which is being evaluated is relatively new and methods of evaluation are still under development (Pols, 1994:4289). Evaluation of these systems are of interest to the

system developers, product buyers and end users who require specifications on system performance in absolute and relative terms. Absolute terms may for instance refer to percentage correct phoneme or word intelligibility scores. Relative terms may refer to factors such as which module sounds more natural for a specific application in a specific language, where different modules are compared to each other (Pols, 1994:4289; Pols, 1997:429; Hirschman & Thompson, 1991:409).

Pols (1991:387) states that: "*the quality of text-to-speech synthesis-by-rule systems is good enough for specific applications but is far from being anywhere near the quality of natural speech.*" Deficiencies occur at the different levels of processing concerning text input and spoken output. It follows that all these levels should be evaluated. The following table is taken from Pols (1994:4289). It lists the components and quality factors to be evaluated.

Table 3.3 Various system components and quality factors of a text-to-speech synthesizer (cf. Pols, 1994:4292).

Possible components
• text complexity, text preprocessing
• lexical search, morphological decomposition, grapheme-to-phoneme conversion, semantic analysis, syntactico-prosodic parsing, phrasing (syntactic boundaries), accentuation (sentence accent), speaking rate and rhythm
• intonation, duration, syllable boundary, word stress
• selection of unit for acoustic realization, spectro-temporal characteristics
• voice characteristics
• sound synthesizer
• system control strategy
Quality factors
• text interpretation, correct focus words, given/new information
• acceptable prosody
• word intelligibility
• phoneme intelligibility
• naturalness, general acceptability

Pols (1994:4292-4295) explains the different tests that can be implemented to evaluate a text-to-speech system by using the tests run on the MITalk-79 and DECtalk TTS systems as examples. Tests were used to measure phoneme and word intelligibility, speech understanding, word-processing and memory load (Pols, 1994:4292). Pols (1994:4293) mentions that very few tests are available at the linguistic level (see also Pols, 1991:392-403).

According to Pols (1994:4294) synthetic speech may still sound unnatural, even if phoneme and word intelligibility has reached an acceptable level of performance. This is due to the lack of natural prosody. In fact, Pols (1991:404) believes that "*the higher the basic phoneme, word and sentence intelligibility becomes, the more important prosody will become in the future for further improvement of the synthesized speech in terms of acceptability and naturalness*".

There is a great lack of standardized prosodic evaluation tests mainly because rule-synthesizers themselves are not yet very advanced in this area (Pols, 1997:430; Pols, 1994:4294; Pols, 1991:403). However, Pols (1997:430) mentions that concatenative synthesis in which units are taken from databases, may prove to be a way to obtain sufficient knowledge concerning detailed rules. Concerning evaluation, the approach followed currently is that of paired comparison procedures where preference judgements are made for one sentence over another by introducing small prosodic changes at a time (Pols, 1991:404; Satza et al., 1996). Other applied methods to quantify system performance are:

- Magnitude estimation (Pols, 1994:4294; Satza et al., 1996:650), where subjects directly estimate one or more aspects (such as acceptability, intelligibility, or naturalness) by assigning a positive number of their own choice to each utterance produced by each system.
- Categorical scaling (Satza et al., 1996: 650), which requires listeners to use interval scales for judging speech output.
- Mean opinion scores (MOS) regarding adequacy, naturalness, and appropriateness of specific aspects, such as final rise in fundamental frequency or segmental duration (Pols, 1994:4294; Satza et al., 1996:650), which uses a five or ten point scale for each question associated with a given voice parameter.

3.6 The Application of TTS Synthesis Systems

TTS synthesis systems can be applied in many different ways. It can be applied in voice-based communications such as:

- Telecommunications services
- Telebanking
- E-mail retrieval
- Database query systems
- Personal navigation systems
- Automated catalogue ordering systems
- Information retrieval systems (Bloothooft et al., 1995; Dutoit, 1993:19-20).

TTS technology could play a very important role in the support of manpower deployment (Laver, 1994:4284) where there is a need for messages (commands in particular) to be transmitted securely. This could be applied in areas ranging form the battlefield and the police service to emergency service operations.

TTS systems can also be applied for language education purposes (Bloothooft et al., 1995; Dutoit, 1993:19). For example, computer aided learning systems and translation output in spoken language interpretation.

Another useful application is that of aids to the handicapped. TTS systems can be used as speaking-aids to the speech-impaired in the form of a telephone relay service as well as reading-aids to the sight-impaired, where it can be coupled with optical character recognition (OCR) systems (Bloothooft et al., 1995; Dutoit, 1993:19-20).

Other applications include talking books and toys, man-machine communication in multimedia, vocal monitoring in the measurement of control systems, as well as fundamental and applied research (Dutoit, 1993:20).

Literally all of these applications may be considered as relevant and useful in our country and there is no doubt that the whole Xhosa speaking community will benefit from the development of a TTS system in one way or the other.

3.7 Chapter Summary

In this chapter speech synthesis and text-to-speech synthesis in particular was discussed. It was established that:

- Speech synthesis is concerned with the generation of speech by a speech output system.
- A TTS system uses textual information as input and converts it into speech output.
- TTS systems generally have two main components namely a natural language processing module (NLP) and a digital signal processing module (DSP).

Suggestions for an architecture for the NLP module of a Xhosa TTS system were presented. The text analysis, automatic phonetization and prosody generation modules were explained as part of the NLP module. Areas in the NLP module were identified where language specific features might influence the approach taken to text processing. The following alternative ideas were presented:

- Due to the fact that English and Afrikaans words might appear frequently in Xhosa, it was suggested that a sub-module be included that would function as a language and spelling checker.
- The morphological complexity of Xhosa should be taken into account when designing a morphological parser. Automatically derived morpho-phonological rules and a well-constructed lexicon may aid in the decomposition process.
- Because tone is a lexically distinctive feature, the units that make up the lexicon should be tone marked and tonological rules should be applied in the tone assignment sub-module.
- Grapheme-to-phoneme conversion for Xhosa can be done successfully using only rules, but it should be noted that the pronunciation problem of tone would prevail if a tone marked lexicon is not implemented.
- An alternative algorithm for grapheme-to-phoneme conversion was proposed. This algorithm maintains a high level of accuracy without sacrificing processing speed. It is also less complex and can easily be extended to other African languages. These properties make the algorithm more suitable for implementation in an automatic phonetization module.

Furthermore the essential part that prosodics play in natural sounding synthetic speech was explained. In order for synthetic speech not to sound monotonous and robot-like, an intonational, duration and loudness model should be designed and this should be used to apply the prosodic structure to the utterance. Designing such a prosodic model for Xhosa imperatives is in fact the main problem addressed in this study and particularly in the following two chapters.

Two methods of speech synthesis were discussed, namely rule-based and corpus-based synthesis. It was established that the corpus-based synthesis method produces high quality and natural sounding speech in real time and with relatively low complexity. Bearing these advantages in mind, it was concluded that corpus-based synthesis may be an appropriate method to use for a Xhosa TTS system in service oriented applications.

An overview regarding the evaluation of speech synthesis was given and it was found that there is a lack of standardized tests to evaluate the prosody modelling of these systems. Finally, a list of possible applications for a Xhosa TTS system was given.

Chapter Four

Acoustic Analysis of Imperatives in Xhosa

4.1 Chapter Overview

This chapter is divided into four main sections. Firstly, an explanation of the morphological structure of imperatives in Xhosa is given in the Introduction to Imperatives. This section also includes a summary of prosodic information (tonal structures in particular) available on this subject.

Secondly, the acoustic analysis that was conducted on a corpus of imperatives is discussed. In this section the systematic progression of the study is described through an explanation of:

- the aims of the acoustic analysis,
- the data used and
- the recording and preparation of the data for the acoustic analysis.

The third section concentrates on the acoustic analysis itself, as well as on the statistical analysis of the acoustic measurements.

In the fourth section the results obtained from the statistical analysis are discussed.

4.2 Introduction to Imperatives

The *imperative* in Xhosa is a verb mood which carries the meaning of a command (Wentzel et al., 1972:25; Pinnock, 1994:145). According to Wentzel et al. (1972:25) the imperative does not contain a subject concord. It does, however, contain a verb stem and a basic suffix /-a/. Other verbal attributes of the imperative are the distinction between the singular and plural form, as well as the distinction between the positive and negative form. As a direct command, the imperative is written with an exclamation mark (Wentzel et al., 1972:25).

Table 4.1 (taken from Davey, 1973:26) gives the grammatical patterns for the imperative, where:

(4.1) yi = prefix functioning as stabilizer
R = root / verb stem
a = basic suffix
OC = objectival concord
ni = plural suffix

e.g. *Bhalani* 'Write ye' → R = bhal
basic suffix = a
ni = plural suffix

Table 4.1 Grammatical patterns for the imperative.

Positive	**Singular**	**(a)**		**R-a**
		(b)	(y)	-R-a
		(c)	(y)i	-R-a
	With an OC		OC	-R-e
	Plural	**(a)**		**R-a-ni**
		(b)	(y)-	R-a-ni
		(c)	(y)i-	R-a-ni
	With an OC	OC-		R-e-ni

Pattern (a) occurs with verb stems that have more than one syllable.

Pattern (b) occurs with vowel-commencing verb stems.

Pattern (c) occurs with monosyllabic verb stems.

Table 4.2 contains examples of imperatives with different grammatical structures:

Table 4.2 Examples of imperatives with different grammatical structures.

Singular	**(a)**		**bhal-a**	'Write!'
	(b)	(y)	-oz-a	'Roast!'
	(c)	(y)i	-ty-a	'Eat!'
With an OC		Wa	-phek-e	'Boil it'
Plural	**(a)**		**bhal-a-ni**	'Write ye!'
	(b)	(y)-	oz-a-ni	'Roast ye!'
	(c)	(y)i-	ty-a-ni	'Eat ye!'
With an OC		Wa-	phek-e-ni	'Boil ye it!'

For the purpose of this thesis only the grammatical patterns written above in bold will be analyzed, i.e. *positive di-syllabic imperatives in the singular and plural context*, as well as the *transitive and intransitive forms without the objectival concord (OC).*

Tonal Features of Imperatives

According to Riordan's (1969:79) grammatical summary of imperatives in Xhosa, the suffix /-a/ always has a low tone in the singular context:

R-à

e.g. *Hámbà kákùhlé!* 'Go nicely - goodbye'

In the plural context the suffix /-a/ has a high tone and the plural affix /-ni/ has a low tone:

R-á-nì

e.g. *Hàmbánì kákùhlé!* 'Go nicely – goodbye'

Davey[8] (1973:30-32) gives the following rules concerning the tonal construction of di-syllabic verb stems:

a) The pluralizing suffix is low toned as with monosyllabic stems.

e.g. *Bhal-á-ni!* 'write ye!'

LL stems

b) The tone of these stems changes from LL to LH in the imperative.

e.g. *Chachá!* 'Recover!'

The aberrant LL stems

c) The LL tone of these stems changes to HL in the singular context and to HH in the plural context. (This was inferred from the tone marked examples in Davey, 1973:30-32.)

e.g. *Súka!* 'Go away!'
Súkáni! 'Go ye away!'

HL stems

d) The basic HL pattern of the stem is retained in the singular when there is no objectival concord present.

e.g. *Thémba ábántu!* 'Trust people!'

e) In the plural the H initial tone of the stem can be repeated on the stem final syllable or there can be tone shift from the first to the last syllable of the stem, i.e. it can

[8] Note that low tones are not explicitly marked in Davey's examples.

either have a HH or a LH sequence. LH is more common after a depressor consonant and HH is more common elsewhere.

e.g. *Zuzáni ímalí!* 'Get ye the money!'
Sálání kákuhlé! 'Stay ye well!'

FL stems

f) These stems retain their FL pattern before a juncture in the singular irrespective of whether or not there is an objectival concord present. In the plural the FL pattern changes to its morphotonemic variation of HH, as it is not final, i.e. the underlying tone is HH throughout.

e.g. *Lînga!* 'Try!'
Phákáni úkutyá! 'Serve ye the food!'

It should be borne in mind that the assignment of these tonal features as given above, are based on impressionistic observations. Neither Wentzel, Riordan nor Davey offer any other prosodic information (fundamental frequency, duration, amplitude) in their above mentioned work and the tonal rules given are not substantiated by any phonetic explanations.

To gather more information on the role of prosody in the production of imperatives, an acoustic analysis was done on a small corpus of verbs in different contexts.

4.3 Aims and Methods of Acoustic Analysis

4.3.1 Aims

Broadly stated, the primary aim of the acoustic analysis was to provide linguistic information at segmental and suprasegmental level concerning the production of commands. More specifically, the aim of this analysis was to determine *which* features speakers use and *how* they use these features to produce commands in different contexts. The expectation was to determine the relationship between pitch, duration and loudness

with reference to the degree of significance of these features in the production of commands. Investigating the relationship between the prosodic features might shed some light on whether information regarding pitch is the minimum information needed in order to convey the meaning of a command, or whether information about duration and loudness is also required.

The secondary aim was to derive prosodic patterns from the commands produced by a group of speakers. Specifically, the aim was to determine how the infinitive (neutral) form of the verb could be altered so that it can be perceived as a command. This information would then be used to formulate a prosodic model which may be implemented in a Xhosa text-to-speech system in order to generate natural sounding and intelligible commands.

As will be shown in the following sections, the corpus consisted of verbs in three different tone groups, each of which was analyzed separately. It should, however, be noted that the acoustic realization and perception of tonal movement within words were not addressed as this was considered to be beyond the scope of this investigation. Rather, average pitch features on phonemes and the relative pitch levels between phonemes were considered.

4.3.2 Data

The preparation of the data for acoustic analysis was done in the phases summarized below:

1. A corpus was selected.
2. The speech was recorded digitally.
3. Tags were added to mark phoneme boundaries.
4. The pitch on the voiced phonemes was calculated.

The actual acoustic analysis was subsequently conducted in different phases and will be discussed in detail in Section 4.4.

4.3.2.1 Corpus

The corpus consisted of 6 verbs in the infinitive mood, as well as the same 6 verbs in the imperative mood in 4 different contexts. Two verbs in each of three tonal groups namely Falling-Low (FL), High-Low (HL) and Low-Low (LL) were chosen. The imperatives were each recorded in Singular (without an object or adverb) and Singular (with an object or adverb) context, as well as Plural (without an objects or adverb) and Plural (with an object or adverb) context. Therefore 6 verbs in 5 different contexts were recorded for each of the eight speakers. This resulted in a total corpus for analysis of 240 words.

The data that were to be analyzed was limited to bi-syllabic (singular) and tri-syllabic (plural) words. Since the different measurements (pitch, duration and loudness) for the data recorded were to be compared to each other, it was important to choose words that were structurally similar. The phonetic structure of a word has particular implications for the measurement of duration, since different sounds have different innate lengths and the more sounds differ from each other, the more difficult it is to compare their lengths reliably. For this reason it was decided to select only words with a /C + ala/ structure. In other words, the stem of each verb would only differ with regard to its initial consonant and each verb would take the verbal suffix /-a/. Having the voiced consonant /l/ in the structure made it possible to measure the pitch contour from the first /a/ through the /l/ up to the end of the word without the interruption of any unvoiced sound. In the case of plural words, the structure was always /C + alani/. The advantages of these structures are that the number of variables is minimized while the number of samples is maximized. This combination enhances the reliability of statistical comparison results.

The 6 verbs and the different contexts in which they were recorded are listed in Table 4.3. The tone marked examples as they appear in Table 4.4 were taken from Westphal (1967:45-46) and Riordan (1969:75-79).

Table 4.3 Data recorded.

Infinitive	**Imperative:** Singular without an object or adverb.	**Imperative:** Singular with an object or adverb. (Command given to one person)	**Imperative:** Plural without an object or adverb. (Command given to more than one person)	**Imperative:** Plural with an object or adverb.
Ukubhala 'To write'	*Bhala !* 'Write !'	*Bhala unobumba !* 'Write a letter !'	*Bhalani !* 'Write ye!'	*Bhalani unobumba !* 'Write ye a letter !'
Ukudlala 'To play'	*Dlala !* 'Play !'	*Dlala apha !* 'Play here !'	*Dlalani !* 'Play ye!'	*Dlalani apha !* 'Play ye here !'
Ukulala 'To sleep'	*Lala !* 'Sleep !'	*Lala apha !* 'Sleep here !'	*Lalani !* 'Sleep ye!'	*Lalani apha !* 'Sleep ye here !'
Ukuhlala 'To sit'	*'Hlala !* 'Sit !'	*Hlala phantsi !* 'Sit down !'	*Hlalani !* 'Sit ye!'	*Hlalani phantsi !* 'Sit ye down !'
Ukubala 'To count'	*Bala !* 'Count !'	*Bala izinto!* 'Count the things !'	*Balani !* 'Count ye!'	*Balani izinto!* 'Count ye the things !'
Ukutsala 'To pull'	*Tsala !* 'Pull !'	*Tsala apha !* 'Pull here !'	*Tsalani !* 'Pull ye!'	*Tsalani apha !* 'Pull ye here !'

Table 4.4 Tonal patterns of the data.

Infinitive	**Imperative: Singular**	**Imperative: Plural**
ukubhâlà FL	*bhâlà* FL	*bhàlánì* LHL
ukudlâlà FL	*dlâlà* FL	*dlàlánì* LHL
ukulálà HL	*lálà* HL	*làlánì* LHL
ukuhlàlà LL	*hlálà* HL	*hlàlánì* LHL
ukubàlà LL	*bàlà* LL	*bàlánì* LHL
ukutsàlà LL	*tsàlà* LL	*tsàlánì* LHL

The phonetic transcription of the words is given in Table 4.5 below.

Table 4.5 Phonetic transcription of the data.

Infinitive	Imperative: Singular	Imperative: Plural
ukubhala [uk'u̥bala]	*bhala* [b̥ala]	*bhalani* [b̥alani]
ukudlala [uk'uɮala]	*dlala* [ɮala]	*dlalani* [ɮalani]
ukulala [uk'ulala]	*lala* [lala]	*lalani* [lalani]
ukuhlala [uk'uɬala]	*hlala* [ɬala]	*hlalani* [ɬalani]
ukubala [uk'uɓala]	*bala* [ɓala]	*balani* [ɓalani]
ukutsala [uk'uts'ala]	*tsala* [ts'ala]	*tsalani* [ts'alani]

4.3.2.2 Recording

A total of eight mother-tongue speakers of Xhosa were used to record the corpus. The subjects were all males below the age of 44 years.

The speakers were recorded directly onto the Computerized Speech Lab[9] (CSL) system of the Phonetics Laboratory at the University of Stellenbosch. The CSL system used is a model 4300B, version 5.04. The data were recorded at a sampling frequency (F_s) of 20 kHz and with 16 bits resolution.

Care was taken not to saturate the system while recording in order to avoid clipping of the digitized signals. The input level was unfortunately varied between speakers and therefore

[9] CSL/Computerized Speech Lab is a registered trademark of Kay Elemetrics Corp.

one can only compare the relative loudness levels of different speakers and not the absolute loudness levels.

The words were presented in random order and read in citation form with breaks between words. To avoid co-articulation effects, no carrier phrase was used.

Each word was saved in a separate file with a unique filename. Each character of the seven character filename serves as an information carrier and is implemented as follows: the second character of the filename is used for the speaker (A, D, J, M, P, T, X, Z) while characters three and four hold the tonal group FL, HL, or LL. Characters five and six hold the example number (01,02). The seventh character is either A for Infinitive, B for Singular Imperative without an object or adverb, C for Singular Imperative with an object or adverb, D for Plural Imperative without an object or adverb or E for Plural Imperative with an object or adverb.

4.3.2.3 Tagging and the Calculation of Duration

In order to calculate the duration of phonemes the boundaries between phonemes must be marked. The Multi-Speech[10] Signal Analysis Workstation (Model 3700, version 1.20), which is a Windows based version of the CSL, was used to annotate the speech signal with tags. These tags store the time instant of the tag as well as arbitrary ASCII text. Tags were inserted to mark the start of every phoneme and, where applicable, to mark Voice Onset Time (VOT) and Closure Duration (CD).

The tags were inserted manually in the Multi-Speech program. The speech signal was displayed in one window and the spectrogram in another, larger window. LPC derived formant tracks were sometimes superimposed on the spectrogram to aid in identifying phoneme boundaries. Tagging was done both visually, using the spectrogram and the signal, and auditorally by listening to the marked speech segments. All words for all eight speakers were tagged in the same manner. Figure 4.1 below illustrates how the signal and spectrogram was displayed in the Multi-Speech program. The small 'T' marks above the signal represent the tags inserted at each phoneme boundary. Tag names and information

[10]Multi-Speech is registered trademark of Kay Elemetrics Corp.

are not displayed in the window, but the tags may be queried at any time, in which case the information will appear in a separate window.

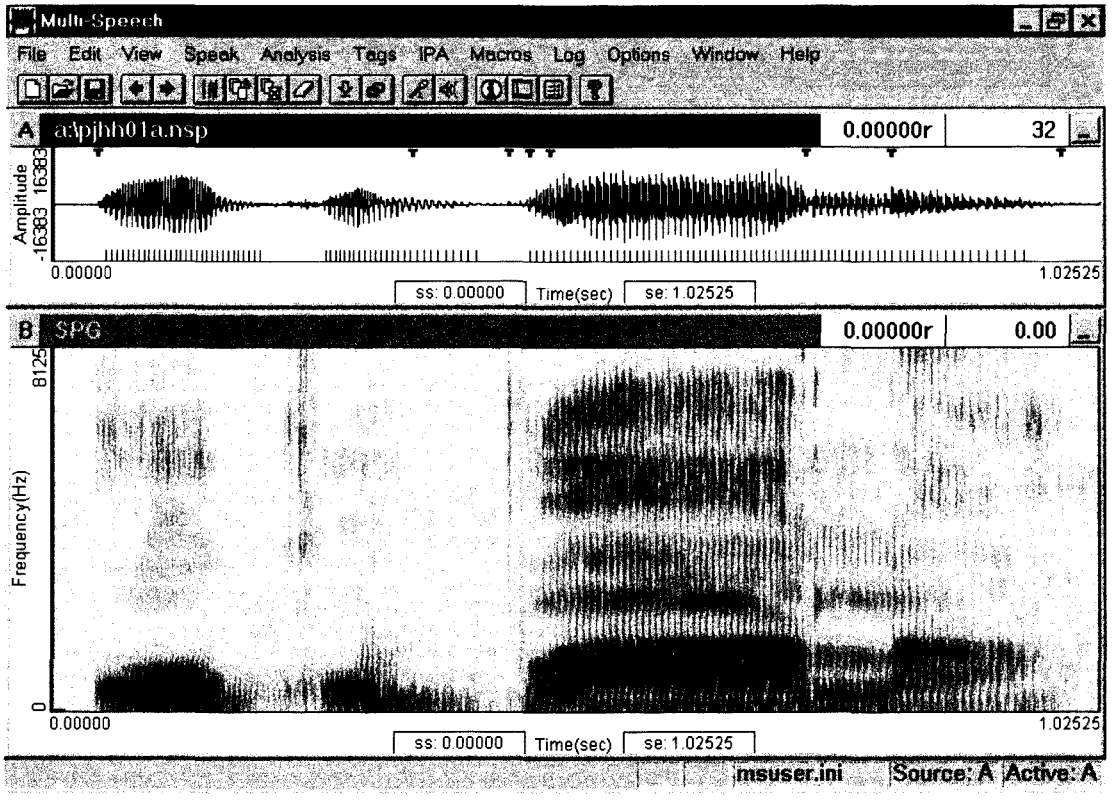


Figure 4.1 Tagging the word *ukubhala* in Multi-Speech.

For all contexts the verb stem was tagged, but not the object or adverb following the verb. In the case of the plural context, the /ni/ morpheme was tagged as well. In order to calculate CD for the consonants the /uku/ morpheme of the infinitive was also tagged.

A basic tagging system was developed for the annotation of the words. Orthographic characters were used to represent the phonemes. When a phoneme occurred more than once, the phonemes were numbered. The '] ' character was used to indicate the end of the word.

It was mentioned in Section 2.3 that the innate length may influence the absolute duration of a phoneme. At the same time, the point and manner of articulation of the phoneme itself, as well as the sounds that surround it, condition the innate length of the sound. As will become clear from the following definitions, CD and VOT are the acoustic features that can actually be measured so as to investigate the innate length of consonants. In the

tagging process, CD and VOT were identified for the initial consonant of each verb stem. CD and VOT were determined based on the following definitions:
According to Kent and Read (1992:106) stop consonants have both a closure phase and a release phase. For a postvocalic stop consonant, the duration of closure or the *stop gap* is the interval between the end of the previous vowel and the point of release of the oral closure. Voice Onset Time, as defined by Kockaert & Godwin (1996:1), is "*the time interval between the release of the oral closure and the onset of regular laryngeal pulsation for the following vowel*". VOT is also described by Lieberman and Blumstein (1998:215) as "*the timing relation between the release of a stop consonant and the onset of glottal excitation*".

Six different word-initial consonants occur in the recorded data namely a bilabial stop consonant /bh/, an implosive stop consonant /b/, two fricatives /dl/ and /hl/, a liquid /l/ and an affricate /ts/. CD was marked for /bh/, /b/ and /ts/ in the infinitive context, as this is the only context in which these consonants are preceded by a vowel. Due to the acoustic nature of the implosive /b/, CD can also be observed for this sound when it is not preceded by a vowel. Therefore CD was marked for the implosive /b/ in the imperative contexts as well. VOT was marked for the two stop consonants /bh/ and /b/. When tagging the two fricatives /dl/ and /hl/ as well as the liquid /l/, CD and VOT were not observed. In these cases the consonants were not tagged in such detail.

Table B.1 in Appendix B, provides details of the annotation system used for each word. The tagging process and the implementation thereof which applies to all words, will now be explained using the example *ukubhala* or *bhala*. Due to the differences between the word-initial consonants, the annotation system was adapted for the different consonants. The detailed tagging of the consonants will also be explained. Section 4.4.1 illustrates how these tags were used to measure duration, pitch and loudness automatically.

Words and syllables

The tags for *ukubhala* would be:

(4.2) 'u1 u2* bh bha1 a1 l a2]'

The description of these tags are given in Table 4.6.

Table 4.6 Description of tags for the words *ukubhala* and *bhalani*.

Tag	**Description** ***ukubhala***	**Tag**	**Description** ***bhalani***
u1	beginning of first /u/ of /uku/	bh	burst of /bh/
u2*	end of second /u/ of /uku/	bha1	beginning of first /a/
bh	burst of /bh/		(tagged at onset of voice)
bha1	beginning of first /a/	a1	first /a/
	(tagged at onset of voice)		(tagged at beginning of steady state)
a1	first /a/	l	beginning of /l/
	(tagged at beginning of steady state)	a2	beginning of second /a/
l	beginning of /l/	n	beginning of /n/
a2	beginning of second /a/	i	beginning of /i/ for all plural words
]	end of word	]	end of word

For the infinitive form of *bhala* the duration of the verb is the time difference between the 'bh' tag and the ']' tag. (Note that the /uku/-morpheme is not included in the total duration.) The duration of *bhala* would be:

(4.3) duration(/bhala/) = starting time(']') – starting time('bh')

The starting time referred to in (4.3) is the time duration relative to the beginning of the speech file. This will be written more compactly as follows:

(4.4) $d_{bhala} = t_{]} - t_{bh}$

The duration of a phoneme is the time difference between the tag for that phoneme and the next tag:

(4.5) $d_{a1} = t_{l} - t_{a1}$

The duration of the /bha/ syllable would then be given by:

(4.6) $d_{bha} = t_{l} - t_{bh}$

The duration of the /ala/ part of each word was calculated from the steady state of the first /a/ to the end of the second /a/. Therefore, the duration of /ala/ would be given by:

(4.7) $d_{ala} = t_{]} - t_{a1}$

or for the plural form:

(4.8) $d_{ala} = t_{n} - t_{a1}$

The duration of the transition from the consonant to the first /a/ was measured from the onset of voice after the consonant to the beginning of the steady state of the first /a/. For *bhala*, for example, the duration of the transition would be given by:

(4.9) $d_{bha1} = t_{a1} - t_{bha1}$

Consonants

To calculate CD for the /bh/ of *bhala*, the infinitive form was tagged at the end of the second /u/ of /uku/ and at the beginning of the burst of /bh/. These tags were then named 'u2*' and 'bh' respectively. The CD of the /bh/ would then be given by:

(4.10) $CD_{bh} = t_{bh} - t_{u2*}$

This method of annotation was also used for the initial consonant of *bala* in the infinitive context. To calculate CD for /b/ in the imperative context the closure of the lips was annotated with a tag called '*b' and the CD would be given by:

(4.11) $CD_{b} = t_{b} - t_{*b}$

To calculate CD for the affricate /ts/ the infinitive form was tagged at the end of the second /u/ of /uku/ and at the beginning of the release of /ts/. The CD of the /ts/ would then be given by:

(4.12) $CD_{ts} = t_{ts} - t_{u2*}$

To calculate VOT for /bh/ a tag was positioned at the onset of voice of the first /a/ of /ala/. This tag was named 'bha1'. The duration of the VOT for /bh/ would then be given by:

(4.13) $d_{VOTbh} = t_{bha1} - t_{bh}$

This method of annotation was also used for the implosive /b/.

Tagging the two fricatives /dl/ and /hl/ as well as the liquid /l/ can be explained by using the /dl/ of *dlala* as an example. The duration of /dl/ would be given by:

(4.14) $d_{dl} = t_{dla1} - t_{dl}$

where d_{dl} is interpreted as the duration of the frication of /dl/. In the same way d_l is interpreted as the duration of the liquid /l/, d_{hl} as the duration of the frication of /hl/ and d_{ts} as the release phase of /ts/.

The tags for one speaker were checked thoroughly by hand. After all mistakes were corrected, the tags for the other seven speakers were compared against that of the corrected speaker. The sequence of tags for a particular word is identical for all speakers. If a specific speaker did not pronounce a phoneme, the tag for that phoneme is still present, but it is located very close to the next tag. For example, speakers often dropped the final vowel. Therefore, the 'a2' tag was placed very close to the ']' tag:

(4.15) $t_] - t_{a2} \approx 0$

The 'Note' field of the Multi-Speech speech file format was used to store the orthographic transcription of the recorded speech. For example, if the word is *Bhala!* the text 'Word=Bhala!' was stored in the Note field.

4.3.2.4 Pitch Calculation

The CSL offers two pitch calculation algorithms. Both operate on the speech signal in the time domain. The *Pitch* command finds the average pitch in a frame by locating positive

and negative peaks, while the *Impulse* command sequentially locates individual pitch periods in the signal.

The *Pitch* command produces fairly accurate pitch estimates in the steady state parts of voiced speech but performance degrades during voiced/unvoiced transitions and in low energy regions. Another major disadvantage of the *Pitch* command is that pitch estimation errors cannot be corrected by hand. Using the *Impulse* command, one can edit pitch markers, but this is a time consuming and error prone task since the initial pitch estimates are often in error or noisy.

For the reasons stated above, a pitch extraction algorithm[11] was developed in a mathematical package called MATLAB. Figure 4.2 below illustrates the output of this algorithm for a voiced speech segment. The algorithm adaptively low pass filters the original speech signal (red) to emphasize pitch periods (blue signal). Pitch period markers (green circles) are then inserted at the negative-to-positive zero crossings of the filtered signal. Some refinement is still necessary to make the voiced/unvoiced decision and to minimize pitch-doubling errors. At this time these errors were corrected manually in MATLAB by deleting pitch markers. The algorithm is able to track rapid changes in pitch accurately.

Suppose the start and end of a particular pitch period is denoted by two tags, t_{p1} and t_{p2}, respectively, then the pitch at time instant t_{p1} , i.e. $F0(t_{p1})$, is the reciprocal of the pitch period duration:

$$F0(t_{p1}) = \frac{1}{t_{p2} - t_{p1}} \text{ Hz} \tag{4.16}$$

The pitch markers were calculated for all words and saved with the speech signals. After pitch errors were corrected the data were saved on CD-ROM. This comprised the final data set used for further analysis.

[11]Unpublished work, Mr. J. A. N. Louw.

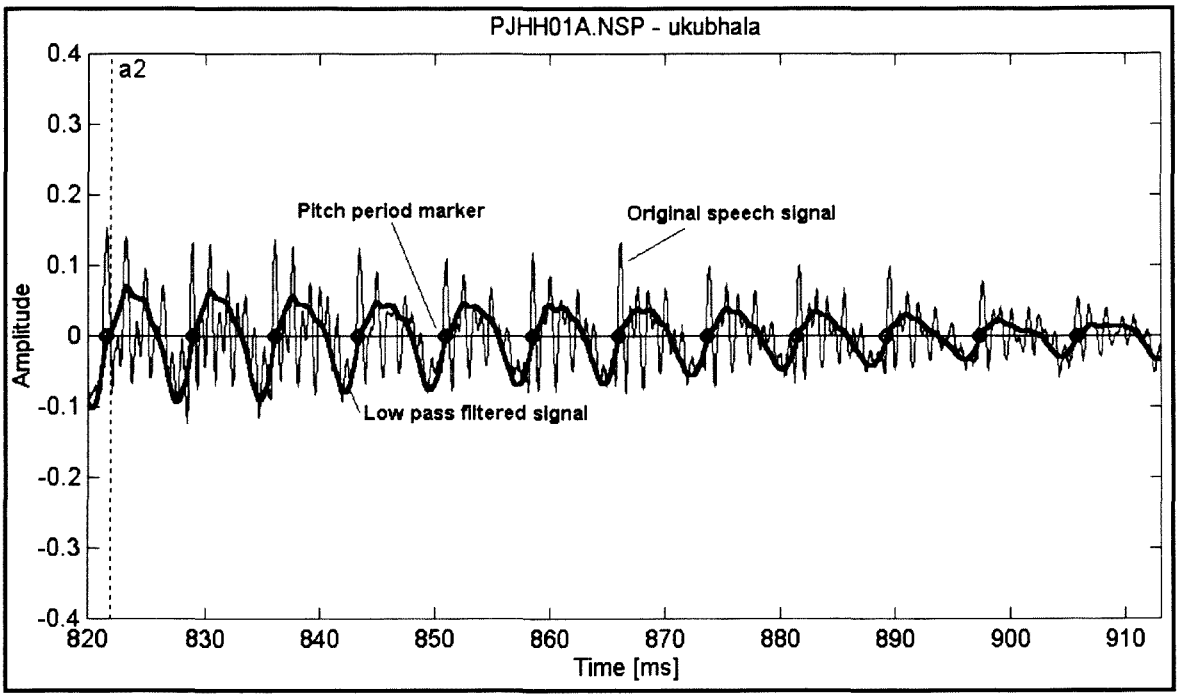


Figure 4.2 Output of the pitch extraction algorithm developed in MATLAB.

4.3.2.5 Loudness Calculation

The energy of the speech signal over a window of length N samples is given by:

$$(4.17)\quad \sum_{n=1}^{N} x^2(n)$$

where $x(n)$ is the speech signal. The energy will be greater for a longer phoneme than for a shorter phoneme of the same intensity. The objective was not to compare the energy, but to compare loudness (amplitude). Therefore, we defined *loudness* as the normalized energy by dividing with the number of samples in the window. The absolute loudness E_k in the kth window of length N samples was therefore calculated as:

$$(4.18)\quad E_k = \frac{1}{N}\sum_{n=1}^{N} x^2(n)$$

Since sound input levels varied between speakers, only *relative* and not absolute loudness levels were compared. The relative loudness e_k was given by:

$$(4.19)\quad e_k = \frac{E_k}{E_{\max}}$$

where

$$(4.20)\quad E_{\max} = \max(E_k) \qquad \text{for all } k$$

Note that the relative loudness is dimensionless.

4.4 Analysis

The acoustic features to be analyzed, namely duration, pitch and loudness are presented first, followed by a discussion of the comparative analyses that were performed. Eight intra-contextual comparisons were made to determine how the infinitive and the four imperative contexts differ for a particular verb stem.

4.4.1 Acoustic Features

The extraction of acoustic features to be analysed are discussed below.

1. Duration
2. Pitch
3. Loudness

4.4.1.1 Duration

A custom written program[12] calculated the duration of phonemes automatically, using the tag positions. The following duration features were calculated for the word *ukubhala* for example:

(4.21) $d_{u1}\ d_{u2*}\ d_{bh}\ d_{bha1}\ d_{a1}\ d_{l}\ d_{a2}$

[12] 'ViewTags' developed by Mr. J.A.N Louw.

The total duration of the verb stem (d_{bhala}) was also calculated. All duration measurements are in seconds. These feature vectors were calculated for all speakers and all comparisons.

Figure 4.3 below illustrates the duration measurements made from the tag inserted for the word *ukubhala*. The red lines indicate the tag positions.

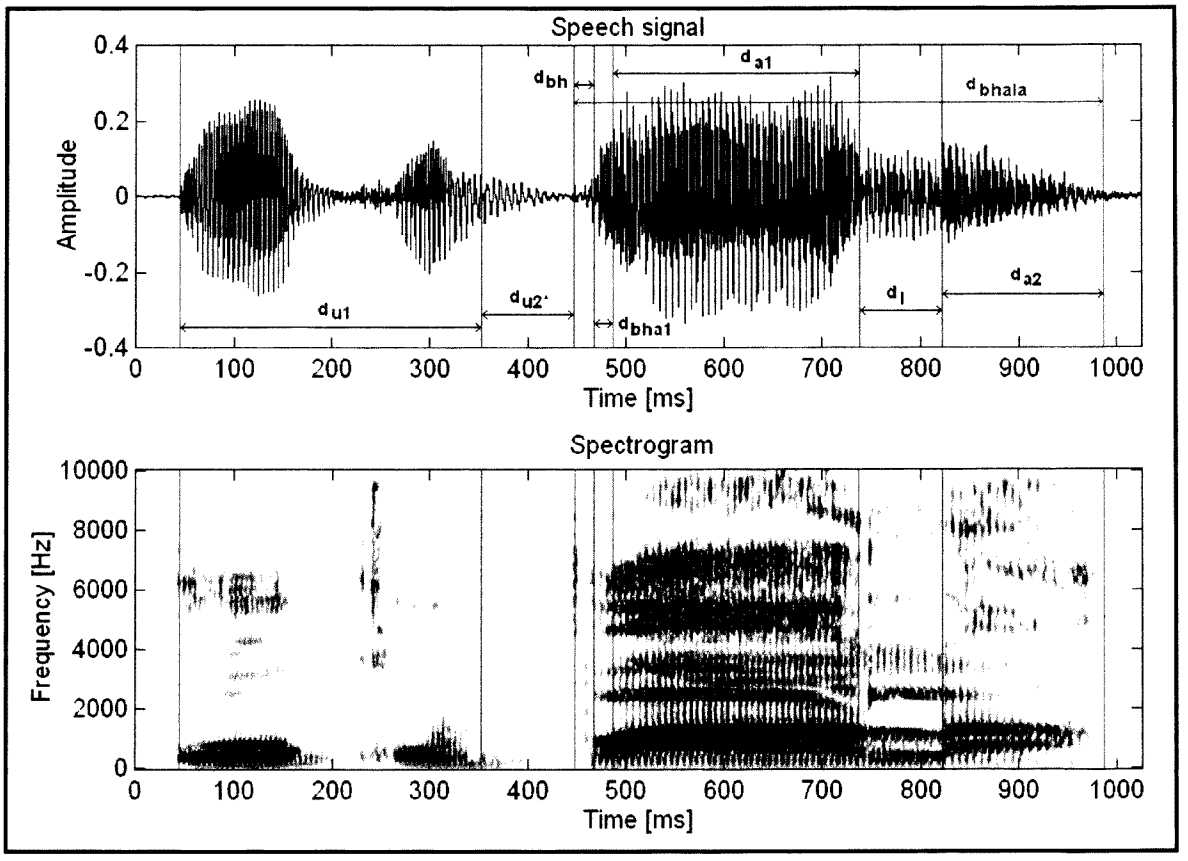


Figure 4.3 Duration measurements and tag positions shown in the speech signal and the spectrogram of the word *ukubhala*.

4.4.1.2 Pitch

Pitch was calculated only on the part of the verb following the initial consonant (i.e. /ala/ or /alani/). Each phoneme, located by means of the tags, yielded several pitch measurements, which are first smoothed with a third order median filter. The average is then computed and taken as the pitch feature for that phoneme.

Occasionally a speaker did not clearly articulate the final vowel of utterances resulting in unreliable pitch measurements for these phonemes. When representing this type of data vector with a single number, the median gives a more reliable representation than the

mean, because it tends to eliminate the effect of outliers. Therefore the phoneme pitch for a particular word over all eight speakers was given by the median of the speakers' mean pitch values.

The following pitch features were calculated for the word *ukubhala* for example:

(4.22) p_{a1} p_l p_{a2}

The average pitch of the word (p_{ala}) was also calculated. All pitch measurements are in Hz. The pitch algorithm employed calculated the pitch on each pitch period. Several pitch values were calculated between the two tags demarkating the boundaries of a particular phoneme. The average of these pitch values were taken as the pitch feature for that phoneme. This is illustrated in Figure 4.4 below where the (green) horizontal lines indicate the average pitch for each voiced phoneme. Additionally, the average pitch (pink) over the /ala/ (or /alani/) part was calculated.

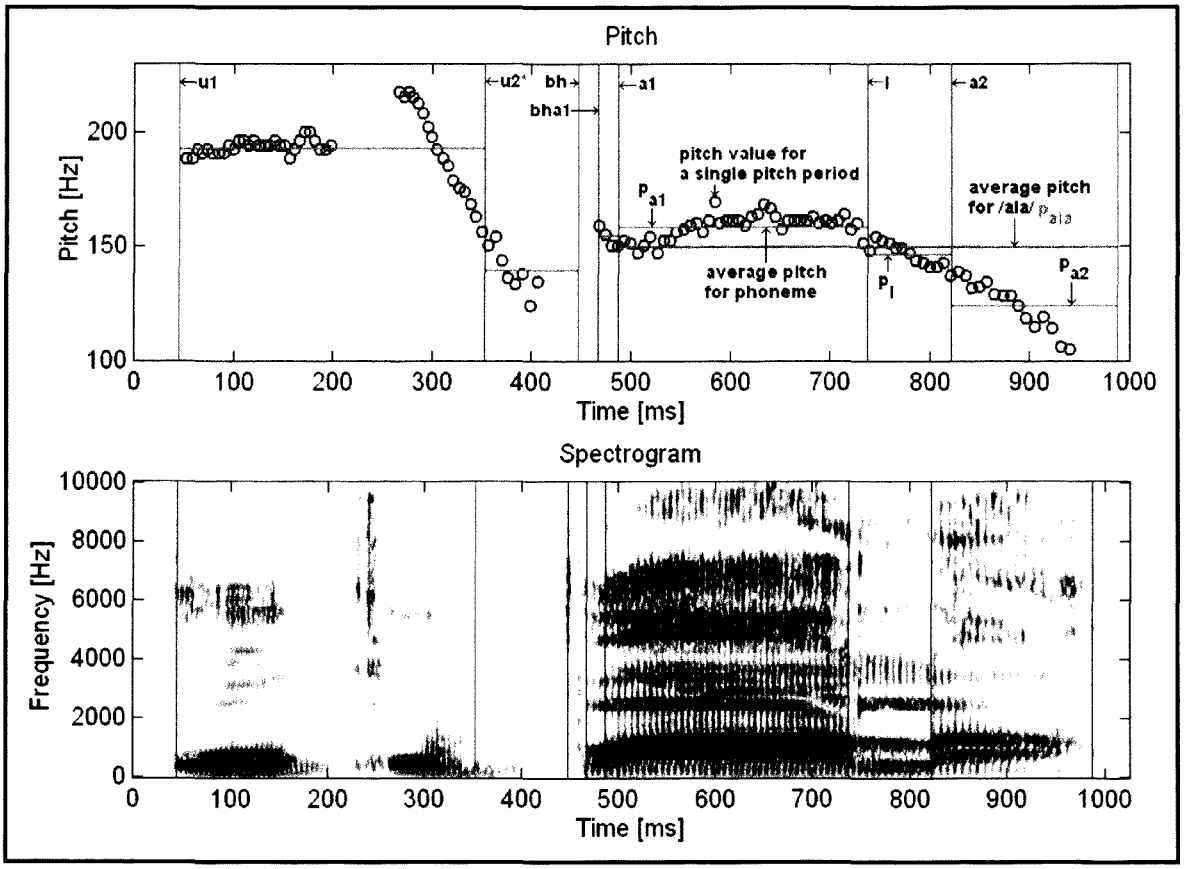


Figure 4.4 Pitch measurements and tag positions shown in the speech signal and the spectrogram of the word *ukubhala*.

4.4.1.3 Loudness

The loudness values were calculated in 20 ms non-overlapping windows. Each phoneme, located by means of the tags, yielded several loudness measurements, of which the average was taken as the loudness feature for that phoneme. These loudness features are dimensionless and indicate the relative loudness of each phoneme within a word.

The following loudness features were calculated for the word *ukubhala* for example:

(4.23) $e_{u1}\ e_{u2*}\ e_{bh}\ e_{bha1}\ e_{a1}\ e_{l}\ e_{a2}$

Figure 4.5 below illustrates these concepts for the word *ukubhala*.

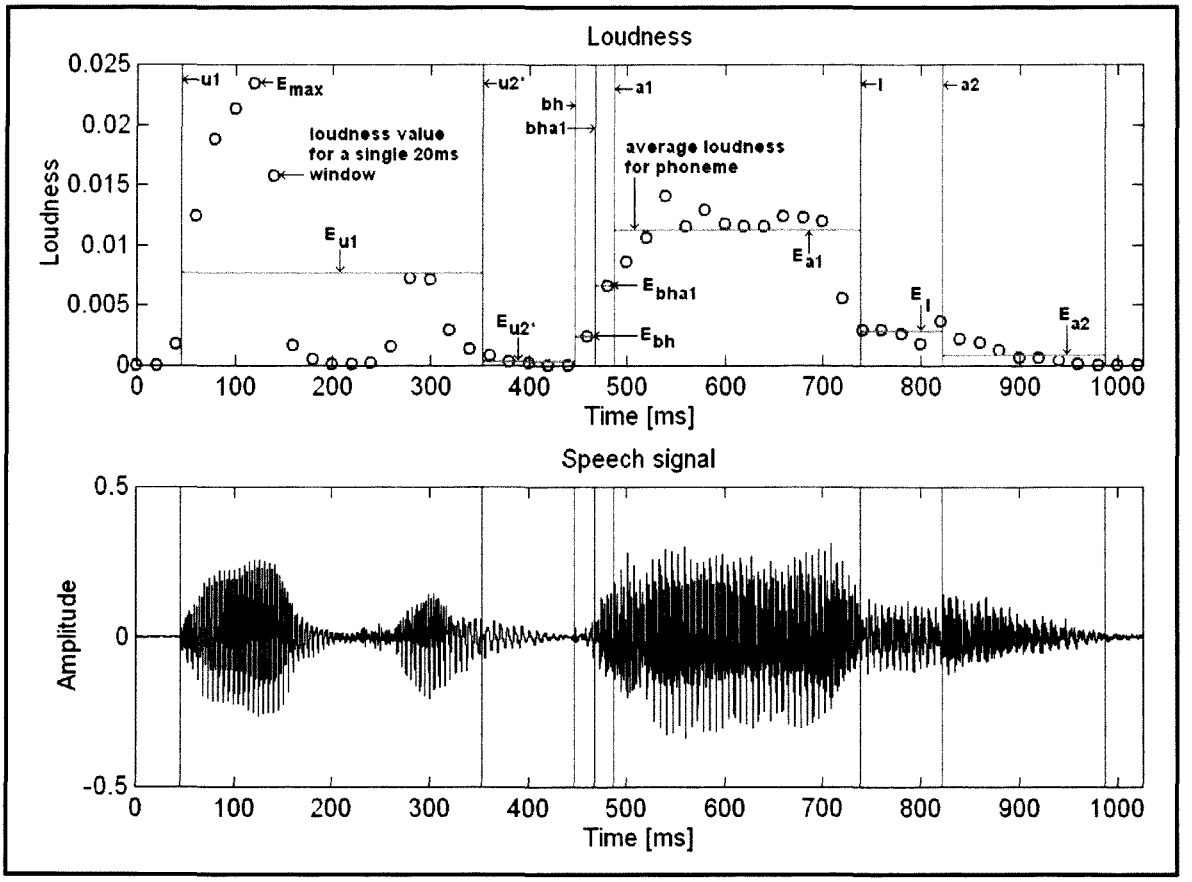


Figure 4.5 Loudness measurements and tag positions shown in the speech signal and the spectrogram of the word *ukubhala*.

4.4.2 Intra-Contextual Comparisons

The aim of the statistical analysis was to determine *which* prosodic features speakers use and *how* they use these features to produce /C+ala/ imperatives in different contexts. The results could ultimately be used to describe how the prosodic features of a verb in the neutral infinitive form could be altered in order to be perceived as an imperative in a specific context or how an imperative in a specific context could be altered to be perceived as an imperative in another context. In other words, this statistical analysis aims to provide rules that can be applied in order to change, for example, *bhala* in the infinitive context, to *bhala* in the singular context without an object or adverb.

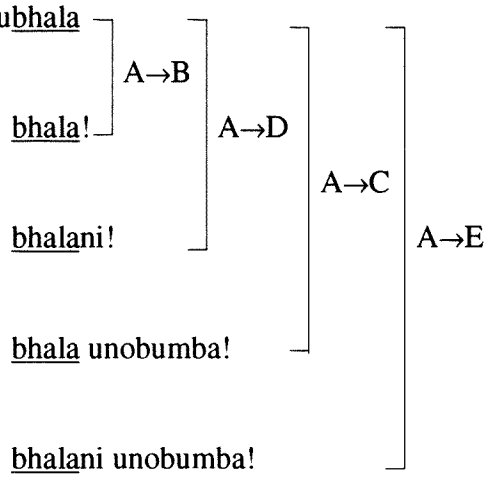
Table 4.7 contains the list of comparisons between the different contexts which were analyzed. The first column of this table gives the contexts that are compared to each other. The arrow can be read as 'compared to'. For example, A→B means that the verb in the infinitive context (called 'A') is compared to the verb in the singular imperative without object or adverb context (called 'B').

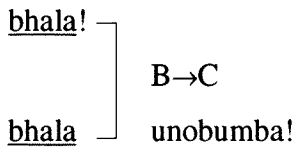
Table 4.7 Intra-contextual comparisons.

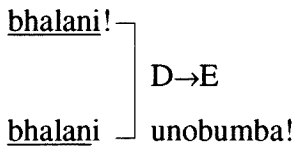
A→B	Infinitive	Compared to:	Singular Imperative without object or adverb
A→C	Infinitive	Compared to:	Singular Imperative with object or adverb
A→D	Infinitive	Compared to:	Plural Imperative without object or adverb
A→E	Infinitive	Compared to:	Plural Imperative with object or adverb
B→C	Singular Imperative without object or adverb	Compared to:	Singular Imperative with object or adverb
B→D	Singular Imperative without object or adverb	Compared to:	Plural Imperative without object or adverb
C→E	Singular Imperative with object or adverb	Compared to:	Plural Imperative with object or adverb
D→E	Plural Imperative without object or adverb	Compared to:	Plural Imperative with object or adverb

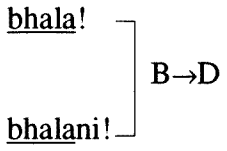
The different contexts in which the verb *bhala*, for example, may occur, was compared to each other in the following ways:

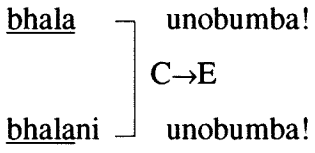
(4.24)











Statistical Features Calculated for Each Context

In order to compare the properties of phonemes between different contexts, features should be extracted from the acoustic signal and ranked in order of significance. For every word w a number of features f can be computed. Different contexts are compared in pairs. Denote the feature data vector for the first context form in the pair being compared as X_{wf}. This feature data vector consists of 8 samples since there are 8 speakers. The corresponding feature data vector for the second context form of the pair is called Y_{wf}.

As an example, consider the comparison between the duration of the infinitive context (A) and that of the singular imperative with object (C) for the phoneme /a1/ of the word *bhala*. In this case the X vector corresponds to A while the Y vector corresponds to C. The word w is *bhala* and the feature f being compared is the duration of the first /a/, i.e. d_{a1}. This notation is illustrated below:

(4.25)

$A_{\text{bhala } d_{a1}}$	$C_{\text{bhala } d_{a1}}$
$X = A$	$Y = C$
$w = \text{bhala}$	$w = \text{bhala}$
$f = d_{a1}$	$f = d_{a1}$

The raw duration measurements for the X_{wf} and Y_{wf} vectors are tabulated below.

Table 4.8 Raw duration measurements of the A and C contexts for the first /a/ of the word *bhala*.

Speaker	$X_{wf} = A_{\text{bhala } d_{a1}}$	$Y_{wf} = C_{\text{bhala } d_{a1}}$
A	337.9 ms	111.7 ms
J	250.3 ms	214.5 ms
L	251.6 ms	155.9 ms
M	271.4 ms	129.4 ms
P	227.9 ms	152.5 ms
T	245.9 ms	162.2 ms
X	257.3 ms	139.8 ms
Z	235. 5 ms	192.9 ms

The following parameters were calculated for each data vector X_{wf} and Y_{wf} by applying a function $f(\ )$:

(4.26)

- minimum
- maximum
- mean $\overline{\mu}$
- median $\tilde{\mu}$
- standard deviation σ

Table 4.9 shows the different functions computed on each data vector X_{wf} and Y_{wf}, as well as the numerical result.

Table 4.9 Parameters computed on the data vectors for the A and C contexts of the first /a/ of the word *bhala*.

Function	$f\left(A_{\text{bhala d}_{\text{a1}}}\right)$	$f\left(C_{\text{bhala d}_{\text{a1}}}\right)$
maximum	337.9 ms	214.5 ms
minimum	227.9 ms	111.7 ms
median $\tilde{\mu}$	251.0 ms	154.2 ms
mean $\overline{\mu}$	259.7 ms	157.3 ms
standard deviation σ	34.2	33.3

Statistical Features Calculated for Context Pairs

A statistical measure is required to determine whether X_{wf} and Y_{wf} originate from different distributions, or otherwise stated, whether X_{wf} and Y_{wf} are significantly different. If X_{wf} and Y_{wf} originate from the same distribution that feature is not significant in distinguishing between the verbs in different contexts. In order to compare the two distributions a new data vector D_{wf} was defined as the difference between the data vectors for each context:

(4.27) $D_{wf} = X_{wf} - Y_{wf}$

The notation A→C in (4.28) signifies that although D_{wf} is a single data vector, it represents a compact comparison between two different data vectors, or two different contexts.

(4.28) $A \rightarrow C_{\text{bhala d}_{\text{al}}}$

$D = \text{A} \rightarrow \text{C}$

$w = \text{bhala}$

$f = \text{d}_{\text{al}}$

Table 4.10 Differences of duration measurements of the A and C contexts for the first /a/ of the word *bhala*.

Speaker	$D_{wf} = A \rightarrow C_{\text{bhala d}_{\text{al}}}$
A	226.2 ms
J	35.8 ms
L	95.7 ms
M	142.1 ms
P	75.5 ms
T	83.7 ms
X	117.5 ms
Z	42.6 ms

The following parameters were calculated for data vector D_{wf}:

(4.29)

- mean $\overline{\mu}$
- median $\tilde{\mu}$
- standard deviation σ
- $\dfrac{\text{mean}}{\text{standard deviation}} = \dfrac{\overline{\mu}}{\sigma}$
- Wilcoxon signed rank test

Parametric statistical methods such as the normal distribution are inappropriate for small sample sizes. In this study with feature pairs of 8 samples each, a non-parametric method was used, namely the *Wilcoxon signed rank test*. The test and its application in finding the most significant features will be discussed next.

Table 4.11 Parameters computed on the difference data vector for the A and C contexts of the first /a/ of the word *bhala*.

Parameter	$f\left(A \rightarrow C_{\text{bhala d}_{\text{al}}}\right)$
Wilcoxon signed rank test	36
median $\tilde{\mu}$	89.7 ms
mean $\overline{\mu}$	102.4 ms
standard deviation σ	61.2
$\frac{\text{mean}}{\text{standard deviation}} = \frac{\overline{\mu}}{\sigma}$	1.67 ms

4.4.2.1 Wilcoxon Signed Rank Test

The Wilcoxon signed rank test is a non-parametric method that can be used to test the null hypothesis that the distribution of the differences between paired samples has a zero median value. The test uses both the signs and magnitudes of the differences (Solomon, 1996:299). If the null hypothesis is rejected, it means in this application that the feature is significant. A magnitude of significance S can be computed. The larger the value of S is, the more significant a particular feature is.

4.4.2.2 Mean-Standard Deviation

There are instances where the Wilcoxon test assigns equal significance values to different features. The magnitude of the mean divided by the standard deviation $\left|\frac{\overline{\mu}}{\sigma}\right|$ may be used to resolve which feature is more important. This value will be larger for features that differ more significantly. The sign of $\frac{\overline{\mu}}{\sigma}$ is also useful: positive values show that the numerical value of this feature is higher for the verb in the first context than for the verb in the second context of the comparison, while the negative implies the opposite.

Example of Comparison Between A and C Contexts

The example to be presented next illustrates the use of the Wilcoxon test and the $\frac{\overline{\mu}}{\sigma}$ in determining the significant differences between the duration values of two contexts for all phonemes of the word *bhala*.

Table 4.12 Parameters computed on the difference data vector for the A and C contexts of the phoneme durations for the word *bhala*.

Parameter	$f\left(A \rightarrow C_{\text{bhala}}\, d_{\text{bh}}\right)$	$f\left(A \rightarrow C_{\text{bhala}}\, d_{\text{bha1}}\right)$	$f\left(A \rightarrow C_{\text{bhala}}\, d_{\text{a1}}\right)$	$f\left(A \rightarrow C_{\text{bhala}}\, d_{\text{l}}\right)$	$f\left(A \rightarrow C_{\text{bhala}}\, d_{\text{a2}}\right)$
Wilcoxon signed rank test S	35	0	36	36	0
median $\tilde{\mu}$	9.2 ms	10.9 ms	89.7 ms	23.9 ms	-16.0 ms
mean $\overline{\mu}$	8.9 ms	4.9 ms	102.4 ms	20.5 ms	-11.9 ms
standard deviation σ	5.8	14.8	61.2	11.7	38.9
$\frac{\text{mean}}{\text{standard deviation}} = \frac{\overline{\mu}}{\sigma}$	1.53 ms	0.33 ms	1.67 ms	1.75 ms	-0.31 ms

The following conclusions may be drawn from Table 4.12:

- Duration features d_{bha1} and d_{a2} are not significant according to the Wilcoxon test ($S = 0$).
- According to the Wilcoxon test duration features d_{a1} and d_l are equally significant ($S = 36$), while d_{bh} is of lesser significance ($S = 35$).
- By considering the $\frac{\overline{\mu}}{\sigma}$ in conjunction with the Wilcoxon test, it is established that d_l is more significant than d_{a1} due to the higher value of $\left|\frac{\overline{\mu}}{\sigma}\right|$ for d_l.
- All phonemes except /a2/ are longer in duration for the A form than those of the C form because the $\frac{\overline{\mu}}{\sigma}$ values are positive in these cases.

All these observations regarding the phoneme duration difference between the A and the C forms may be compactly represented by the following notation to be discussed in detail later:

(4.30) $\mathbf{bh}_{3+}$ bha1 $\mathbf{a1}_{2+}$ $\mathbf{l}_{1+}$ a2

The significance of a feature in a specific comparison has particular relevance for the synthesis of stimuli. The following chapter will show how the prosodic features of imperatives in one context are changed so that they acquire the prosodic features of those imperatives in another context. The stimuli are generated by applying calculated percentage changes to the phonemes of a word. Phonemes that exhibited greater significance in the statistic analysis of the acoustic results take priority when it comes to deciding which phonemes should be changed and which not.

In the acoustic analysis the difference in duration, pitch and loudness for phonemes in different contexts were measured. The greater the significance of the difference is for a particular phoneme, the more important that difference is and the more likely it is that a percentage change will be applied to that phoneme when its context is changed in the process of stimuli generation.

4.5 Results of the Acoustic Analysis

This section includes a summary of the duration, pitch and loudness results of the acoustic analysis. The results given below are summarized in tables within this chapter or in Appendices C, D and E.

For duration, detailed results are summarized in the form of tables that can be found in Appendix C. Table C.1 presents the mean duration (in ms) of each durational segment measured for each word. In Table C.2 the mean consonant duration (in ms) for each word analyzed can be found. Table C.3 presents the mean duration (in ms) for /ala/ or /alani/ for all words analyzed and Table C.4 presents the percentage changes in duration from one verb context to another. Table 4.13 in this chapter presents an even shorter summary of the duration results, showing the duration patterns that were observed.

Similarly, detailed results are summarized for pitch in Appendix D. Table D.1 presents the median pitch values for initial consonants in the infinitive context. Table D.2 gives the median pitch values of /ala/ or /alani/ for each word. Table D.3 presents the mean pitch values for the words in their different tonal groups. A summary of the percentage pitch changes between verb forms can be found in Table D.4.

Tables C.4 and D.4 show to what degree a certain phoneme of the verb in a specific context differs in duration and pitch respectively, from that same phoneme in another context (see 4.27 for calculation of difference). The first column shows the contexts of the verbs being compared to each other. For example, A→B means the verb in the infinitive context (A) is compared to the imperative in the singular (-object/adverb) context (B).
The second column in the table shows the order of significance of the differences between the phonemes that were compared to each other. When a phoneme appears in bold print, it means that the difference regarding that particular phoneme was found to be significant.

The numeric subscript indicates the order of significance where a one denotes a strong significance and a two (or three) lesser significance. In addition to the order of significance, positive and negative signs are used to identify the change as an increase or a decrease. When a phoneme is marked with a '+' it can be interpreted as '*the verb in the first context* is longer/has a higher pitch value than *the verb in the second context* for this phoneme'. Conversely, when a phoneme is marked with a '-' it can be interpreted as '*the verb in the first context* is shorter/has a lower pitch value than *the verb in the second context* for this phoneme'.

The following example serves to demonstrate how the tables are interpreted:

(4.31) Duration results: **A→B:** $\boldsymbol{a1}_{1+}$ $\boldsymbol{l}_{2-}$ *a2*

Duration results are dealt with in this example. It shows that the *verb in the infinitive* context (A) was compared to *the verb in the singular imperative without object/adverb* context (B). In this case the most significant difference for duration in the comparison was found to be on the first /a/ of the verb. The number one indicates greatest significance and the positive sign indicates that the /a1/ for A was longer in duration (the difference was positive) than that of the /a1/ for B. The second most significant difference occurred on the phoneme /l/ as indicated by the number two. The negative sign indicates that /l/ for the verb A was shorter in duration (the difference was negative) than that of the /l/ for the verb B. The phoneme /a2/ does not appear in bold print, which means that the duration difference for this phoneme in this comparison was not found to be significant. The tables summarizing the results for pitch can be interpreted in the same way.

Table E.1 in Appendix E gives the combined results for duration, pitch and loudness with results for each verb given separately. When a D_+, D_-, P_+ or P_- is present, it means that the duration and/or pitch of the /ala/ part of the word showed a significant difference when two contexts were compared. Similarly, TD and TP represents the Total Duration and Total Pitch respectively of the word, excluding the first consonant and including the /ni/ (i.e. d_{alani} or p_{alani}).

4.5.1 Duration

A→B:

The following results are only applicable to the /ala/ part of the infinitive (A) and the imperative in singular (-object/adverb) context (B).

- The duration of the /ala/ part of the verb in the infinitive context was found to be significantly **shorter** compared to the imperative in the singular (-object/adverb) context of the verb for all verb stems (D.). An exception was observed for the verb *bhala*, where no significant change in duration was found.
- No significant change in the duration of a specific phoneme occurred when the infinitive form of the verb was compared with the imperative in singular (-object/adverb) context.

For future stimuli generation the implications are as follows: to change the neutral infinitive form of the verb to an imperative in the singular (-object/adverb) context, the total duration of the infinitive form should be lengthened.

A→C, A→D, A→E:

The following results are only applicable to the /ala/ part of the infinitive (A) and the imperative in the particular contexts singular (+object/adverb) (C), plural (-object/adverb) (D) and plural (+object/adverb) (E).

- The duration of the /ala/ part of the infinitive is **longer** than that of the imperative for **all** these cases (D_+).
- Of the phonemes that differed significantly in duration, 75.7% were on the /al/ part. In all these instances either the /a/ or the /l/ (or both) was longer for the infinitive than for the other contexts (C, D, E). The other 24.3% of the phonemes were shorter for the infinitive and the shortening only occurred on the second /a/ of the plural contexts.
- The duration difference of the second /a/ of the singular (+object/adverb) context (C) was never found to be significant.
- When the infinitive was compared to the plural forms (D) and (E), the second /a/ and in particular the /la/ syllable as a whole, was always relatively shorter for the infinitive than for the plural forms. With /la/ being the penultimate syllable of the

plural forms, this observation agrees with the phonological rule in Xhosa that states that the penultimate syllable is relatively longer than the other syllables in a word.

- In 88.9% of the cases the duration of the first /a/ was the **most** significant feature (**a1**$_1$), while the /l/ (**l**$_1$) was most significant in 11.1% of the cases.
- The order of significance for the singular (+object/adverb) context is '**a1**$_{1+}$ **l**$_{2+}$ a2'. For the plural contexts the order is '**a1**$_{1+}$ **l**$_{3+}$ **a2**$_{2-}$'.

B→C:

The following results are only applicable to the /ala/ part of the singular (-object/adverb) context (B) and the singular (+object/adverb) context (C).

- The duration of the /ala/ part is longer for all singular imperative (-object/adverb) contexts than for the singular imperative (+object/adverb) contexts (D_+). In other words the tempo of /ala/ becomes quicker when the verb is followed by and object or an adverb.
- The most significant difference in duration is on the first /a/ phoneme, followed in order of significance by the /l/ phoneme (**a1**$_{1+}$ **l**$_{2+}$ a2).

For future stimuli generation the implications are the following: to change the imperative from singular (-object/adverb) to singular (+object/adverb), the most important change should occur on the first /a/ which should be shortened. Secondly the /l/, can be shortened, but /a2/ does not have to be changed (**a1**$_{1+}$ **l**$_{2+}$ a2).

D→E:

The following results are only applicable to the /alani/ part of the plural (-object/adverb) context (D) and the plural (+object/adverb) context (E).

- The duration of the /alani/ part is longer for all plural imperative (-object/adverb) contexts than for the plural imperative (+object/adverb) contexts (TD_+).
- The penultimate syllable of D was found to be relatively longer than that of the E form.
- The order of significance is 'a1 l **a2**$_{1+}$ **n**$_{2+}$ i'.

From these observations one can conclude that, if a word is said in isolation, its duration is longer than when the word is followed by another word (+object/adverb). Again it seems that speakers raised the tempo for the plural (+object/adverb) context.

B→D:

The following results are only applicable to the /ala/ part of the singular (-object/adverb) context (B) and the plural (-object/adverb) context (D).

- The duration of the /ala/ part is longer for all singular (-object/adverb) contexts than for the plural (-object/adverb) contexts (D_+).
- The first /a/ of the singular is longer (**a1**$_{1+}$) than that of the plural context, while the second /a/ is shorter (**a2**$_{2-}$). In the plural context, the second /a/ is part of the penultimate syllable and is consequently longer, as expected.
- Order of significance: '**a1**$_{1+}$ l **a2**$_{2-}$'.

C→E:

The following results are only applicable to the /ala/ part of the singular (+object/adverb) context (C) and the plural (+object/adverb) context (E).

- The duration of the /ala/ part is longer for all singular (+object/adverb) contexts than for the plural (+object/adverb) contexts (D_+).
- No distinct order of significance pattern in duration differences between these two contexts was found, except that the first /a/ of the singular context is longer than that of the plural context (**a1**$_+$), while the /l/ is shorter (**l**$_-$).

The general rule regarding duration in Xhosa states that the penultimate syllable of the word, phrase or sentence is lengthened. Since two phrases (i.e. a verb plus an object/adverb) were compared to each other in the C→E comparison, neither /a1/ or /la/ was in the penultimate syllable position. This may account for the fact that no distinct pattern of significant duration differences between these two contexts was found.

The results summarized above are represented in Table 4.13.

Table 4.13 Duration patterns observed.

	Order of significance	d_{ala}	d_{alani}
A→B	a1 l a2	D_-	
A→C	**a1**$_{1+}$ **l**$_{2+}$ a2	D_+	
A→D	**a1**$_{1+}$ **l**$_{3+}$ **a2**$_{2-}$	D_+	
A→E	**a1**$_{1+}$ **l**$_{3+}$ **a2**$_{2-}$	D_+	
B→C	**a1**$_{1+}$ **l**$_{2+}$ a2	D_+	
D→E	a1 l **a2**$_{1+}$ **n**$_{2+}$ i	D_+	TD_+
B→D	**a1**$_{1+}$ l **a2**$_{2-}$	D_+	
C→E	**a1**$_+$ l a2 *	D_+	

* No consistent order of significance pattern was observed.

4.5.2 Pitch

This section includes a summary of the pitch results of the acoustic analysis.

A→B, A→C, A→D, A→E:

The following results are only applicable to the /ala/ part of the infinitive (A) and the particular contexts B, C, D, and E.

- The average pitch of the /ala/ part of the infinitive is lower than that of the imperative for 75% of the cases (P_-).

From this observation one may conclude that the register of the /ala/ part raises when the verb changes from the infinitive context to the imperative.

- When the infinitive was compared to the singular form of the imperative (A→B, A→C) for FL and HL, the pitch difference of the /l/ of /ala/ was always found to be more significant than the second /a/ (a1 **l**$_1$ **a2**$_2$). The only exceptions occurred in the A→B comparison of the FL group, where the first /a/ was found to show the most significant difference (**a1**$_1$ **l**$_2$ a2).

This observation may have implications for how tone marking is conducted in the future, especially regarding voiced consonants. To date, vowels are the only phonemes considered to carry tone and therefore only vowels are tone marked. This comparison between A and B does however show significant pitch movement on the voiced consonant /l/ giving us the indication that this consonant also carries tone.

- When the infinitive was compared to the plural form of the imperative (A→D, A→E), the second /a/ of /ala/ always showed the most significant difference, followed by the /l/ (a1 $\mathbf{l}_2$ $\mathbf{a2}_1$).

For all the tone groups the second /a/ was originally tone marked as L (e.g. FL: /Câlà/, HL: /Cálà/, LL: /Càlà/). The significant pitch differences for the /a2/ phoneme between the A and D/E contexts act as confirmation of the tonological rule stating that the verb in the infinitive context, tone marked as FL, HL or LL, becomes LHL in the plural imperative context. The second /a/ therefore changes from L to H.

- For verbs in the LL group, the comparison between the infinitive context and the imperative singular (A→B, A→C) never showed a significant difference in average pitch. However, significant differences were found on the /ala/ part, where the second /a/ always showed the most significant difference, followed by the /l/ (a1 $\mathbf{l}_2$ $\mathbf{a2}_1$).

In the case of singular verbs in the LL group, the lack of significant pitch difference between the two /ala/ parts of the verbs acts as confirmation of the tonological rule stating that the verb in infinitive context, tone marked as LL, stays LL in the imperative singular context. In other words, no change in tone marking occurred, therefore no significant pitch difference was expected.

- When the infinitive form of the verb, *lala* was compared to the singular forms (B) and (C), the second syllable /la/ changed significantly, becoming higher for the singular imperative forms (a1 $\mathbf{l}_1$ $\mathbf{a2}_2$). The imperative, *lala* is classified as HL and is tone marked as HL in the infinitive form.
- For both plural forms of the verb, *lala*, the second /a/ changed most significantly ($\mathbf{a2}_1$) becoming higher for the D and E forms.

Again this significant pitch difference is a confirmation of the tonal rule stating that the verb in the infinitive context, tone marked as HL, becomes LHL in the plural imperative context, where /a2/ changes from L to H.

- For the verb, *hlala* the most significant change occurred on the /l/ although the /a1/ also changed significantly where A was compared to B and C ($\mathbf{a1}_3$ $\mathbf{l}_1$ $\mathbf{a2}_2$).

The verb, *hlala* is tone marked as LL in the infinitive context and as HL in the singular imperative context, therefore the increase in pitch for /a1/ and /l/ is expected in the tonal change from L to H.

B→C:

- The difference between the /ala/ for the imperative singular (-object/adverb) context (B) and the /ala/ for the imperative singular (+object/adverb) context (C) was only found to be significant for FL ($a1_3$+ l_1- $a2_2$-).
- For HL and LL no significant pitch differences were found.

For this comparison the only difference is that the B context has no object or adverb, but the C context does have an object or adverb. One would therefore not expect a major change in pitch.

D→E:

- No consistent pattern of significant differences was found between the imperative plural (-object/adverb) (D) and imperative plural (+object/adverb) (E) contexts for the /alani/ part of the word.

B→D:

- The overall pitch for the /ala/ part of the imperative plural (-object/adverb) context (D) was found to be significantly higher than that of the /ala/ part of the imperative singular (-object/adverb) context (B) (P.), but no consistent pattern was found as to where in the word the change occurred.

C→E:

- Like the B→D comparison, the overall pitch for the /ala/ part of the imperative plural (+object/adverb) (E) context was found to be significantly higher than that of the /ala/ part of the imperative singular (+object/adverb) context (C) (P.). A consistent pattern was only found for the LL group, where the most significant change occurred on the second /a/ of /ala/ (a1 $\mathbf{l_2}$- $\mathbf{a2_1}$-).

For changing the neutral infinitive form of the verb to the imperative in different contexts, a consistent pattern was observed within the different tone groups, both in the phoneme order of significance and overall pitch. It was found that the pitch raised significantly on the second syllable /la/ from the A form to the imperative in B, C, D, and E contexts.
For the comparisons B→C, D→E, B→D and C→E the patterns of phoneme changes observed, were not consistent enough to derive pitch rules. However, overall pitch increased for most words in most comparisons ($P_.$), except for the comparison between the imperative in singular (-object/adverb) context and the imperative in singular (+object/adverb) context (B→C).

The results listed above are summarized in Table 4.14.

Table 4.14 Pitch patterns observed.

	FL		**HL**		**LL**	
	Order of significance	p_{ala}	**Order of significance**	p_{ala}	**Order of significance**	p_{ala}
A→B	**a1$_1$- l$_2$-** a2	**$P_.$**	a1 **l$_1$- a2$_2$-**	**$P_.$**	a1 **l$_2$- a2$_1$-**	*
A→C	a1 **l$_1$- a2$_2$-**	**$P_.$**	a1 **l$_1$- a2$_2$-**	*	a1 **l$_2$- a2$_1$-**	*
A→D	a1 **l$_2$- a2$_1$-**	**$P_.$**	a1 **l$_2$- a2$_1$-**	**$P_.$**	a1 **l$_2$- a2$_1$-**	*
A→E	a1 **l$_2$- a2$_1$-**	**$P_.$**	a1 **l$_2$- a2$_1$-**	**$P_.$**	a1 **l$_2$- a2$_1$-**	*
B→C	a1 **l$_1$- a2$_2$-**	*	a1 l a2	**$P_.$**	a1 l a2	*
D→E	a1 l a2 n i	*	a1 l a2 n i	*	a1 l a2	**$P_.$**
B→D	a1 l a2	**$P_.$**	**a1$_{1+}$ l$_{2+}$** a2	**$P_.$**	a1 l a2	*
C→E	a1 l a2	**$P_.$**	**a1$_{1+}$** l a2	*	a1 **l$_2$- a2$_1$-**	**$P_.$**

* Average pitch on /ala/ (p_{ala}) was not statistically significant.

4.5.3 Loudness

As mentioned earlier, the absolute loudness values were not used in the analysis since the input levels for loudness were varied during the recording process. For this reason it was decided to statistically analyze the relative loudness results (cf. Section 4.4.1.3).

Table E.1 in Appendix E gives a summary of the combined results for duration, pitch and loudness. As can be observed in the 'Loudness' column of this table, the results regarding relative loudness did not display any consistent patterns of significance. The conclusion was therefore drawn that loudness, as calculated in this study, does not play a significant

role in the production of imperatives. As shall be noted in the following chapter, the prosodic feature loudness did also not prove to be a prerequisite from a perceptual point of view.

4.6 Chapter Summary

In this chapter an overview was given of the prosodic features, grammatical structures and tone marking for imperatives in Xhosa. It was noted that not much information regarding prosodics exist and that the tonal rules given are not substantiated by any phonetic explanations.

An acoustic analysis was conducted on a small corpus of verbs in different contexts, in order to gather more information on the role of prosody in the production of imperatives. The corpus was digitally recorded and annotated, using speech analysis software and applying a consistent tagging method. Pitch, duration and loudness were measured for the corpus. In order to extract the relevant information from the large amount of data that was produced, the data was statistically analyzed.

Duration results that may be highlighted are as follows:

- To change the neutral infinitive form of the verb to an imperative in the singular (-object/adverb) context, the total duration of the infinitive form should be lengthened.
- To change the imperative from (-object/adverb) to (+object/adverb), the most important change should occur on the first /a/ which should be shortened. Secondly the /l/, can be shortened, but /a2/ does not have to be changed ($\mathbf{a1}_{1+}$, $\mathbf{l}_{2+}$).
- Speakers seemed to raise the tempo for the imperative (+object/adverb) context.
- Evidence of penultimate syllable lengthening was found.

Pitch results that may be highlighted are as follows:

- Significant pitch movement occurred on the voiced consonant /l/ where the verb in the A context was compared to the verb in the B context. This may possibly have implications as to how tone marking is conducted in the future.

- The following tonological rules were confirmed by the observation of significant pitch differences for vowels, when compared in different contexts:

Verbs in the infinitive context, tone marked as FL, HL or LL, become LHL in the plural imperative context;

Verbs in the infinitive context, tone marked as LL stays LL in the imperative singular context;

The verb in infinitive context, *-hlala,* that is tone marked as LL becomes HL in the singular imperative context.

- For changing the neutral infinitive form of the verb to the imperative in different contexts, a consistent pattern was observed within the different tone groups, both in the phoneme order of significance and overall pitch. It was found that the pitch raised significantly on the second syllable /la/ from the A form to imperative in B, C, D, and E contexts.
- For the comparisons B→C, D→E, B→D and C→E the patterns of phoneme changes observed were not consistent enough to derive pitch rules.

The relative loudness results were statistically analyzed, but no consistent patterns of significance were observed for this feature.

The results obtained in this chapter will be used to synthesize imperatives in different contexts. These synthesized imperatives should, however, be perceptually tested in order to determine the relationship between the production and perception of the prosodic features.

Chapter Five

Perceptual Experiment

5.1 Chapter Overview

In this chapter the perceptual experiment done on imperatives in Xhosa is discussed. This experiment is primarily based on the acoustic analysis described in the previous chapter. In the following subsections the aims of the experiment are explained as well as the methods used, the preparation of stimuli, the compilation and presentation of the perception tests and the results obtained. Two examples are also given of how the results can be applied in the prosody generation module of a Xhosa TTS system.

5.2 Introduction

This study investigates the prosodics of imperatives in Xhosa, and in particular, the implication such information on prosodics might have for the development of a text-to-speech system for Xhosa. Speech technology is moving towards making synthetic sound as natural as possible and the research done here is aimed to contribute to this process.

The acoustic analysis conducted previously, aimed to obtain acoustic information on the role of the prosodic features pitch, duration and loudness in the production of commands. This information was successfully obtained through the systematic measurement of the parameters in question and the subsequent statistical analysis thereof.

So far the emphasis was on the production of commands, however, we do not as yet know the relationship between the production and perception processes in this respect. It was therefore decided not to rely only on the acoustic data obtained, when giving guidelines for the synthesis of commands, but also to account for the role of perception by mother tongue listeners.

On the subject of perceptual studies pertaining to prosodics and tone in particular, in the African languages, Roux (1995b) points out the shortcomings of certain approaches

followed in the past. He specifically argues against the impressionistic auditory approach, and maintains that this approach cannot aid in the validation and quantification of prosodic data. Roux (1995b:197) does however support the so-called IPO-approach (Institute of Perception Research, Einhoven) with the following remark:

> "*Manipulating parameters in natural speech electronically, and subjecting a number of listeners to series* (sic) *of perception tests in a controlled manner, result in much more credible data sets on the language. There is a vast difference between this type of data reflecting the perceptual judgements of speaker-listeners of the language in a controlled manner, and data merely representing the impressionistic judgement of a 'trained phonetician'.*"

A similar approach is followed in this perceptual experiment, where stimuli are presented on multimedia computers to a group of subjects within a controlled environment. It is believed that the results obtained from this experiment will provide us with a better understanding of how commands in Xhosa are perceived and at the same time aid in determining whether the measurements done in the acoustic analysis can be applied successfully for synthesis. The opportunity is also taken to test the method of synthesis applied.

5.3 Aims

The aim of the experiment was to devise, through perceptual testing, guidelines on the generation of natural prosodics for imperatives in a Xhosa TTS system. Another aim was to evaluate the method of prosodic parameter extraction discussed in Chapter 4. The possibility of generalizing the prosodic parameters applied for this set of stimuli to similarly structured imperatives, was also investigated. The experiment was divided into three tests, the specific aims of each test being as follows:

Perception Test 1:

1. To determine whether the results of the acoustic analysis conducted previously could be applied to generate plural and singular commands with perceptually acceptable prosodics.

2. To determine whether a combination of both pitch and duration information plays a role in the perception of imperatives or whether the one prosodic element in fact plays a more important role than the other.
3. To determine whether subjects can distinguish between the prosodics of the infinitive (i.e. neutral form) and that of a command.

Perception Test 2:

1. To determine whether commands with objects/adverbs could be generated based on the results of the acoustic analysis.
2. To determine whether the tempo and pitch differences between the imperative-with-object/adverb forms and without object/adverb as revealed in the acoustic analysis are also perceptually relevant.

Perception Test 3:

1. To determine whether some deviation from the mean duration and pitch synthesis parameters are perceptually acceptable.

5.4 Method

Stimuli for the perceptual experiment were generated by applying the mean duration and pitch values for eight speakers as derived from the statistic analysis of the acoustic data[13]. The stimuli represented different versions of commands, as different combinations of these prosodic values were applied. These new synthesized utterances were then tested perceptually.

The method applied for all three perception tests is that of the *preference test on pairs of stimuli* (Salza et al., 1996; Pols, 1991:404). The different versions of commands were presented to the subjects in pairs and subjects were asked to choose the utterance that sounded the most like a command. Subjects could only choose 'Utterance 1' or 'Utterance 2'. Provision was not made for the subjects to answer 'Unsure' or 'Utterance 1 equivalent

[13]Since the acoustic analysis showed that loudness does not play a significant role in the production of imperatives, no particular application was made for this feature during the stimuli generation process.

to Utterance 2'. However, the answer was automatically recorded as 'Undecided' if the subjects did not choose either one of the two options within 4 seconds. This way the subjects were forced to choose one preferred utterance. According to Salza et al. (1996:653) the subjects "*are likely to pay more attention to stimuli in the forced choice condition and comparison results could become more reliable*". Through these preferences one could infer which combination of prosodic values the subjects considered to be the most acceptable for a specific command.

The results of this experiment were then statistically analyzed so that it could be determined which range of prosodic values could be applied to synthesize perceptually acceptable commands.

5.5 Preparation of Stimuli

5.5.1 Method of Stimuli Generation

In order to generate commands using predefined duration and pitch parameters, some method of synthesis had to be decided upon. The Kay Elemetrics suite of speech analysis tools was available in the Research Unit for Experimental Phonology (RUEPUS). This system was used for the recording and analysis of speech data as discussed in Chapter 4. The package includes a program called Analysis/Synthesis Lab[14] (ASL Version 3.01) that allows the user to manipulate the duration, pitch and formants of a speech signal using linear predictive coding (LPC) based analysis and synthesis methods. In our application the ASL had some disadvantages, as discussed below.

In an informal experiment the PSOLA technique was compared to the LPC based pitch and duration modification provided by the ASL and it was found that the PSOLA technique resulted in higher quality synthetic speech. PSOLA type methods are widely used in the speech synthesis community (Dutoit, 1997a).

While the CSL and ASL packages are very powerful tools, it is sometimes necessary to augment these packages with custom written programs to optimally achieve a certain goal.

[14] Analysis/Synthesis Lab is a registered trademark of Kay Elemetrics Corp.

This was indeed the route taken. A PSOLA system was developed in the Research Unit[15] for this application. A major benefit was that the implemented programs required minimal manual intervention, making the approach suitable for automatic manipulation by computer. Furthermore, our approach required only the specification of one pitch value per phoneme, whereas ASL required the user to manually draw the pitch contour with the mouse.

Our implementation of PSOLA pitch and duration manipulation proved to cause distortions in quality when natural speech was used to synthesize (either by concatenating syllables or by synthesizing the word as a whole). This would in turn cause unwanted interference when presenting the stimuli to naïve subjects. Consequently it was decided to have monotonous verbs recorded by one speaker, since it is easier to impose a desired prosodic pattern on a neutral base than on completely natural speech. This approach agrees with the recommendations of Dutoit (1997b) and Dutoit and Leich (1993:438) stating that the best results are achieved when the corpus is read with the most monotonic intonation possible. It is also recommended that "*even the end of words should maintain their fundamental frequency constant*", however unnatural it may sound (Dutoit, 1997b).

Since the quality of synthesis was found to be better if the pitch was raised in stead of lowered and the duration was shortened in stead of lengthened, the speaker was instructed to speak in low tones and to produce sounds with long durations. Therefore, in effect, the speaker produced speech with over-all monotonous prosody (little pitch variation, long sounds, unstressed). Even though it is not humanly possible to sustain a perfectly monotonous pitch level and while it is true that this monotonous speaking style inadvertently influences, amongst others, the formants of the speech signal, these adverse effects were not found to degrade the intelligibility significantly.

In a trial run before the actual experiment, it was also found that, even after manipulation, the signals retained enough speaker specific qualities for the subjects to recognize the 'voice' as that of the original speaker.

---

[15]Developed by Mr. J.A.N. Louw.

5.5.2 Corpus Recorded

The same verbs used in the acoustic analysis were recorded, as well as three extra verbs that had not been acoustically analyzed previously. The speaker was also asked to utter natural command sentences. This was done in order to have a small database of naturally produced objects and adjectives to concatenate to the synthesized verbs.

Each monotonous verb as well as the naturally pronounced objects and adverbs were annotated using the same method as for the acoustic analysis. Pitch and duration was calculated for each word. The corpus of words recorded is listed in Table 5.1.

Table 5.1 Corpus of words recorded.

Pronounced Monotonously					
Included in acoustic analysis				**Not included in acoustic analysis**	
Bhala	'Write'	*Dlala*	'Play'	*Qala*	'Begin'
Bhalani	'Write ye'	*Dlalani*	'Play ye'	*Qalani*	'Begin ye'
Lala	'Sleep'	*Hlala*	'Sit'	*Sala*	'Sit'
Lalani	'Sleep ye'	*Hlalani*	'Sit ye'	*Salani*	'Sit ye'
Bala	'Count'	*Tsala*	'Pull'	*Vala*	'Close'
Balani	'Count ye'	*Tsalani*	'Pull ye'	*Valani*	'Close ye'
Pronounced Naturally					
Bhala unobumba!	'Write a letter!'			*Tsala apha!*	'Pull here!'
Lala apha!	'Sleep here!'			*Qala ngoku!*	'Begin now!'
Bala izinto!	'Count the things!'			*Sala apha!*	'Sit down!'
Dlala apha!	'Play here!'			*Vala ucango!*	'Close the door!'
Hlala phantsi!	'Sit down!'				

5.5.3 Method of Manipulation

As mentioned above, a PSOLA technique was used to modify the pitch contour and duration of stimuli. Specifically, the *Time-Domain Pitch Synchronous Overlap-Add* (TD-PSOLA) method (Moulines & Charpentier, 1990) was used. This algorithm is relatively easy to implement. The only prerequisite is that the pitch periods be marked. The pitch extraction algorithm explained in Section 4.3.2.4 was employed to this end.

The process to modify the pitch contour of a voiced segment with the TD-PSOLA method consists of the following steps (cf. Figure 5.1):

1. A set of short-term signals (red) is obtained from the original signal (black) by windowing each pitch period with a Hanning window (green). These windows are centred around pitch marks and may include portions of the previous and successive pitch periods.
2. The windowed short-term signals (purple) are time-shifted so that the delays between successive short-term signals correspond to the pitch periods of the desired pitch contour.
3. These overlapping short-term signals are then summed to yield the manipulated speech signal (pink).

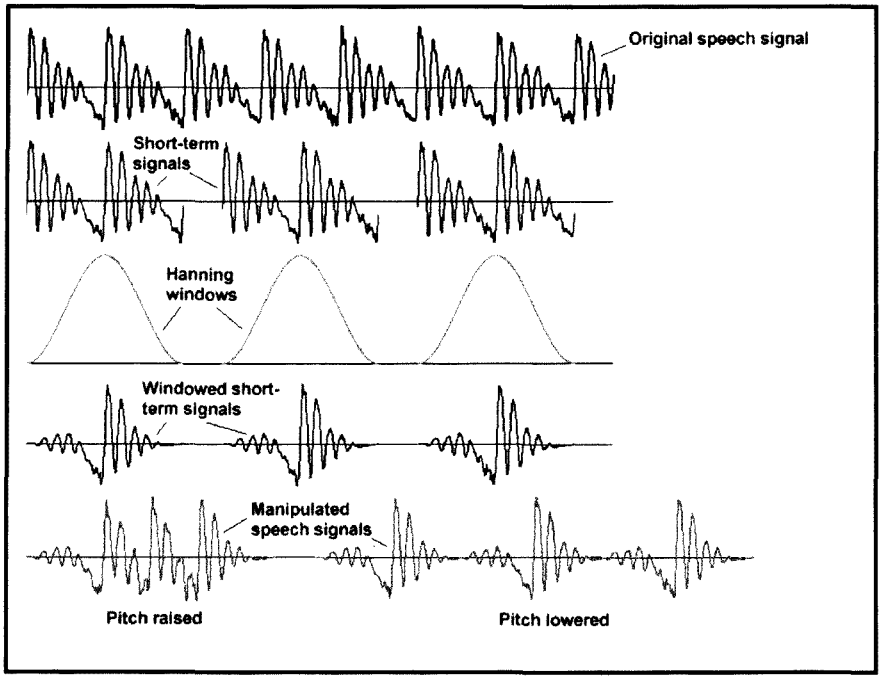


Figure 5.1 The TD-PSOLA process for raising or lowering the pitch of a voiced speech segment.

For each stimulus, the pitch manipulation was done first, followed by a correctional step to obtain the desired duration. The pitch and the duration values could be varied independently of one another. A limitation was that the range of possible manipulation depended on the value of the parameter in the original signal. The pitch manipulation program allowed for the insertion or removal of every second pitch period at most. Therefore, manipulation was limited to increases of no more than a factor two or decreases of no less than half. Raising the pitch automatically causes a shortening in duration and

this sometimes made it impossible to adhere to the statistically determined target duration values.

TD-PSOLA Example

The example below illustrates how duration and pitch values would be specified. The original and the desired values are shown in Tables 5.2 and 5.3 below. The program was not able to simultaneously reach the duration target of 270 ms and the pitch target of 168 Hz for the /aI/ phoneme. Instead the duration and pitch values that could be specified without error were limited to 260 ms and 144 Hz respectively. From the tables it is clear though, that the program was able to modify the original utterance within close range of the specified target values.

Table 5.2 Example of the duration manipulation with TD-PSOLA.

	d_{dI}	d_{aI}	d_{I}	d_{a2}
Original duration	183.3 ms	325.6 ms	77.8 ms	238.3 ms
Desired duration		270 ms	84 ms	140 ms
Specified duration		260 ms	84 ms	140 ms
Actual duration after manipulation	183.3 ms	258.3 ms	87.3 ms	135.3 ms

Table 5.3 Example of pitch manipulation with TD-PSOLA.

	p_{aI}	p_{I}	p_{a2}
Original pitch	100 Hz	101.8 Hz	103.6 Hz
Desired pitch	168 Hz	125 Hz	108 Hz
Specified pitch	144 Hz	125 Hz	108 Hz
Actual pitch after manipulation	142.9 Hz	123.5 Hz	107.5 Hz

Figure 5.2 below illustrates several aspects of the manipulation of duration and pitch. The mean pitch values of the phonemes of the original monotonously produced utterance (red circles), as well as the actual pitch contour (red line) are shown. From the specified means for each phoneme (green circles) the program calculated a smooth desired pitch contour (green line). After the manipulation has been done, the mean pitch values of the phonemes (purple circles) were close to the target values. Also notice that the modified pitch contour (purple line) resembles the shape of the desired pitch contour, but it is shorter in duration. The reason for this is that the desired pitch contour is specified in terms of the duration of the original signal. Since the target pitch is higher than the original, the raising of the pitch inherently causes a shortening of the utterance. Therefore, the final manipulated utterance

will be shorter than the original. After the pitch manipulation was done, the duration of the phonemes was adjusted towards the target duration values. In some cases these target values were less than half or more than twice the duration of the phoneme at that stage (i.e. after pitch manipulation). With our implementation of the TD-PSOLA algorithm, in these instances, it was not possible to reach these target values.

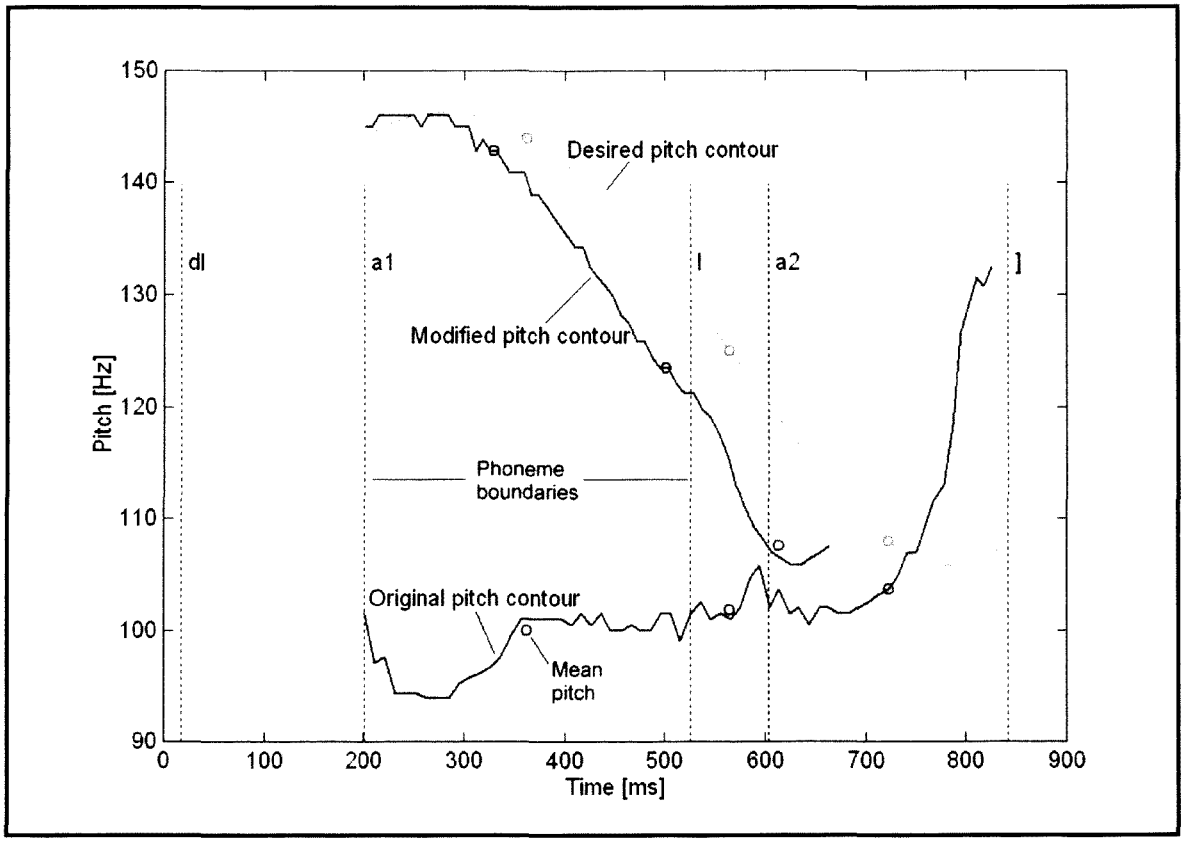


Figure 5.2 Pitch contours of original utterance and after manipulation with TD-PSOLA.

The disadvantage of using only one set of monotonously produced words was that some target values could not be reached when large changes in pitch and duration were required. A practical solution to this problem could be to record the data in different pitch sets, in other words, to record more than one monotonous version of a word, each pronounced on a different register (pitch level). The monotonous word with the average pitch that is the closest to the desired pitch values could then be used as the basis of the synthesis process.

5.5.4 Calculation of Prosodic Parameters of Stimuli

A total of fifteen command versions were synthesized through various degrees of manipulation. These fifteen versions were generated for each of the nine monotonously pronounced verbs, giving a total of 135 stimuli to be presented in the three different perception tests. Each of the command versions were generated by applying the duration and pitch values as derived from the acoustic data. Since no significant patterns were observed in the acoustic analysis, from which to extract loudness parameters, no manipulation of loudness was done.

1. Five basic command versions, namely version $\tilde{A}$, $\tilde{B}$, $\tilde{C}$, $\tilde{D}$ and $\tilde{E}$, were synthesized. The names allocated to these synthesized command versions are written with a tilde (e.g. $\tilde{A}$) so as not to be confused with the originally recorded and acoustically analyzed commands A,B,C,D and E.

   Since the mean duration and pitch values taken from the statistical analysis were applied, each of these five versions was considered to represent the verb with the so-called 'best' prosodics for the specific context. In other words, where the prosodics are concerned, these synthesized versions would represent the best approximation of the naturally produced commands. For the sake of clarification these versions were called 'best' versions.

   - For version $\tilde{A}$ the mean pitch and duration values of the infinitive form of the verb were applied. This version would then represent the singular imperative with the prosodic values of the infinitive or rather a so-called 'neutral' command. In other words, the prefix /uku-/ is dropped but the pitch and duration of the verb stem is that of the infinitive (e.g. *Bhala*).
   - For version $\tilde{B}$ the mean pitch and duration values of the singular imperative without object/adverb were applied. This version would then represent the verb with the best prosodics for the singular (-object/adverb) context (e.g. *Bhala!*).
   - For version $\tilde{C}$ the mean pitch and duration values of the singular imperative with object/adverb were applied; with these values representing the best prosodics for the imperative in this particular context (e.g. *Bhala unobumba!*).

- For version $\tilde{D}$ the mean pitch and duration values of the plural imperative without object/adverb were applied; with these values representing the best prosodics for the imperative in the plural without object/adverb context (e.g. *Bhalani!*).
- For version $\tilde{E}$ the mean pitch and duration values of the plural imperative with object/adverb were applied. This version would then represent the verb with the best prosodics in the plural context with object/adverb (e.g. *Bhalani unobumba!*).

The calculation of the prosodic parameters will be explained next.

Duration

The duration value used in the perception tests for a phoneme x was the mean over all tonal groups of the speakers' mean. For example, the duration of the first /a/ of the word *bhala*, as used for synthesis, was calculated as follows:

(5.1)

$$d_{\text{a1 bhala}} = \frac{1}{8}\left(d_{\text{a1 bhala A}} + d_{\text{a1 bhala J}} + d_{\text{a1 bhala L}} + d_{\text{a1 bhala M}} + d_{\text{a1 bhala P}} + d_{\text{a1 bhala T}} + d_{\text{a1 bhala X}} + d_{\text{a1 bhala Z}}\right)$$

$$d_{\text{a1}} = \frac{1}{6}\left(d_{\text{a1 bhala}} + d_{\text{a1 dlala}} + d_{\text{a1 lala}} + d_{\text{a1 hlala}} + d_{\text{a1 bala}} + d_{\text{a1 tsala}}\right)$$

The duration values used for the five basic command versions are given in Table **5.4**.

Table 5.4 Duration values used in the perception tests.

	a1 [1]	**l**	**a2**	**n**	**i**
$\tilde{A}$	256 ms	86 ms	102 ms		
$\tilde{B}$	260 (270) ms [2]	84 ms	140 ms		
$\tilde{C}$	149 ms	64 ms	114 ms		
$\tilde{D}$	124 ms	79 ms	134 ms [3]	100 ms	110 ms
$\tilde{E}$	113 ms	73 ms	142 ms	72 ms	81 ms

[1] The duration of /a1/ was $d_{Ca1}+d_{a1}$.

[2] The desired duration is shown in parentheses in cases where the target value could not be reached.

[3] The mean obtained from the acoustic analysis was 259 ms, which was impossible to achieve. In this case the percentage change was applied: $d_{\tilde{D}:a2}$ is 31% higher than $d_{A:a2}$ $d_{\tilde{D}:a2}$ = 1.31*102 ms = 134 ms.

Pitch

The pitch value used in the perception tests for a phoneme *x* was the mean over a tonal groups of the speakers' median values. For example, the pitch of the first /a/ of the word *bhala* as used for synthesis was calculated as follows:

(5.2)

$$p_{\text{a1 bhala}} = median(p_{\text{a1 bhala A}} + p_{\text{a1 bhala J}} + p_{\text{a1 bhala L}} + p_{\text{a1 bhala M}} + p_{\text{a1 bhala P}} + p_{\text{a1 bhala T}} + p_{\text{a1 bhala X}} + p_{\text{a1 bhala Z}})$$

$$p_{\text{a1 FL}} = \frac{1}{2}(p_{\text{a1 bhala}} + p_{\text{a1 dlala}})$$

$$p_{\text{a1 HL}} = \frac{1}{2}(p_{\text{a1 lala}} + p_{\text{a1 hlala}})$$

$$p_{\text{a1 LL}} = \frac{1}{2}(p_{\text{a1 bala}} + p_{\text{a1 tsala}})$$

A different set of pitch values was used for each tonal group. The pitch values used for the five basic command versions in each tonal group are given in Table 5.5.

Table 5.5 Pitch values used in the perception tests (in Hz).

FL					
	p_{a1}	p_l	p_{a2}	p_n	p_i
$\tilde{A}$	131 Hz	101 Hz	101 Hz		
$\tilde{B}$	144 Hz (168 Hz)	125 Hz	108 Hz		
$\tilde{C}$	151 Hz	156 Hz	148 Hz		
$\tilde{D}$	134 Hz	148 Hz	172 Hz	175 Hz	151 Hz
$\tilde{E}$	131 Hz	143 Hz	174 Hz	193 Hz	181 Hz
HL					
	p_{a1}	p_l	p_{a2}	p_n	p_i
$\tilde{A}$	155 Hz	125 Hz	114 Hz		
$\tilde{B}$	176 Hz	179 Hz	156 Hz		
$\tilde{C}$	171 Hz	183 Hz	172 Hz		
$\tilde{D}$	155 Hz	159 Hz	180 Hz	185 Hz	158 Hz
$\tilde{E}$	149 Hz	158 Hz	178 Hz	186 Hz	173 Hz
LL					
	p_{a1}	p_l	p_{a2}	p_n	p_i
$\tilde{A}$	157 Hz	103 Hz	90 Hz		
$\tilde{B}$	147 Hz	151 Hz	174 Hz		
$\tilde{C}$	160 Hz	158 Hz	160 Hz		
$\tilde{D}$	157 Hz	160 Hz	178 Hz	179 Hz	151 Hz
$\tilde{E}$	162 Hz	172 Hz	190 Hz	196 Hz	186 Hz

2. Two variations based on version B were generated, the one focussing on duration and the other focussing on pitch.

   - Version *BD* was generated using the mean pitch values of the verb in the infinitive mood and the mean duration values of the singular imperative without object/adverb. This word would then have so-called 'neutral' pitch values and so-called 'best' duration values.
   - Version *BP* was generated using mean pitch values of the singular imperative without object/adverb and the mean duration values of the verb in the infinitive mood. This word would then have so-called 'best' pitch values and so-called 'neutral' duration values.

Table 5.6 provides a summary of how the parameters were specified for the $\tilde{A}$, $\tilde{B}$, *BD* and *BP* versions. Tables 5.7 and 5.8 respectively, provide the numeric duration and pitch values applied to generate these stimuli.

Table 5.6 Parameters specified for $\tilde{A}$, $\tilde{B}$, *BD* and *BP*.

Command version	Duration	Pitch
$\tilde{A}$	d_A	p_A
$\tilde{B}$	d_B	p_B
BD	d_B	p_A
BP	d_A	p_B

Table 5.7 Duration values applied to generate $\tilde{A}$, $\tilde{B}$, *BD* and *BP*.

	a1 $_1$	**l**	**a2**
$\tilde{A}$	256 ms	86 ms	102 ms
$\tilde{B}$	260 (270) ms	84 ms	140 ms
BD	260 (270) ms	84 ms	140 ms
BP	256 ms	86 ms	102 ms

Table 5.8 Pitch values applied to generate $\tilde{A}$, $\tilde{B}$, *BD* and *BP*.

	p_{a1}	p_l	p_{a2}
$\tilde{A}$	131 Hz	101 Hz	101 Hz
$\tilde{B}$	144 Hz (168 Hz)	125 Hz	108 Hz
BD	131 Hz	101 Hz	101 Hz
BP	144 Hz (168 Hz)	125 Hz	108 Hz

3. Imperative versions with object/adverb of version B and D were also generated using the prosodic values of the B and D versions and concatenating a suitable object/adverb. These stimuli were created in such a way that the *BT* version should have the syntactic structure of a singular imperative with object/adverb and the *DT* version should have the structure of a plural imperative with object/adverb.

Table 5.9 provides a summary of how the parameters were specified for the BT, $\tilde{C}$, $\tilde{E}$ and *DT* versions.

Table 5.9 Parameters specified for BT, $\tilde{C}$, $\tilde{E}$ and *DT*.

Command version	**Duration**	**Pitch**	**Object/ adverb**
BT	d_B	p_B	C
$\tilde{C}$	d_C	p_C	C
$\tilde{E}$	d_E	p_E	E
DT	d_D	p_D	E

4. Finally, three more variations of version B and D each were generated by reducing or raising the so-called 'best' pitch and duration values simultaneously in steps. The stimuli produced would then represent command versions with a specified *range* of prosodic parameters.

   - Version *BM1* (B-more-1-step) was generated by raising the 'best' percentage changes for pitch and duration with 33.3%.
   - Version *BM2* (B-more-2-steps) was generated by raising the 'best' percentage changes for pitch and duration with 100%.
   - Version *BL1* (B-less-1-step) was generated by lowering the 'best' pitch and duration percentage changes with 33.3%.
   - Lowering these values in a second step proved to distort the quality of synthesis, therefore it was decided not to generate a *BL2* (B-less-2-steps) version.
   - In the same way three more versions of version D were also generated altering the pitch and duration values simultaneously in steps yielding versions *DM1*, *DM2* and *DL1*.

The calculation of the values of the duration and pitch parameters applied to generate command versions within a specified range was done as follows:

Suppose the duration (or pitch) of a phoneme of a particular stimulus is v. The value of v was calculated as:

$$v = v_A + n\frac{v_B - v_A}{3} \tag{5.3}$$

where v_A and v_B are the values of the parameter for the A and B forms respectively as calculated in the acoustic analysis. The value of n was as shown in Table 5.10 below.

Table 5.10 Calculation of steps.

	n
L1	2
M1	4
M2	6 [1]

[1] If $v < 0$ then $n = 5$.

Therefore the percentage changes shown in Table F.1 and Table F.2 in Appendix F was reduced by 33.3% for the L1 step and increased by 33.3% and 100% for the M1 and M2 steps respectively.

Table 5.11 provides a summary of how the parameters were specified for the $\tilde{B}$, *BL1*, *BM1*, *BM2*, $\tilde{D}$, *DL1*, *DM1* and *DM2* versions. Tables 5.12 and 5.13 respectively, provide the numeric duration and pitch values applied to generate these stimuli.

Table 5.11 Parameters specified for $\tilde{B}$, *BL1*, *BM1*, *BM2*, $\tilde{D}$, *DL1*, *DM1* and *DM2*.

Command version	**Duration**	**Pitch**
$\tilde{B}$	d_B	p_B
BL1	$d_A+0.67(d_B-d_A)$	$p_A+0.67(p_B-p_A)$
BM1	$d_A+1.33(d_B-d_A)$	$p_A+1.33(p_B-p_A)$
BM2	$d_A+2(d_B-d_A)$	$p_A+2(p_B-p_A)$
$\tilde{D}$	d_D	p_D
DL1	$d_A+0.67(d_D-d_A)$	$p_A+0.67(p_D-p_A)$
DM1	$d_A+1.33(d_D-d_A)$	$p_A+1.33(p_D-p_A)$
DM2	$d_A+2(d_D-d_A)$	$p_A+2(p_D-p_A)$

Table 5.12 Duration values applied to generate $\tilde{B}$, *BL1*, *BM1*, *BM2*, $\tilde{D}$, *DL1*, *DM1* and *DM2*.

	a1	**l**	**a2**	**n**	**i**
$\tilde{A}$	256 ms	86 ms	102 ms		
BL1	259 ms	85 ms	127 ms		
$\tilde{B}$	260 (270) ms	84 ms	140 ms		
BM1	261 ms	83 ms	153 ms		
BM2	264 ms	82 ms	178 ms		
DL1	168 ms	81 ms	123 ms	100 ms	110 ms
$\tilde{D}$	124 ms	79 ms	134 ms	100 ms	110 ms
DM1	80 ms	77 ms	145 ms	100 ms	110 ms
DM2	36 ms	72 ms	166 ms	100 ms	110 ms

Table 5.13 Pitch values applied to generate $\tilde{B}$, *BL1*, *BM1*, *BM2*, $\tilde{D}$, *DL1*, *DM1* and *DM2*.

FL					
	$\mathbf{p_{a1}}$	$\mathbf{p_l}$	$\mathbf{p_{a2}}$	$\mathbf{p_n}$	$\mathbf{p_i}$
$\tilde{A}$	131 Hz	101 Hz	101 Hz		
BL1	140 Hz	117 Hz	106 Hz		
$\tilde{B}$	144 (168) Hz	125 Hz	108 Hz		
BM1	148 Hz	133 Hz	110 Hz		
BM2	157 Hz	149 Hz	115 Hz		
DL1	133 Hz	132 Hz	148 Hz	175 Hz	151 Hz
$\tilde{D}$	134 Hz	148 Hz	172 Hz	175 Hz	151 Hz
DM1	135 Hz	164 Hz	196 Hz	175 Hz	151 Hz
DM2	137 Hz	195 Hz	243 Hz	175 Hz	151 Hz
HL					
	$\mathbf{p_{a1}}$	$\mathbf{p_l}$	$\mathbf{p_{a2}}$	$\mathbf{p_n}$	$\mathbf{p_i}$
$\tilde{A}$	155 Hz	125 Hz	114 Hz		
BL1	169 Hz	161 Hz	142 Hz		
$\tilde{B}$	176 Hz	179 Hz	156 Hz		
BM1	183 Hz	197 Hz	170 Hz		
BM2	197 Hz	233 Hz	198 Hz		
DL1	155 Hz	148 Hz	158 Hz	185 Hz	158 Hz
$\tilde{D}$	155 Hz	159 Hz	180 Hz	185 Hz	158 Hz
DM1	155 Hz	170 Hz	202 Hz	185 Hz	158 Hz
DM2	155 Hz	193 Hz	246 Hz	185 Hz	158 Hz
LL					
	$\mathbf{p_{a1}}$	$\mathbf{p_l}$	$\mathbf{p_{a2}}$	$\mathbf{p_n}$	$\mathbf{p_i}$
$\tilde{A}$	157 Hz	103 Hz	90		
BL1	150 Hz	135 Hz	146 Hz		
$\tilde{B}$	147 Hz	151 Hz	174 Hz		
BM1	144 Hz	167 Hz	202 Hz		
BM2	137 Hz	199 Hz	258 Hz		
DL1	157 Hz	141 Hz	149 Hz	179 Hz	151 Hz
$\tilde{D}$	157 Hz	160 Hz	178 Hz	179 Hz	151 Hz
DM1	157 Hz	179 Hz	207 Hz	179 Hz	151 Hz
DM2	157 Hz	217 Hz	266 Hz	179 Hz	151 Hz

5.6 Compilation of Perception Tests

The three perception tests were compiled in such a way that the subject would hear two different synthesized versions of a verb. The subject would then have to choose which one of the two verbs sounds the most like a command. Stimuli were selected for the compilation of the three perception tests according to the different aims of each test (cf. Section 5.3).

5.6.1 Perception Test 1

As noted in Section 5.3, the aims of the first perception test were (i) to determine whether the results of the acoustic analysis conducted previously could be applied to generate plural and singular commands with perceptually acceptable prosodics, (ii) to determine the role of pitch and duration in the perception of imperatives and (iii) to determine whether subjects can distinguish between the prosodics of the infinitive and that of a command.

For this perception test, the stimuli were compiled in the following way: each time the different versions of the singular imperative without object/adverb ($\tilde{B}$ versions) were played together with the so-called 'neutral' verbs ($\tilde{A}$ versions).

Three verbs for each of the 3 tonal groups (FL, HL and LL) were presented as summarized below:

$\tilde{A}$ vs. $\tilde{B}$
$\tilde{A}$ vs. BD
$\tilde{A}$ vs. BP

This amounted to a total of 27 stimuli presented to a group of 33 subjects.

5.6.2 Perception Test 2

The aims of the second perception test were (i) to determine whether commands with objects/adverbs could be generated based on the results of the acoustic data and (ii) to determine whether the tempo and pitch differences between the imperative-with-

object/adverb forms and without object/adverb, as revealed in the acoustic analysis, are also perceptually relevant.

For the this perception test, the stimuli were compiled in the following way: each time the singular imperative with object/adverb ($\tilde{C}$ version) was played against the singular imperative with the 'best' prosodic values of a verb *without* object/adverb but with the syntactic structure of an imperative with object or adverb ($\tilde{B}$ version + object/adverb). In the same way, the $\tilde{E}$ version was played against the *DT* version.
Three verbs for each of the 3 tonal groups were presented as summarized below:

$\tilde{C}$ vs. BT
$\tilde{E}$ vs. DT

This amounted to a total of 18 stimuli presented to a group of 35 subjects.

5.6.3 Perception Test 3

The aim of the third perception test was to determine whether some deviation from the mean duraion and pitch synthesis parameters are perceptually acceptable.

For this perception test, the so-called 'best' versions of the singular and plural imperative without object/adverb were played against their counterparts where the prosodic levels were raised or lowered in steps. For each of the 3 tonal groups, 3 stimuli were presented for each verb in the singular context and 3 stimuli for each verb in the plural context, consisting of the following pairs:

$\tilde{B}$ vs. *BL1*
$\tilde{B}$ vs. *BM1*
$\tilde{B}$ vs. *BM2*
$\tilde{D}$ vs. *DL1*
$\tilde{D}$ vs. *DM1*
$\tilde{D}$ vs. *DM2*

This amounted to a total of 54 stimuli that were presented to a group of 33 subjects (except for $\tilde{D}$ vs. $DM2$ that was presented to 17 subjects).

5.7 Presentation

In the perceptual experiment 90 stimuli were presented to 37 subjects. This amounted to 3330 individual responses. Clearly it would not be ideal to transfer these responses from paper to computer readable form by hand for further analysis. Since a system to do this automatically was not readily available, software was developed locally[16] (i) to present the stimuli to subjects and (ii) to collect their responses automatically. Additional advantages to this approach are that the subjects' response times could be measured and that a maximum limit could be imposed on the time available to respond.

A subject interacted with the perception test program using only the mouse. Firstly the subject started the test by clicking on either 'Start' button, as shown in Figure 5.3. The program played two stimuli in succession over headphones while the text on the screen instructed the subject to listen (see Figure 5.4). While the subject was listening, the cursor showed an hourglass and the subject was prohibited from responding using the mouse. The subject was then expected to select the more natural stimulus as being the first or the second utterance by clicking on the appropriate button on the screen as exemplified in Figure 5.5. If the subject did not click on any button within 4 seconds the response of that stimulus was taken as *Undecided* and the test continued. The cycle repeated itself until all stimuli had been presented.

[16] 'PTest' developed by Mr. J.A.N. Louw.

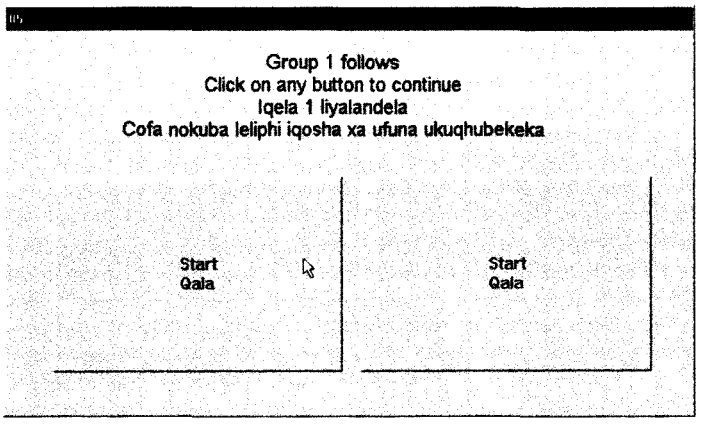


Figure 5.3 Screen layout prompting the subject to start a perception test.

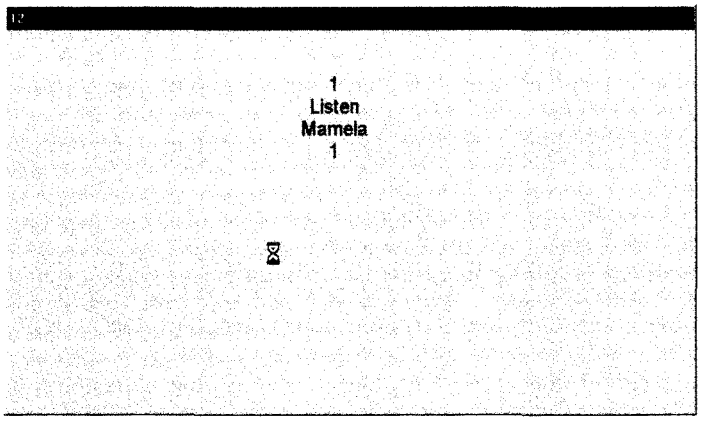


Figure 5.4 Screen layout while a stimulus is being played.

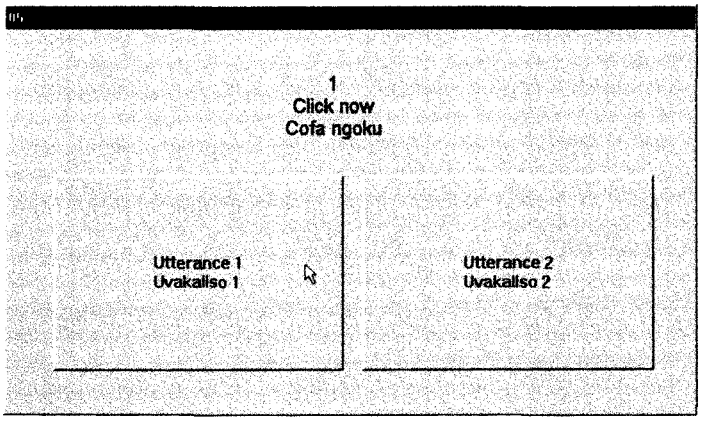


Figure 5.5 Screen layout prompting the subject to click on one of the two buttons.

The perception tests were administered on multimedia computers[17] in a language laboratory accommodating twenty pupils at a time. An introductory lecture on the subject of imperatives was conducted. When it was established that the subjects had a good understanding of the kind of imperatives dealt with in this study, several practice runs were conducted on the computers. The purpose of the practise runs was to familiarize subjects with the manner of presentation of the data and to practise interacting with the computer, using a mouse. The subjects in the perceptual experiments were grade 11 Xhosa mother-tongue speakers.

5.8 Results of Perceptual Experiment

5.8.1 Calculating Results

Graphs such as Figures G.1, G.2, G.3 and G.4 presented in Appendix G, of the percentage First Utterance, Second Utterance and Undecided responses for a given set of stimuli aided in the interpretation of the results of the perception tests. However, the *chi square test* may be used to formally evaluate the *statistical significance* of a majority response of one class.

The *chi square test with one degree of freedom* was applied to every stimulus in isolation to test whether the particular majority was significant.

Chi Square Test

Consider the number of responses obtained for a stimulus numbered *i*. The number of First Utterance, Second Utterance and Undecided responses are $n_{i,F}$, $n_{i,S}$ and $n_{i,U}$ respectively. The *observed frequency* for First Utterance, $f_{i,F}$, is defined as:

$$f_{i,F} = n_{i,F} + {}^{n_{i,U}}\!/_{2} \tag{5.4}$$

Similarly the observed frequency of the Second Utterance responses $f_{i,S}$ is defined as:

$$f_{i,S} = n_{i,S} + {}^{n_{i,U}}\!/_{2} \tag{5.5}$$

[17]Pentium, SoundBlaster 16 sound card, Bayer microphone/earphone headsets, Microsoft Windows 95.

Note that the Undecided responses are divided evenly between the First Utterance and Second Utterance responses in order to use the chi square test as indicated.

The chi square value χ_i^2 for the *i*th stimulus in the *chi square test with one degree of freedom* (Kreyszig, 1988) is calculated as shown below:

$$\chi_i^2 = \frac{(f_{i,F} - e_i)^2}{e_i} + \frac{(f_{i,S} - e_i)^2}{e_i} \tag{5.6}$$

where e_i is the *expected frequency*:

$$e_i = \frac{n_{i,F} + n_{i,S} + n_{i,U}}{2} \tag{5.7}$$

If χ_i^2 is greater than some threshold, it means that the particular majority, be it First Utterance or Second Utterance, is significant.

5.8.2 Perception Test 1

The results obtained for Perception Test 1 are summarized in Table 5.14. Results for mean response time (MRT) are presented in Table 5.15. The following findings were derived from these tables:

1. A percentage of 3.3 undecided responses were observed.
2. For the three tonal groups, on average, the subjects preferred $\tilde{B}$, *BP* and *BD* over $\tilde{A}$.
3. There was a particularly strong preference for $\tilde{B}$ and *BP* over $\tilde{A}$.
4. For the $\tilde{A}$ vs. $\tilde{B}$ and $\tilde{A}$ vs. *BP* sections of the perception test, the average MRT was lower than that of the $\tilde{A}$ vs. *BD* section.
5. *BD* achieved a statistically insignificant majority over $\tilde{A}$.
6. In 6 out of 9 instances either *BP* or *BD* was preferred over $\tilde{A}$.
7. Subjects preferred $\tilde{A}$ for the FL tone group. According to the chi square test, however, $\tilde{A}$ only achieved a statistically significant majority once i.e. over *BP*.
8. The average MRT was higher for the FL stimuli than for the HL and LL stimuli.

Conclusions

It was established that, on average, the subjects preferred the singular command versions where the 'best' values for both pitch and duration were used, followed by versions where either the 'best' pitch or duration values were applied. Since the subjects accepted the 'best' versions as being commands, it was concluded that the results of the acoustic analysis could indeed be applied successfully to generate plural and singular commands with perceptually acceptable prosodics.

It was found that both pitch and duration play a role in the perception of singular imperatives, with pitch being the more significant factor of the two. The best synthesis results will, however, be obtained when both pitch and duration are modelled.

For two out of three tonal groups, subjects were able to clearly detect the difference in prosodics between the infinitive and the command.

It is inferred that subjects took longer to respond to stimuli of which the prosodics were not convincingly that of acceptable commands.

With only 3.3% undecided responses overall, it is clear that subjects made an effort to respond to all stimuli. Otherwise the number of undecided responses would have been higher. The question may arise whether these responses are in fact meaningful. Although it is not easy to quantify the understanding and dedication of subjects to their task, the results indicate definite trends in preferences that correlate with response times. This would not have been possible if subjects responded randomly.

For high quality synthesis it is recommended that the mean values derived from the acoustic analysis be applied.

Table 5.14 Results of Perception Test 1.

Observation	**Percentage of stimuli**			
	FL	**HL**	**LL**	**Average**
Subjects prefer $\tilde{B}$ over $\tilde{A}$	40.4%	***94.9%***	***88.9%***	***74.7%***
Subjects prefer *BP* over $\tilde{A}$	***29.3%***	***96.0%***	***91.9%***	***72.4%***
Subjects prefer *BD* over $\tilde{A}$	40.4%	***60.6%***	50.5%	50.5%

* Percentages shown in bold italic print are significant according to the chi square test.

Table 5.15 Mean response time for Perception Test 1.

Mean Response Time	
min	1.017 s
max	1.673 s
mean	1.403 s
mean FL	1.511 s
mean HL	1.328 s
mean LL	1.369 s
mean $\tilde{A}$ vs. $\tilde{B}$	1.324 s
mean $\tilde{A}$ vs. *BD*	1.528 s
mean $\tilde{A}$ vs. *BP*	1.356 s

5.8.3 Perception Test 2

The results obtained for Perception Test 2 are summarized in Table 5.16. MRT results are presented in Table 5.17. The following findings were derived from these tables:

1. A percentage of 1.3 undecided responses were observed.
2. The subjects preferred the $\tilde{C}$ and $\tilde{E}$ versions to the *BT* and *DT* versions, albeit with only a very slight statistically insignificant majority.
3. In 4 out 6 instances the subjects preferred the utterance with the pitch and duration values that belong to the imperative with verb/object (i.e. . $\tilde{C}$ or $\tilde{E}$).
4. There were no significant differences between the average response times for different tone groups or for the sets of singular and plural stimuli.

Conclusions

Since the subjects' preference for the so-called 'best' singular and plural imperative with object/adverb showed a statistically insignificant majority score, albeit only a slight majority, the conclusion was drawn that the subjects were not able to convincingly distinguish between the two versions.

In general, subjects did accept the so-called 'best' versions as commands followed by an object or adverb and therefore the following proposal could be made for synthesis purposes. Preferably the mean prosodic values, as derived from the acoustic analysis, may be used, but if that level of complexity poses problems in a practical design, the values of the word in isolation may be used without a significant degradation in perceptual quality.

In other words, sub-optimal, but intelligible speech can be synthesized from an acoustic analysis of isolated words instead of sentences. However, it must be stressed that higher quality speech will be obtained if the prosodics of the sentence as a whole is modeled, albeit at a higher computational cost.

The lack of strong preference trends observed correlates with the minimal variation in MRT.

Table 5.16 Results of Perception Test 2.

Observation	**Percentage of stimuli**			
	FL	**HL**	**LL**	**Average**
Subjects prefer $\tilde{C}$ over *BT*	58.1%	***70.5%***	***22.9%***	50.5%
Subjects prefer $\tilde{E}$ over *DT*	46.7%	52.4%	54.3%	51.1%

*Percentages shown in bold italic print are significant according to the chi square test.

Table 5.17 Mean response time for Perception Test 2.

Mean Response Time	
min	1.300 s
max	1.636 s
mean	1.448 s
mean FL	1.466 s
mean HL	1.434 s
mean LL	1.442 s
mean $\tilde{C}$ vs. *BT*	1.415 s
mean $\tilde{E}$ vs. *DT*	1.481 s

5.8.4 Perception Test 3

The results obtained for Perception Test 3 are summarized in Table 5.19. MRT results are presented in Table 5.17. The following findings were derived from these tables:

1. A percentage of 0.2 undecided responses were observed.
2. In all cases the subjects preferred $\tilde{B}$ to *BL1/BM1/BM2* and $\tilde{D}$ to *DL1/DM1/DM2*. On average, this preference was statistically significant.
3. Subjects very strongly (80.5% and 96.7%) preferred $\tilde{B}$ over *BM2* and $\tilde{D}$ over *DM2*.
4. A slight decrease in MRT was observed for *BL1-BM1- BM2* and *DL1-DM1-DM2*.

Conclusions

Subjects preferred the command versions where mean values, as derived from the acoustic analysis, were applied, although there does seem to be room for slight deviation. It is assumed that the short response time for the *M2* steps is a result of subjects' strong and immediate rejection of these prosodic properties.

Due to the scope of this experiment, no strong recommendations could be made for synthesis purposes other than that deviations from the mean values as obtained through the acoustic analysis should not exceed 33.3%. In a dedicated experiment with a greater variety of stimuli, more accurate results considering perceptual thresholds could be obtained.

Table 5.18 Results of Perception Test 3.

Observation	**Percentage of stimuli**			
	FL	**HL**	**LL**	**Average**
Subjects prefer $\tilde{B}$ over *BL1*	55.6%	***62.6%***	***64.6%***	***60.9%***
Subjects prefer $\tilde{B}$ over *BM1*	***63.6%***	56.6%	***79.8%***	***66.7%***
Subjects prefer $\tilde{B}$ over *BM2*	***71.7%***	***76.8%***	***92.9%***	***80.5%***
Subjects prefer $\tilde{D}$ over *DL1*	***68.7%***	***73.7%***	***77.8%***	***73.4%***
Subjects prefer $\tilde{D}$ over *DM1*	***72.7%***	57.6%	***81.8%***	***70.7%***
Subjects prefer $\tilde{D}$ over *DM2*	***100%***	***96.1%***	***94.1%***	***96.7%***

* Percentages shown in bold italic print are significant according to the chi square test.

Table 5.19 Mean response time for Perception Test 3.

Mean Response Time	
min	1.300 s
max	1.733 s
mean	1.391 s
mean FL	1.376 s
mean HL	1.392 s
mean LL	1.405 s
mean $\tilde{B}$ vs. *BL1*	1.436 s
mean $\tilde{B}$ vs. *BM1*	1.430 s
mean $\tilde{B}$ vs. *BM2*	1.359 s
mean $\tilde{D}$ vs. *DL1*	1.411 s
mean $\tilde{D}$ vs. *DM1*	1.383 s
mean $\tilde{D}$ vs. *DM2*	1.325 s

5.8.5 Generalization of Acoustic Results

The same method of stimuli generation was used for verbs that were acoustically analyzed previously, as for the tree verbs that were not previously analyzed (ref. Table 5.1). Up to this point, the perceptual results were calculated without distinguishing between these two sets of stimuli. To determine how subjects responded to the words that were excluded from the analyses, compared to those that were analysed, the results were separated for these two sets.

The averaged results shown in Tables 5.20, 5.21 and 5.22 below indicate that similar trends occurred for the two sets of stimuli. For this limited data set, this proves that the synthesis parameters are applicable to other imperatives with the /C + ala/ and /C + alani/ structures. Moreover, acoustic and synthesis parameters may be extracted from a relatively small set of a particular structure and may then be generalized to other variations of that structure.

Table 5.20 Generalization results for Perception Test 1.

Words in analysis set	
Observation	**Percentage of stimuli**
Subjects prefer $\tilde{B}$ over $\tilde{A}$	***74.7%***
Subjects prefer *BP* over $\tilde{A}$	***72.7%***
Subjects prefer *BD* over $\tilde{A}$	53%
Words not in analysis set (generalization)	
Observation	**Percentage of stimuli**
Subjects prefer $\tilde{B}$ over $\tilde{A}$	***74.7%***
Subjects prefer *BP* over $\tilde{A}$	***71.7%***
Subjects prefer *BD* over $\tilde{A}$	45.5%

* Percentages shown in bold italic print are significant according to the chi square test.

Table 5.21 Generalization results for Perception Test 2.

Words in analysis set	
Observation	**Percentage of stimuli**
Subjects prefer $\tilde{C}$ over *BT*	48.6%
Subjects prefer $\tilde{E}$ over *DT*	50%
Words not in analysis set (generalization)	
Observation	**Percentage of stimuli**
Subjects prefer $\tilde{C}$ over *BT*	54.3%
Subjects prefer $\tilde{E}$ over *DT*	53.3%

* Percentages shown in bold italic print are significant according to the chi square test.

Table 5.22 Generalization results for Perception Test 3.

Words in analysis set	
Observation	**Percentage of stimuli**
Subjects prefer $\tilde{B}$ over *BL1*	***59.1%***
Subjects prefer $\tilde{B}$ over *BM1*	***69.2%***
Subjects prefer $\tilde{B}$ over *BM2*	***76.8%***
Subjects prefer $\tilde{D}$ over *DL1*	***79.3%***
Subjects prefer $\tilde{D}$ over *DM1*	***72.2%***
Subjects prefer $\tilde{D}$ over *DM2*	***98%***
Words not in analysis set (generalization)	
Observation	**Percentage of stimuli**
Subjects prefer $\tilde{B}$ over *BL1*	***64.6%***
Subjects prefer $\tilde{B}$ over *BM1*	***61.6%***
Subjects prefer $\tilde{B}$ over *BM2*	***87.9%***
Subjects prefer $\tilde{D}$ over *DL1*	***61.6%***
Subjects prefer $\tilde{D}$ over *DM1*	***67.7%***
Subjects prefer $\tilde{D}$ over *DM2*	***94.1%***

* Percentages shown in bold italic print are significant according to the chi square test.

5.9 General Conclusions

The fact that meaningful as well as positive results could be obtained from the perceptual experiment proved firstly, that an appropriate method of analysis was applied and secondly, that the acoustic results were applied successfully to generate commands in different contexts. An advantage of this method of prosodic parameter extraction is that the analysis of a small database will be sufficient, since the parameters extracted for a particular structure will be applicable to variations of that structure.

Although no loudness manipulation was performed, the subjects still accepted the stimuli as being commands. From this we may infer that loudness is not a prerequisite for the synthesis of commands from a perceptual point of view.

The results of all three tests indicated that trends in preferences correlated with response times. If the prosodic parameters of stimuli were perceptually acceptable, the MRT for these stimuli were relatively short. Conversely, if the prosodics were not perceptually acceptable, the MRT for these stimuli were relatively long. The lack of strong preference trends in turn, correlated with a minimum variation in MRT.

Combining the results of the acoustic analysis and the perceptual experiment made it possible to present a prosodic model for the generation of perceptually acceptable imperatives in a practical Xhosa TTS system. A demonstration of the application of this model is presented in the following section.

5.10 Application of Results

A combination of the results of the acoustic analysis and the perceptual experiment will now be applied to demonstrate how prosodic parameters can be calculated in a Xhosa TTS system. The prosody generation module of the TTS system will be illustrated for the same two sentences as were used in Chapter 3 (cf. 3.3.4).

Example 1: *Lala!*

Taking all the information given in the text analysis, language and spelling checker and automatic phonetization modules into account, the system generates a singular imperative without an object/adverb (in other words a verb with a structure that is similar to that of B) by applying the percentage duration and pitch changes to the infinitive form, as calculated in the acoustic analysis.

The duration values for synthesis are calculated as follows:
The duration of the first consonant /l/, $d_{\bar{B}:l1}$, is taken as the mean duration of the /l/ of the infinitive context $d_{A:l1}$ as found in Table C.2:

(5.8) $d_{\bar{B}:l1} = d_{A:l1} = 73.8$ ms

Table F.1 shows the percentage changes in duration for the remaining vowels and consonants. Duration values for the verb in the infinitive context (A), as shown in Table C.3, are used and the percentage changes are applied to these values. According to Table F.1 the duration of /a1/, for example, is calculated by increasing the value for /a1/ in the infinitive context by 11%. The duration values for /a1/, /l/ and /a2/ are calculated as described below:

(5.9) $d_{\tilde{B}:a1} = d_{A:a1} + \Delta d_{\tilde{B}:a1} = d_{A:a1} + d_{A:a1}\Delta d_{A\to\tilde{B}:a1} = d_{A:a1}\left(1 + \Delta d_{A\to\tilde{B}:a1}\right)$

(5.10) $d_{\tilde{B}:a1} = 231.9\text{ ms}\left(1+0.11\right) = 231.9\text{ ms}\times 1.11 = 257.9\text{ ms}$

(5.11) $d_{\tilde{B}:l2} = 86\text{ ms}\left(1+0.11\right) = 86\text{ ms}\times 1.11 = 95.5\text{ ms}$

(5.12) $d_{\tilde{B}:a2} = 101.8\text{ ms}\left(1+0.11\right) = 101.8\text{ ms}\times 1.11 = 113.0\text{ ms}$

Pitch is calculated as follows: for the first consonant /l/ the mean pitch of the verb in the infinitive context is used. This value can be found in Table D.1 in Appendix D.

(5.13) $p_{\tilde{B}:l1} = p_{A:l1} = 171.2\text{ Hz}$

Table F.2 in Appendix F provides the percentage changes in pitch for the remaining vowels and consonants in the verb. Pitch values for the infinitive form A as shown in Table D.3 in Appendix D are used and the percentage changes are applied to these values. According to Table F.2 the pitch of /a1/ is the same as that of /a1/ in the infinitive context, therefore no changes apply for this value. However, the table shows that the pitch of /l2/ is 43% higher than that of /l2/ in the infinitive context.

(5.14) $p_{\tilde{B}:a1} = p_{A:a1} = 154.9\text{ Hz}$

(5.15) $p_{\tilde{B}:l2} = p_{A:l2} + \Delta p_{\tilde{B}:l2} = p_{A:l2} + p_{A:l2}\Delta p_{A\to\tilde{B}:l2} = p_{A:l2}\left(1 + \Delta p_{A\to\tilde{B}:l2}\right)$

(5.16) $p_{\tilde{B}:l2} = 125.4\text{ Hz}\left(1+0.43\right) = 125.4\text{ Hz}\times 1.43 = 179.3\text{ Hz}$

(5.17) $p_{\tilde{B}:a2} = 113.9\text{ Hz}\left(1+0.37\right) = 113.9\text{ Hz}\times 1.37 = 156.0\text{ Hz}$

Example 1: *Bhalani* page 3!

Taking all the information given in the text analysis, language and spelling checker and automatic phonetization modules into account, the system generates a plural imperative followed by an object as well as an adjective, by applying the percentage pitch and duration changes to the infinitive form, as calculated in the acoustic analysis.

Duration values for *bhalani* is calculated as follows: for the first consonant /bh/ the mean consonant duration of the verb in the infinitive context is used. This value can be found in Table C.2 in Appendix C.

(5.18) $d_{\tilde{E}:bh} = d_{A:bh} = 102.4 \text{ ms}$

Table F.1 in Appendix F provides the percentage changes in duration for the remaining vowels and consonants. Duration values for the verb in the infinitive context (A), as shown in Table C.3, are used and the percentage changes are applied to these values. According to Table F.1 the duration of /a1/, for example, is calculated by decreasing the value for /a1/ in the infinitive context by 59%. The duration values for /a1/, /l/, /a2/, /n/ and /i/ are calculated as described below:

(5.19) $d_{\tilde{E}:a1} = d_{A:a1} + \Delta d_{\tilde{E}:a1} = d_{A:a1} + d_{A:a1}\Delta d_{A\rightarrow\tilde{E}:a1} = d_{A:a1}\left(1 + \Delta d_{A\rightarrow\tilde{E}:a1}\right)$

(5.20) $d_{\tilde{E}:a1} = 231.9 \text{ ms}\left(1 - 0.59\right) = 231.9 \text{ ms} \times 0.41 = 95.079 \text{ ms}$

(5.21) $d_{\tilde{E}:l} = 86 \text{ ms}\left(1 - 0.15\right) = 86 \text{ ms} \times 0.85 = 73.1 \text{ ms}$

(5.22) $d_{\tilde{E}:a2} = 101.8 \text{ ms}\left(1 + 0.39\right) = 101.8 \text{ ms} \times 1.39 = 141.502 \text{ ms}$

(5.23) $d_{\tilde{E}:n} = d_{E:n} = 71.5 \text{ ms}^*$

(5.24) $d_{\tilde{E}:i} = d_{E:i} = 81 \text{ ms}^*$

* Note that for the duration of /n/ and /i/ the values of the verb in the imperative, plural context followed by an object/adverb (E) is used, as calculated in the acoustic analysis. These values can also be found in Table C.3.

Pitch values for *bhalani* are calculated as follows: for the first consonant /bh/ the mean pitch of the verb in the infinitive context is used. This value can be found in Table D.1 in Appendix D.

(5.25) $p_{\tilde{E}:bh} = p_{A:bh} = 0Hz^*$

* In this case the consonant is voiceless.

Table F.2 in Appendix F provides the percentage changes in pitch for the remaining vowels and consonants in the verb. Pitch values for the infinitive form A as shown in Table D.3 are used and the percentage changes are applied to these values. According to Table F.2 the pitch of /a1/ is the same as that of /a1/ in the infinitive context. The value of

/l/ in the infinitive context should, however, be increased by 42% and the value of /a2/ in the infinitive context should be increased by 72%. The values used for the /n/ and the /i/ are those calculated for the verb in the imperative, plural with object/adverb context. These values can also be found in Table D.3.

(5.26) $p_{\tilde{E}:a1} = p_{A:a1} + \Delta p_{\tilde{E}:a1} = p_{A:a1} + p_{A:a1}\Delta p_{A\to\tilde{E}:a1} = p_{A:a1}\left(1 + \Delta p_{A\to\tilde{E}:a1}\right)$

(5.27) $p_{\tilde{E}:a1} = p_{A:a1} = 154.9\ \text{Hz}$

(5.28) $p_{\tilde{E}:l} = 125.4\ \text{Hz}\left(1 + 0.42\right) = 125.4\ \text{Hz} \times 1.42 = 178.0\ \text{Hz}$

(5.29) $p_{\tilde{E}:a2} = 113.9\ \text{Hz}\left(1 + 0.72\right) = 113.9\ \text{Hz} \times 1.72 = 195.9\ \text{Hz}$

(5.30) $p_{\tilde{E}:n} = p_{E:n} = 193.4\ \text{Hz}$ *

(5.31) $p_{\tilde{E}:i} = p_{E:i} = 181.0\ \text{Hz}$ *

* Note that for the pitch of /n/ and /i/ the values of the verb in the imperative, plural context followed by an object/adverb (E) are used. These values can also be found in Table D.3.

5.11 Chapter Summary

In this chapter the perceptual experiment conducted on a corpus of synthesized imperatives was discussed. Stimuli for the perceptual experiment were generated by applying the prosodic parameters extracted from the acoustic analysis. Monotonous verbs were recorded by one speaker and the prosodic features, duration and pitch, of this recorded corpus were modified with the TD-PSOLA method. An advantage of this approach was that the implemented programs required minimal manual intervention, making it suitable for automatic manipulation by computer.

These new synthesized utterances were then tested perceptually by means of three different preference tests on pairs of stimuli. The test procedures were automated so that the stimuli could be presented on multimedia computers to a group of mother tongue listeners within a controlled environment. The subjects' responses and response time were automatically collected and subjected to statistical analyses.

Results that may be highlighted are as follows:

- On average, the subjects preferred the singular command versions where the 'best' values for both pitch and duration were used, followed by versions where either the 'best' pitch or duration values were applied.
- Since the subjects accepted the so-called 'best' versions as being commands, it was concluded that the results of the acoustic analysis could indeed be applied successfully to generate plural and singular commands with perceptually acceptable prosodics.
- Both pitch and duration play a role in the perception of singular imperatives, with pitch being the more significant factor of the two. The best synthesis results will, however, be obtained when both pitch and duration are modelled.
- For synthesizing commands with objects or adverbs, the mean prosodic values, as derived from the acoustic analysis, is preferable, but if that level of complexity poses problems in a practical design, the values of the word in isolation may be used without a significant degradation in perceptual quality. In other words, sub-optimal, but intelligible speech can be synthesized from an acoustic analysis of isolated words instead of sentences.
- Strong recommendations could not be made regarding perceptual thresholds except that, for synthesis purposes, deviations from the mean values as obtained through the acoustic analysis should not exceed 33.3%.
- Definite trends in preferences (and even the lack thereof) could be correlated with response times.
- An advantage of the method of prosodic parameter extraction is that the analysis of a subset will be sufficient, since the parameters extracted for a particular structure will be applicable to variations of that structure.

Combining the results of the acoustic analysis and the perceptual experiment made it possible to present a prosodic model for the generation of perceptually acceptable imperatives in a practical Xhosa TTS system. Finally, the application of this model was demonstrated.

Chapter Six

Conclusions

This chapter concludes the study with an overview of results, accomplishments and contributions. Finally, recommendations for future research are given.

In this study it was re-established that speakers use suprasegmental features such as duration, pitch and loudness to convey important information additional to that conveyed by the segmental composition of the utterance. It was shown how the relationship between duration, pitch and loudness, as manifested in the production and perception of Xhosa imperatives in particular, could be determined through acoustic analyses and perceptual experiments. An experimental phonetic approach proved to be essential for the acquisition of substantial and reliable prosodic information.

An extensive acoustic analysis was conducted to acquire prosodic information on the production of imperatives by eight Xhosa mother tongue speakers. A corpus of bi-syllabic imperatives in three different tone groups were digitally recorded and annotated. A consistent and accurate tagging method using speech analysis software was introduced. A range of software was developed to automatically measure duration, pitch and loudness features. Subsequently, various statistical parameters were calculated on the raw acoustic data (i) to establish patterns of significance and (ii) to represent the large amount of numeric data generated, in a compact manner.

The linguistic properties of imperatives in Xhosa were described at various levels. At segmental level, the structure of imperatives were determined as follows: for verb stems that have more than one syllable, the imperative consists of a root, a verbal suffix /-a/ and a suffix /-ni/ in the plural form. At suprasegmental level speakers employed duration, pitch and loudness features to encode commands. Pitch and duration results of linguistic value emerged. For instance, evidence of penultimate syllable lengthening was found and it was established that the tempo of verbs were raised when these verbs were followed by objects or adverbs. Observations regarding statistically significant pitch values on the voiced

consonant /l/ were considered to have implications for tone marking principles. Furthermore significant pitch differences for vowels confirmed certain tonological rules.

A perceptual experiment was conducted to investigate the perception of imperatives. The prosodic parameters that were extracted from the acoustic analysis were applied to synthesize imperatives in different contexts. A novel approach to Xhosa speech synthesis was adopted. Monotonous verbs were recorded by one speaker and the pitch and duration of these words were then manipulated with the TD-PSOLA technique. An advantage of this technique was that the implemented programs required minimal manual intervention, making the approach suitable for automatic manipulation by computer.

In the perceptual experiment the synthesized stimuli were presented to more than thirty Xhosa mother tongue listeners within a controlled environment. The large scale of the experiment required the automation of testing procedures, hence perception test software was developed (i) to present the stimuli to subjects on multimedia computers and (ii) to collect their responses automatically. The method applied was that of the preference test on pairs of stimuli. Through these preferences one could infer which combination of prosodic values the subjects considered to be the most perceptually acceptable for a specific command.

Meaningful and positive results were obtained from the perceptual experiment proving firstly, that an appropriate method of analysis was applied and secondly, that the acoustic results were applied successfully to generate commands in different contexts. The advantage of this method of prosodic parameter extraction is that the analysis of a subset of a particular structure will be sufficient, since these parameters will be applicable for the synthesis of variations of that structure.

Although it was found that pitch plays a more significant role than duration in the perception of singular commands, modelling both these parameters simultaneously, provided for the most natural synthesized commands. For synthesizing commands with objects or adverbs, the mean prosodic values, as derived from the acoustic analysis, are preferable, but if that level of complexity poses problems in a practical design, the values of the word in isolation may be used without a significant degradation in perceptual quality.

Combining the results of the acoustic analysis and the perceptual experiment made it possible to present a prosodic model for the generation of perceptually acceptable imperatives in a practical Xhosa TTS system.

Prosody generation in a natural language processing module and its place within the larger framework of text-to-speech synthesis was discussed. It was shown that existing architectures for TTS synthesis would not be appropriate for Xhosa without some adaptation. Hence, a unique architecture was proposed and subsequently illustrated. Of particular importance was the development of an alternative algorithm for grapheme-to-phoneme conversion. This new algorithm maintains the high level of accuracy without sacrificing processing speed. It is also less complex and can easily be extended to other African languages. These properties make the algorithm suitable for implementation in an automatic phonetization module.

With regard to speech synthesis methods, recommendations for choosing an appropriate synthesis system for Xhosa were also presented. It was established that the corpus-based synthesis method produces intelligible and natural sounding speech with low computational cost and complexity. Bearing these advantages in mind, it was concluded that corpus-based synthesis may be an appropriate method to use for a Xhosa TTS system with service oriented applications.

The prosodic model derived in this study may not only have implications for TTS synthesis; it may also be useful for the development of automatic speech recognition systems. For instance, the synthesis and recognition of commands may be applied in the support of manpower deployment and speech based information retrieval systems.

It is believed that this study may have made a significant contribution towards the establishment of spoken language technology for Xhosa and the advancement of this field in South Africa.

Future Research

Future research and development of the following aspects were considered to be of particular importance for the advancement of spoken language technology and the field of linguistic studies in general:

- There is a great need for the establishment of standardized, national databases of digital text and speech material.
- Automatic tagging of both text and speech data is essential, considering the amount of data required.
- The derivation of prosodic models should be extended to include all structures of the Xhosa language.
- Sub-word concatenative TTS systems holds the promise of the implementation of practical, unlimited vocabulary, natural sounding speech synthesis systems for use in various domains.

Appendix A

TTS systems on the Internet

Table A.1 Speech Synthesis Systems

NAME	TYPE	COMPANY	LANGUAGE
Soft Voice TTS System	Rule-based (formant)	Soft Voice, Inc	English Spanish
Available: http://www.webcom.com/tts/welcome.html			
ETI-Eloquence	Rule-based (formant)	Eloquent Technology, Inc	English
Available: http://www.eloq.com/			
Speak	Rule-based Articulatory	National Centre for Voice and Speech	
Was available: http://ncvs.shc.uiowa.edu/research/speak/index.html			
ASY	Rule-based Articulatory	Haskins Laboratories	
Available: http://www.haskins.yale.edu/haskins/MISC/ASY/ASY.html			
Yorktalk	Rule-based Formant	University of York	English
Available: http://www-users.york.ac.uk/~lang4/Yorktalk.html			
MBROLA	Concatenative	TCTS Laboratory	11 European languages
Available: http://tcts.fpms.ac.be/			
Festival	Concatenative	CSTR	British English American English Spanish Welsh
Available: http://www.cstr.ed.ac.uk/projects/festival/			
Bell Labs TTS system	Concatenative	Bell Laboratories	American English
Available: http://www.bell-labs.com/project/tts/			
Eurovocs		Technologie & Revalidatie at ELIS Speech Lab	Dutch French German American English
Available: http://www.tni.be/product.html			
Laureate TTS system		BT Laboratories	British English American English Prototype versions for several European languages
Available: Available: http://innovate.bt.com/showcase/laureate/index.htm			

Appendix B

System of Phoneme Annotation

Table B.1 System of phoneme annotation.

bhala	
Filename	**Tags**
PAFL01A.NSP	u1 u2* bh bha1 a1 l a2]
PAFL01B.NSP	bh bha1 a1 l a2]
PAFL01C.NSP	bh bha1 a1 l a2]
PAFL01D.NSP	bh bha1 a1 l a2 n i]
PAFL01E.NSP	bh bha1 a1 l a2 n i]
Tag	**Tag position**
u1	beginning of first /u/ of /uku/
u2*	end of second /u/ of /uku/
bh	burst of /bh/
bha1	beginning of first /a/ (tagged at onset of voice)
a1	tagged at beginning of steady state of first /a/
l	beginning of /l/
a2	beginning of second /a/
n	beginning of /n/
i	beginning of /i/ for all plural words
]	end of word
dlala	
Filename	**Tags**
PAFL02A.NSP	u1 dl dla1 a1 l a2]
PAFL02B.NSP	dl dla1 a1 l a2]
PAFl02C.NSP	dl dla1 a1 l a2]
PAFL02D.NSP	dl dla1 a1 l a2 n i]
PAFL02E.NSP	dl dla1 a1 l a2 n i]
Tag	**Tag position**
u1	beginning of first /u/ of /uku/
dl	end of second /u/ of /uku/, beginning of /dl/ (tagged at beginning of high frequency noise)
dla1	beginning of first /a/
. . .	

Table B.1 System of phoneme annotation (continued).

lala	
Filename	**Tags**
PAHL01B.NSP	u1 l1 a1 l2 a2]
PAHL01A.NSP	l1 a1 l2 a2]
PAHL01C.NSP	l1 a1 l2 a2]
PAHL01D.NSP	l1 a1 l2 a2 n i]
PAHL01E.NSP	l1 a1 l2 a2 n i]
Tag	**Tag position**
u1	beginning of first /u/ of /uku/
l1	end of second /u/ of /uku/, beginning of first /l/
. . .	
hlala	
Filename	**Tags**
PAHL02A.NSP	u1 hl hla1 a1 l a2]
PAHL02B.NSP	hl hla1 a1 l a2]
PAHL02C.NSP	hl hla1 a1 l a2]
PAHL02D.NSP	hl hla1 a1 l a2 n i]
PAHL02E.NSP	hl hla1 a1 l a2 n i]
Tag	**Tag position**
u1	beginning of first /u/ of /uku/
hl	end of second /u/ of /uku/, beginning of /hl/ (tagged at beginning of high frequency noise)
hla1	beginning of first /a/
. . .	
bala	
Filename	**Tags**
PALL01A.NSP	u1 u2* b ba1 a1 l a2]
PALL01B.NSP	*b b ba1 a1 l a2]
PALL01C.NSP	*b b ba1 a1 l a2]
PALL01D.NSP	*b b ba1 a1 l a2 n i]
PALL01E.NSP	*b b ba1 a1 l a2 n i]
Tag	**Tag position**
u1	beginning of first /u/ of /uku/
u2*	end of second /u/ of /uku/ and beginning of implosive /b/ (tagged at closure of lips) (only for infinitive form)
*b	beginning of implosive /b/ (tagged at closure of lips)
b	burst of implosive /b/
ba1	beginning of first /a/ (tagged at onset of voice)
. . .	

Table B.1 System of phoneme annotation (continued).

Tsala	
Filename	**Tags**
PALL02A.NSP	u1 u2* ts tsa1 a1 l a2]
PALL02B.NSP	ts tsa1 a1 l a2]
PALL02C.NSP	ts tsa1 a1 l a2]
PALL02D.NSP	ts tsa1 a1 l a2 n i]
PALL02E.NSP	ts tsa1 a1 l a2 n i]
Tag	**Tag position**
u1	beginning of first /u/ of /uku/
u2*	end of second /u/ of /uku/
ts	release of /ts/
tsa1	beginning of first /a/
. . .	

Appendix C

Tables Summarizing the Acoustic and Statistic Results for Duration Features

Table C.1 Mean duration (in ms).

bhala (FL01)											
	d_{uku}	CD_{bh}	VOT_{bh}	d_{bha1}	d_{a1}	d_{l}	d_{a2}	d_{n}	d_{i}	d_{ala}	d_{alani}
A	229.2	86.5	22.9	22.7	259.7	91.5	91.6			488.5	
B			13.2	17.3	282.6	89	93.9			496	
C			14	17.9	157.3	71	103.4			363.6	
D			15.7	18.7	111.6	76.4	272.3	109.9	92.3	494.6	696.8
E			13.8	18.1	105.7	80.4	151.7	77.9	69.9	369.7	517.6
dlala (FL02)											
	d_{uku}	d_{dl}	d_{dla1}	d_{a1}	d_{l}	d_{a2}	d_{n}	d_{i}	d_{ala}	d_{alani}	
A	250.9	131.2	25.6	227.4	86.4	109.8			580.3		
B		155.8	22	257.3	95.4	126.1			656.6		
C		141.5	19.5	144.6	63.3	130			499		
D		127	19.4	111.9	80.3	276.3	102.8	112.4	614.8	830	
E		134	21	91.6	71.6	152.4	69.9	85.6	470.6	626.1	
lala (HL01)											
	d_{uku}	d_{l1}	d_{a1}	d_{l2}	d_{a2}	d_{n}	d_{i}	d_{ala}	d_{alani}		
A	231.3	79.1	233.5	80.2	111.9			504.6			
B		83.2	247.2	89	142.1			561.6			
C		72.8	136.8	64.2	125.5			399.3			
D		67.9	107.5	83.2	266.6	96.3	105	525.2	726.4		
E		66.2	100.3	75.8	136.8	72.3	69.6	379	520.8		
hlala (HL02)											
	d_{uku}	d_{hl}	d_{hla1}	d_{a1}	d_{l}	d_{a2}	d_{n}	d_{i}	d_{ala}	d_{alani}	
A	207.4	143.1	22.5	209.6	81.7	89.7			546.6		
B		173.1	20.8	243.9	78.2	146.2			662.1		
C		149.3	17	88.9	59	106.8			420.9		
D		148.1	17.2	90.2	75.6	235.3	92	127.7	566.4	786.1	
E		129.2	13.9	78.4	67.4	117.2	63.6	86	405.9	555.5	

Table C.1 Mean duration (continued).

bala (LL01)											
	d_{uku}	**CD_b**	**VOT_b**	**d_{ba1}**	**d_{a1}**	**d_l**	**d_{a2}**	**d_n**	**d_i**	**d_{ala}**	**d_{alani}**
A	243	67.2	5.9	23	236.1	84.3	118.6			529.4	
B		62.6	6	17.7	245.1	77	160.6			569	
C		37.9	7.3	19.8	133.1	67	100.4			365.5	
D		49	7.2	20	115.9	79.7	253.6	97.6	129.5	525.4	752.5
E		46.4	8	17.1	100.4	73.7	155.6	74.8	89.6	401.1	565.5
tsala (LL02)											
	d_{uku}	**CD_{ts}**	**d_{ts}**	**d_{tsa1}**	**d_{a1}**	**d_l**	**d_{a2}**	**d_n**	**d_i**	**d_{ala}**	**d_{alani}**
A	224.2	48.5	76.7	27.7	225.3	91.9	88.9			510.4	
B			76	27.1	217.5	77.9	172.4			570.8	
C			64.6	24.3	116.7	60.4	118.3			384.4	
D			74	15.1	100.1	78.7	250.9	98.2	93.7	518.8	710.7
E			66.5	22.3	89.3	71.1	137.3	70.5	85.4	386.5	542.4

Table C.2 Mean consonant duration (in ms).

	d_{uku}	**CD_C**	**VOT_C**	**d_C**	**Td_C**	**d_{Ca1}**
bh	229.2	86.5	15.9		102.4 [1]	18.9
dl	250.9			137.9	137.9	21.5
l	231.3			73.8	73.8	
hl	207.4			148.6	148.6	18.3
b	243	52.6	6.9		59.5 [1]	19.5
ts	224.2	48.5		71.6	120.1 [2]	23.3

[1] Total consonant duration $Td_C = CD_C + VOT_C$
[2] Total consonant duration $Td_C = CD_C + d_C$

Table C.3 Mean duration for /ala/ or /alani/ for all words (in ms).

	d_{a1}	**d_l**	**d_{a2}**	**d_n**	**d_i**	**d_{ala}**	**d_{alani}**
A	231.9	86	101.8			526.6	
B	248.9	84.4	140.2			586	
C	129.6	64.2	114.1			405.5	
D	106.2	79	259.2	99.5	110.1	540.9	750.4
E	94.3	73.3	141.8	71.5	81	402.1	554.7

Table C.4 Percentage duration changes calculated from acoustic data.

	Order of significance	Δd_{a1}	Δd_{l}	Δd_{a2}	Δd_{n}	Δd_{i}	Δd_{ala}	Δd_{alani}
A→B	a1 l a2, **D**$_{-}$	7.3%	-1.9%	37.7%			11.3%	
A→C	**a1**$_{1+}$ **l**$_{2+}$ a2, **D**$_{+}$	-44.1%	-25.3%	12.1%			-23%	
A→D	**a1**$_{1+}$ **l**$_{3+}$ **a2**$_{2-}$, **D**$_{+}$	-54.2%	-8.1%	154.6%			+2.7%	
A→E	**a1**$_{1+}$ **l**$_{3+}$ **a2**$_{2-}$, **D**$_{+}$	-59.3%	-14.8%	39.3%			-23.6%	
B→C	**a1**$_{1+}$ **l**$_{2+}$ a2, **D**$_{+}$	-47.9%	-23.9%	-18.6%			-30.8%	
D→E	a1 l **a2**$_{1+}$ **n**$_{2+}$ i, **D**$_{+}$, **TD**$_{+}$	-11.2%	-7.2%	-45.3%	-28.1%	-26.4%	-25.7%	-26.1%
B→D	**a1**$_{1+}$ l **a2**$_{2-}$	-57.3%	-6.4%	84.9%			-7.7%	
C→E	**a1**$_{+}$ l a2 *	-27.2%	14.2%	24.3%			-0.8%	

* No consistent order of significance pattern was observed.

Appendix D

Tables Summarizing the Acoustic and Statistic Results for Pitch Features

Table D.1 Median pitch for consonants in the infinitive context (in Hz).

	bh	dl	l	hl	b	ts
A	unvoiced	115.1	171.2	194.1	180.1	unvoiced

Table D.2 Median pitch of /ala/ or /alani/ (in Hz).

bhala (FL01)								
	p_{bha1}	p_{a1}	p_{l}	p_{a2}	p_{n}	p_{i}	p_{ala}	p_{alani}
A	137.4	136.1	104.9	103.3			130.3	
B	143.3	163.9	122.2	108.5			151.2	
C	113.2	133.3	145.8	151.9			141.6	
D	131	133.4	146.8	173.5	179.8	158.2	161.8	165.5
E	136.3	129.7	144	174.5	196.4	185.2	154.9	165.8
dlala (FL02)								
	p_{dla1}	p_{a1}	p_{l}	p_{a2}	p_{n}	p_{i}	p_{ala}	p_{alani}
A	122.2	125.1	97	99.2			120.2	
B	142.3	172.3	127.1	107.6			153.2	
C	140.4	168.4	166.5	144.5			161.1	
D	133.3	134.2	148.1	170.3	169.1	142.9	156.3	155.6
E	131	132.1	142.6	173.3	190.3	176.7	153.6	162.4
lala (HL01)								
	p_{a1}	p_{l2}	p_{a2}	p_{n}	p_{i}	p_{ala}	p_{alani}	
A	153.4	148	139.2			154.4		
B	168.4	173.9	152.6			162.8		
C	162.2	178.8	166.7			163.6		
D	143.7	152.9	177.9	186.9	159.4	167.8	171.6	
E	153.9	167	190.9	199.9	184.5	173.6	179.4	

Table D.2 Median pitch of /ala/ or /alani/ (continued).

hlala (HL02)								
	p_{hla1}	p_{a1}	p_l	p_{a2}	p_n	p_i	p_{ala}	p_{alani}
A	195.4	156.3	102.8	88.6			153.3	
B	178.9	183.9	183.5	159.4			180.9	
C	179	179.5	187	177			179.3	
D	169.6	165.6	165.7	182.1	183.5	156.9	174.9	174.2
E	146.7	143.2	149.2	164.1	173	162	154.7	159.9
bala (LL01)								
	p_{ba1}	p_{a1}	p_l	p_{a2}	p_n	p_i	p_{ala}	p_{alani}
A	180.1	157.6	102.2	98.2			152.8	
B	150.8	152.9	148.5	169.8			155.7	
C	159.5	160.9	160.6	167.3			161.9	
D	155.6	158.3	162.2	177.7	182.9	160.1	170.4	172.3
E	157.1	162.5	171.6	189.6	195.9	184.9	178.2	181.7
tsala (LL02)								
	p_{tsa1}	p_{a1}	p_l	p_{a2}	p_n	p_i	p_{ala}	p_{alani}
A	187.5	156	103.9	81.6			141.5	
B	131.8	141.6	154	178.8			166.8	
C	167.5	159.8	155.6	153.2			158	
D	124.4	154.9	157.6	177.3	175.4	142.4	169.4	169.8
E	156.5	162.3	171.5	189.4	196	186.2	178.8	182.9

Table D.3 Mean pitch for tonal groups (in Hz).

Mean pitch for FL01 and FL02								
	p_{Ca1}	p_{a1}	p_{l}	p_{a2}	p_{n}	p_{i}	p_{ala}	p_{alani}
A	129.8	130.6	100.95	101.25			125.25	
B	142.8	168.1	124.65	108.05			152.2	
C	126.8	150.85	156.15	148.2			151.35	
D	132.15	133.8	147.45	171.9	174.45	150.55	159.05	160.55
E	133.65	130.9	143.3	173.9	193.35	180.95	154.25	164.1
Mean pitch for HL01 and HL02								
	p_{Ca1}	p_{a1}	p_{l}	p_{a2}	p_{n}	p_{i}	p_{ala}	p_{alani}
A	195.4	154.85	125.4	113.9			153.85	
B	178.9	176.15	178.7	156			171.85	
C	179	170.85	182.9	171.85			171.45	
D	169.6	154.65	159.3	180	185.2	158.15	171.35	172.9
E	146.7	148.55	158.1	177.5	186.45	173.25	164.15	169.65
Mean pitch for LL01 and LL02								
	p_{Ca1}	p_{a1}	p_{l}	p_{a2}	p_{n}	p_{i}	p_{ala}	p_{alani}
A	183.8	156.8	103.05	89.9			147.15	
B	141.3	147.25	151.25	174.3			161.25	
C	163.5	160.35	158.1	160.25			159.95	
D	140	156.6	159.9	177.5	179.15	151.25	169.9	171.05
E	156.8	162.4	171.55	189.5	195.95	185.55	178.5	182.3

Table D.4 Percentage pitch changes calculated from acoustic data.

Percentage pitch changes for FL								
	Order of significance	Δp_{a1}	Δp_{l}	Δp_{a2}	Δp_{n}	Δp_{i}	Δp_{ala}	Δp_{alani}
A→B	**a1**$_{1}$- **l**$_{2}$- a2, **P.**	28.7%	23.5%	6.7%			21.5%	
A→C	a1 **l**$_{1}$- **a2**$_{2}$-, **P.**	15.5%	54.7%	46.4%			20.8%	
A→D	a1 **l**$_{2}$- **a2**$_{1}$-, **P.**	2.5%	46.1%	69.8%			27%	
A→E	a1 **l**$_{2}$- **a2**$_{1}$-, **P.**	0.2%	42%	71.8%			23.2%	
B→C	a1 **l**$_{1}$- **a2**$_{2}$-	-10.3%	25.3%	37.2%			-0.6%	
D→E	a1 l a2 n i	-2.2%	-2.8%	1.2%	10.8%	20.2%	-3%	2.2%
B→D	a1 l a2, **P.**	-20.4%	18.3%	59.1%			4.5%	
C→E	a1 l a2, **P.**	-13.2%	-8.2%	17.3%			1.9%	
Percentage pitch changes for HL								
	Order of significance	Δp_{a1}	Δp_{l}	Δp_{a2}	Δp_{n}	Δp_{i}	Δp_{ala}	Δp_{alani}
A→B	a1 **l**$_{1}$- **a2**$_{2}$-, **P.**	13.8%	42.5%	37%			11.7%	
A→C	a1 **l**$_{1}$- **a2**$_{2}$-	10.3%	45.9%	50.9%			11.4%	
A→D	a1 **l**$_{2}$- **a2**$_{1}$-, **P.**	-0.1%	27%	58%			11.4%	
A→E	a1 **l**$_{2}$- **a2**$_{1}$-, **P.**	-4.1%	26.1%	55.8%			6.7%	
B→C	a1 l a2, **P-**	-3%	2.4%	10.2%			-0.2%	
D→E	a1 l a2 n i	-3.9%	-0.8%	-1.4%	0.7%	9.5%	-4.2%	-1.9%
B→D	**a1**$_{1+}$ **l**$_{2+}$ a2, **P.**	-12.2%	-10.9%	15.4%			-0.3%	
C→E	**a1**$_{1+}$ l a2	-13.1%	-13.6%	3.3%			-4.3%	
Percentage pitch changes for LL								
	Order of significance	Δp_{a1}	Δp_{l}	Δp_{a2}	Δp_{n}	Δp_{i}	Δp_{ala}	Δp_{alani}
A→B	a1 **l**$_{2}$- **a2**$_{1}$-	-6.1%	46.8%	93.9%			9.6%	
A→C	a1 **l**$_{2}$- **a2**$_{1}$-	2.3%	53.4%	78.3%			8.7%	
A→D	a1 **l**$_{2}$- **a2**$_{1}$-	-0.1%	55.2%	97.4%			15.5%	
A→E	a1 **l**$_{2}$- **a2**$_{1}$-	3.6%	66.5%	110.8%			21.3%	
B→C	a1 l a2	8.9%	4.5%	-8.1%			-0.8%	
D→E	**P.**, **TP.**	3.7%	7.3%	6.8%	9.4%	22.7%	5.1%	6.6%
B→D	a1 l a2	6.3%	5.7%	1.8%			5.4%	
C→E	a1 **l**$_{2}$- **a2**$_{1}$-, **P.**	1.3%	8.5%	18.3%			11.6%	

Appendix E

Table Summarizing the Combined Statistic Results for Duration, Pitch and Loudness Features

Table E.1 Combined results for duration, pitch and loudness.

	$\bar{\mu}/\sigma$ **Duration**	$\bar{\mu}/\sigma$ **Pitch**	$\bar{\mu}/\sigma$ **Loudness**
bhala (FL01)			
A→B	**bh**$_{1+}$ bha l a l a2	**a1**$_{1-}$ **l**$_{2-}$ a2, **P**$_{-}$	bh **bha1**$_{1+}$ a l l a2
A→C	**bh**$_{3+}$ bha l **a1**$_{2+}$ **l**$_{1+}$ a2, **D**$_{+}$	a l **l**$_{1-}$ **a2**$_{2-}$, **P**$_{-}$	**bh**$_{3+}$ bha l a l **l**$_{2-}$ **a2**$_{1-}$
A→D	**bh**$_{3+}$ bha l **a1**$_{1+}$ **l**$_{4+}$ **a2**$_{2-}$, **D**$_{+}$	a l **l**$_{2-}$ **a2**$_{1-}$, **P**$_{-}$	bh **bha1**$_{3+}$ a l **l**$_{2-}$ **a2**$_{1-}$
A→E	**bh**$_{2+}$ bha l **a1**$_{1+}$ l **a2**$_{3-}$, **D**$_{+}$	a l **l**$_{2-}$ **a2**$_{1-}$, **P**$_{-}$	**bh**$_{4+}$ bha l **a1**$_{3+}$ **l**$_{2-}$ **a2**$_{1-}$
B→C	bh bha l **a1**$_{1+}$ **l**$_{2+}$ a2, **D**$_{+}$	**a1**$_{3+}$ **l**$_{1-}$ **a2**$_{2-}$	bh bha l a l **l**$_{2-}$ **a2**$_{1-}$
B→D	bh bha l **a1**$_{1+}$ **l**$_{3+}$ **a2**$_{2-}$, **D**$_{+}$	**a1**$_{1+}$ **l**$_{3-}$ **a2**$_{2-}$, **P**$_{-}$	**bh**$_{4-}$ bha l **a1**$_{3+}$ **l**$_{2-}$ **a2**$_{1-}$
C→E	bh bha l **a1**$_{2+}$ l **a2**$_{1-}$, **D**$_{+}$	a l l **a2**$_{1-}$, **P**$_{-}$	bh bha l **a1**$_{1+}$ l a2
D→E	bh bha l a l l **a2**$_{1+}$ **n**$_{2+}$ i, **D**$_{+}$, **TD**$_{+}$	a l l a2 n **i**$_{1-}$	**bh**$_{2+}$ bha l a l l a2 n **i**$_{1-}$
dlala (FL02)			
A→B	dl dla l a l l a2, **D**$_{-}$	**a1**$_{1-}$ **l**$_{2-}$ a2, **P**$_{-}$	**dl**$_{1+}$ dla l **a1**$_{2-}$ l a2
A→C	dl dla l **a1**$_{2+}$ **l**$_{1+}$ a2	**a1**$_{3-}$ **l**$_{1-}$ **a2**$_{2-}$, **P**$_{-}$	**dl**$_{4+}$ dla l **a1**$_{3-}$ **l**$_{2-}$ **a2**$_{1-}$
A→D	dl **dla1**$_{3+}$ **a1**$_{1+}$ l **a2**$_{2-}$, **D**$_{+}$	a l **l**$_{2-}$ **a2**$_{1-}$, **P**$_{-}$	dl dla l a l **l**$_{2-}$ **a2**$_{1-}$
A→E	dl dla l **a1**$_{1+}$ **l**$_{2+}$ a2, **D**$_{+}$	a l **l**$_{2-}$ **a2**$_{1-}$, **P**$_{-}$	dl dla l a l **l**$_{2-}$ **a2**$_{1-}$
B→C	dl dla l **a1**$_{2+}$ **l**$_{1+}$ a2, **D**$_{+}$	**a1**$_{3+}$ **l**$_{1-}$ **a2**$_{2-}$	dl dla l a l **l**$_{1-}$ **a2**$_{2-}$
B→D	**dl**$_{3+}$ dla l **a1**$_{1+}$ **l**$_{4+}$ **a2**$_{2-}$, **D**$_{+}$	**a1**$_{3+}$ **l**$_{2-}$ **a2**$_{1-}$, **P**$_{-}$	**dl**$_{3-}$ dla l **a1**$_{4+}$ **l**$_{2-}$ **a2**$_{1-}$
C→E	dl dla l **a1**$_{1+}$ **l**$_{2-}$ a2, **D**$_{+}$	**a1**$_{1+}$ **l**$_{2+}$ a2, **P**$_{-}$	**dl**$_{2-}$ dla l **a1**$_{3+}$ l **a2**$_{1-}$
D→E	dl dla l **a1**$_{3+}$ l **a2**$_{2+}$ **n**$_{1+}$ i, **D**$_{+}$, **TD**$_{+}$	a l l a2 n i, **P**$_{+}$	dl dla l a l l a2 **n**$_{1-}$ **i**$_{2-}$
lala (HL01)			
A→B	l l a l l2 a2, **D**$_{-}$	a l **l2**$_{1-}$ **a2**$_{2-}$, **P**$_{-}$	**l l**$_{1+}$ a l l2 a2
A→C	l l **a1**$_{1+}$ l2 a2, **D**$_{+}$	a l **l2**$_{1-}$ **a2**$_{2-}$	**l l**$_{3+}$ **a1**$_{4-}$ **l2**$_{2-}$ **a2**$_{1-}$
A→D	l l **a1**$_{1+}$ l2 **a2**$_{2-}$, **D**$_{+}$	**a1**$_{2+}$ l2 **a2**$_{1-}$, **P**$_{-}$	**l l**$_{2+}$ a l l2 **a2**$_{1-}$
A→E	l l **a1**$_{1+}$ l2 a2, **D**$_{+}$	a l **l2**$_{2-}$ **a2**$_{1-}$, **P**$_{-}$	**l l**$_{2+}$ a l l2 **a2**$_{1-}$
B→C	l l **a1**$_{1+}$ **l2**$_{2+}$ a2, **D**$_{+}$	a l l2 a2	l l a l l2 **a2**$_{1-}$
B→D	**l l**$_{3+}$ **a1**$_{1+}$ l2 **a2**$_{2-}$, **D**$_{+}$	**a1**$_{1+}$ **l2**$_{2+}$ **a2**$_{3-}$, **P**$_{-}$	l l **a1**$_{2+}$ l2 **a2**$_{1-}$
C→E	l l **a1**$_{2+}$ **l2**$_{1-}$ a2, **D**$_{+}$	**a1**$_{1+}$ **l2**$_{3+}$ **a2**$_{2-}$, **P**$_{-}$	l l a l l2 **a2**$_{1-}$
D→E	l l a l l2 **a2**$_{1+}$ **n**$_{2+}$ **i**$_{3+}$, **D**$_{+}$, **TD**$_{+}$	**a1**$_{4-}$ **l2**$_{3-}$ **a2**$_{2-}$ **n**$_{1-}$ **i**$_{5-}$, **TP**$_{-}$	l l a l l2 a2 **n**$_{2-}$ **i**$_{1-}$

Table E.1 Combined results for duration, pitch and loudness (continued).

hlala (HL02)			
A→B	**hl**$_{1-}$ hla1 a1 l **a2**$_{2-}$, **D**$_{-}$	**a1**$_{3-}$ **l**$_{1-}$ **a2**$_{2-}$, **P**$_{-}$	hl hla1 **a1**$_{1-}$ **l**$_{2-}$ **a2**$_{3-}$
A→C	hl hla1 **a1**$_{1+}$ **l**$_{2+}$ a2, **D**$_{+}$	**a1**$_{3-}$ **l**$_{1-}$ **a2**$_{2-}$, **P**$_{-}$	hl hla1 **a1**$_{3-}$ **l**$_{1-}$ **a2**$_{2-}$
A→D	hl hla1 **a1**$_{1+}$ l **a2**$_{2-}$, **D**$_{+}$	a1 **l**$_{2-}$ **a2**$_{1-}$, **P**$_{-}$	hl hla1 a1 **l**$_{2-}$ **a2**$_{1-}$
A→E	hl **hla1**$_{2+}$ **a1**$_{1+}$ **l**$_{4+}$ **a2**$_{3-}$, **D**$_{+}$	a1 **l**$_{2-}$ **a2**$_{1-}$, **P**$_{-}$	**hl**$_{3+}$ **hla1**$_{4+}$ a1 **l**$_{2-}$ **a2**$_{1-}$
B→C	hl hla1 **a1**$_{1+}$ **l**$_{2+}$ **a2**$_{3+}$, **D**$_{+}$	a1 l a2	hl **hla1**$_{2+}$ a1 l **a2**$_{1-}$
B→D	**hl**$_{3+}$ hla1 **a1**$_{1+}$ l **a2**$_{2-}$, **D**$_{+}$	**a1**$_{1+}$ **l**$_{2+}$ a2, **P**$_{-}$	hl **hla1**$_{3+}$ **a1**$_{2+}$ l **a2**$_{1-}$
C→E	**hl**$_{2+}$ hla1 a1 **l**$_{1-}$ a2, **D**$_{+}$	**a1**$_{1+}$ **l**$_{2+}$ a2	hl hla1 **a1**$_{2+}$ l **a2**$_{1-}$
D→E	**hl**$_{4+}$ hla1 **a1**$_{5+}$ **l**$_{3+}$ **a2**$_{2+}$ **n**$_{1+}$ **i**$_{6+}$, **D**$_{+}$, **TD**$_{+}$	a1 l a2 n i	**hl**$_{2+}$ hla1 a1 **l**$_{3-}$ a2 n **i**$_{1-}$
bala (LL01)			
A→B	*b b bal a1 l **a2**$_{1-}$, **D**$_{-}$	a1 **l**$_{2-}$ **a2**$_{1-}$	***b**$_{4+}$ **b**$_{5-}$ **bal**$_{2+}$ a1 **l**$_{1-}$ **a2**$_{3-}$
A→C	***b**$_{3+}$ b bal **a1**$_{1+}$ **l**$_{2+}$ a2, **D**$_{+}$	a1 **l**$_{2-}$ **a2**$_{1-}$	*b b **bal**$_{4+}$ **a1**$_{1-}$ **l**$_{3-}$ **a2**$_{2-}$
A→D	*b b bal **a1**$_{1+}$ l **a2**$_{2-}$, **D**$_{+}$	a1 **l**$_{2-}$ **a2**$_{1-}$, **P**$_{-}$	***b**$_{4+}$ b **bal**$_{2+}$ a1 **l**$_{3-}$ **a2**$_{1-}$
A→E	*b b bal **a1**$_{1+}$ **l**$_{2+}$ a2, **D**$_{+}$	a1 **l**$_{2-}$ **a2**$_{1-}$, **P**$_{-}$	***b**$_{4+}$ b **bal**$_{2+}$ a1 **l**$_{3-}$ **a2**$_{1-}$
B→C	*b b bal **a1**$_{1+}$ **l**$_{2+}$ **a2**$_{3+}$, **D**$_{+}$	a1 l a2	*b b bal a1 l **a2**$_{1-}$
B→D	*b b bal **a1**$_{1+}$ l **a2**$_{2-}$, **D**$_{+}$	a1 l a2, **P**$_{-}$	*b **b**$_{3+}$ bal **a1**$_{2+}$ l **a2**$_{1-}$
C→E	*b b bal **a1**$_{1+}$ l **a2**$_{2-}$, **D**$_{+}$	**a1**$_{3-}$ **l**$_{2-}$ **a2**$_{1-}$, **P**$_{-}$	***b**$_{3+}$ b bal **a1**$_{2+}$ l a2$_{1-}$
D→E	*b b bal a1 **l**$_{4+}$ **a2**$_{1+}$ **n**$_{2+}$ **i**$_{3+}$, **D**$_{+}$, **TD**$_{+}$	a1 **l**$_{2-}$ **a2**$_{1-}$ **n**$_{3-}$ **i**$_{4-}$, **P**$_{-}$, **TP**$_{-}$	***b**$_{3+}$ b bal a1 l **a2**$_{1-}$ n **i**$_{2-}$
tsala (LL02)			
A→B	ts tsal a1 **l**$_{2+}$ **a2**$_{1-}$, **D**$_{-}$	a1 **l**$_{2-}$ **a2**$_{1-}$	**ts**$_{2-}$ tsal a1 **l**$_{1-}$ **a2**$_{3-}$
A→C	**ts**$_{3+}$ tsal **a1**$_{1+}$ **l**$_{2+}$ a2, **D**$_{+}$	a1 **l**$_{2-}$ **a2**$_{1-}$	ts tsal a1 **l**$_{2-}$ **a2**$_{1-}$
A→D	ts **tsal**$_{3+}$ **a1**$_{1+}$ l **a2**$_{2-}$, **D**$_{+}$	a1 **l**$_{2-}$ **a2**$_{1-}$	**ts**$_{3-}$ tsal a1 **l**$_{2-}$ **a2**$_{1-}$
A→E	**ts**$_{4+}$ tsal **a1**$_{1+}$ **l**$_{2+}$ **a2**$_{3-}$, **D**$_{+}$	a1 **l**$_{2-}$ **a2**$_{1-}$, **P**$_{-}$	ts **tsal**$_{3+}$ a1 **l**$_{2-}$ **a2**$_{1-}$
B→C	**ts**$_{4+}$ tsal **a1**$_{1+}$ **l**$_{2+}$ **a2**$_{3+}$, **D**$_{+}$	a1 l **a2**$_{1+}$	**ts**$_{1+}$ tsal a1 l a2
B→D	ts **tsal**$_{3+}$ **a1**$_{1+}$ l **a2**$_{2-}$, **D**$_{+}$	a1 l **a2**$_{1+}$	ts tsal a1 l **a2**$_{1-}$
C→E	ts tsal **a1**$_{2+}$ **l**$_{1-}$ a2, **D**$_{+}$	a1 **l**$_{2-}$ **a2**$_{1-}$, **P**$_{-}$	ts **tsal**$_{2+}$ a1 l **a2**$_{1-}$
D→E	ts tsal **a1**$_{3+}$ l **a2**$_{1+}$ **n**$_{2+}$ i, **D**$_{+}$, **TD**$_{+}$	**a1**$_{4-}$ **l**$_{5-}$ **a2**$_{3-}$ **n**$_{2-}$ **i**$_{1-}$, **P**$_{-}$, **TP**$_{-}$	ts tsal a1 l **a2**$_{1+}$ n i

Appendix F

Percentage Changes for the Generation of Commands in Different Contexts from the Infinitive

Table F.1 Rounded percentage duration changes for the generation of commands from the infinitive.

Percentage duration changes (rounded)				
	Order of significance	**Δd_{a1}**	**Δd_{l}**	**Δd_{a2}**
A→B	a1 l a2	11% [1]	11% [1]	11% [1]
A→C	**$a1_{1+}$ l_{2+}** a2	-44%	-25%	0
A→D	**$a1_{1+}$ l_{3+} $a2_{2-}$**	-54%	-8%	31% [2]
A→E	**$a1_{1+}$ l_{3+} $a2_{2-}$**	-59%	-15%	39%

[1] $\Delta d_{\tilde{B}:a1} = \Delta d_{\tilde{B}:l} = \Delta d_{\tilde{B}:a2} = \Delta d_{\tilde{B}:ala} = 11\%$

[2] Not consistent with statistics:

$$\left|\Delta d_{\tilde{D}:a2}\right| = \left|\Delta d_{\tilde{D}:l}\right| + \left(\left|\Delta d_{\tilde{D}:a1}\right| - \left|\Delta d_{\tilde{D}:l}\right|\right)/2 = 8 + (54-8)/2 = 31\%$$

Table F.2 Rounded percentage pitch changes for the generation of commands from the infinitive.

Percentage pitch changes for FL (rounded)				
	Order of significance	**Δp_{a1}**	**Δp_{l}**	**Δp_{a2}**
A→B	**$a1_{1-}$ l_{2-}** a2, **$P_{.}$**	29%	24%	0
A→C	a1 **l_{1-} $a2_{2-}$**, **$P_{.}$**	0	55%	46%
A→D	a1 **l_{2-} $a2_{1-}$**, **$P_{.}$**	0	46%	70%
A→E	a1 **l_{2-} $a2_{1-}$**, **$P_{.}$**	0	42%	72%
Percentage pitch changes for HL				
	Order of significance	**Δp_{a1}**	**Δp_{l}**	**Δp_{a2}**
A→B	a1 **l_{1-} $a2_{2-}$**, **$P_{.}$**	0	43%	37%
A→C	a1 **l_{1-} $a2_{2-}$**	0	46%	51%
A→D	a1 **l_{2-} $a2_{1-}$**, **$P_{.}$**	0	27%	58%
A→E	a1 **l_{2-} $a2_{1-}$**, **$P_{.}$**	0	26%	56%
Percentage pitch changes for LL				
	Order of significance	**Δp_{a1}**	**Δp_{l}**	**Δp_{a2}**
A→B	a1 **l_{2-},$a2_{1-}$**	0	47%	94%
A→C	a1 **l_{2-},$a2_{1-}$**	0	53%	78%
A→D	a1 **l_{2-},$a2_{1-}$**	0	55%	97%
A→E	a1 **l_{2-},$a2_{1-}$**	0	67%	111%

Appendix G

Results of Perceptual Experiment

Graphs showing response percentages

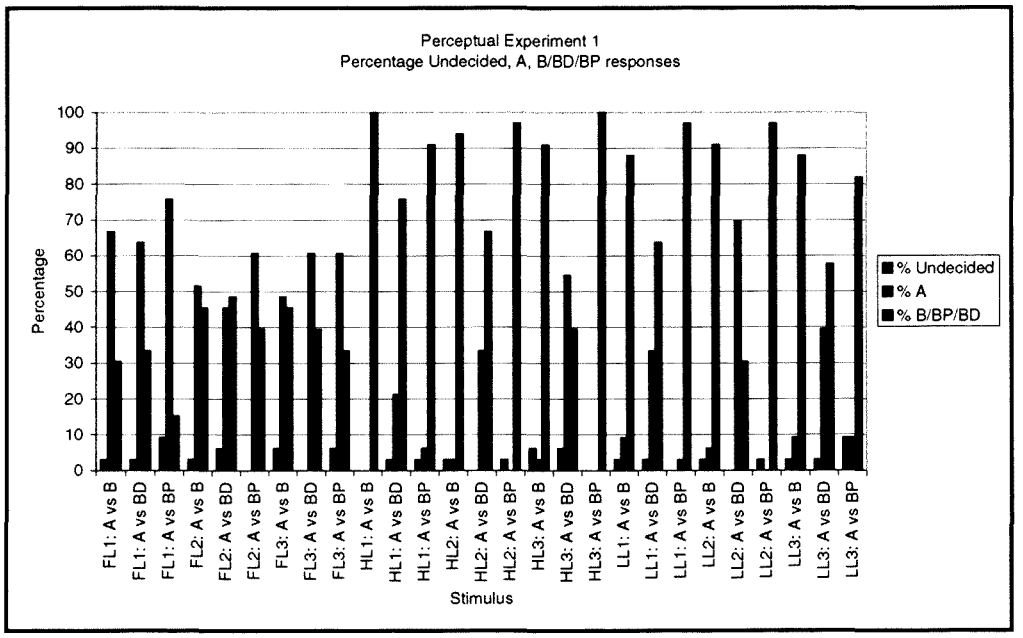


Figure G.1 Graph of the percentage Undecided, $\tilde{A}$ and $\tilde{B}$ */BD/BP* responses for a given set of stimuli.

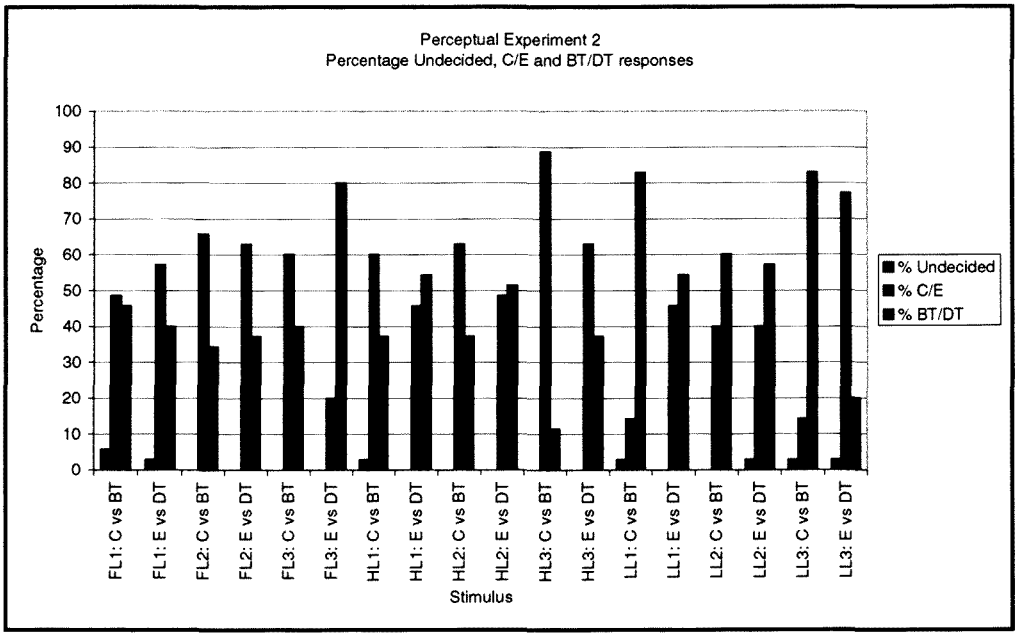


Figure G.2 Graph of the percentage Undecided, $\tilde{C}$, $\tilde{E}$ and *BT/DT* responses for a given set of stimuli.

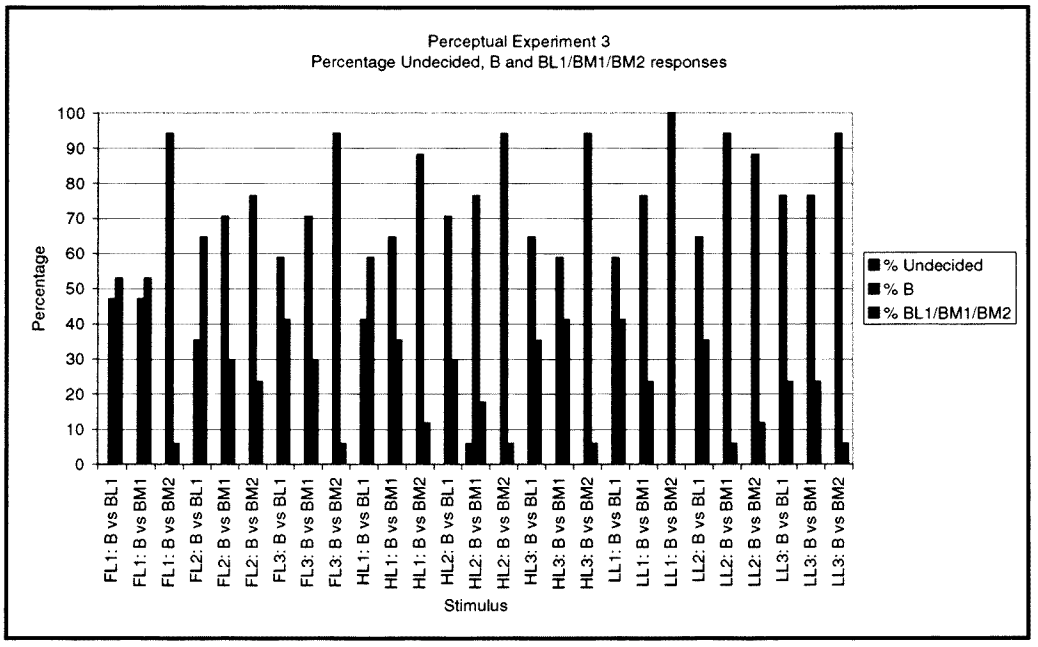


Figure G.3 Graph of the percentage Undecided, $\tilde{B}$ and *BL1/BM1/BM2* responses for a given set of stimuli.

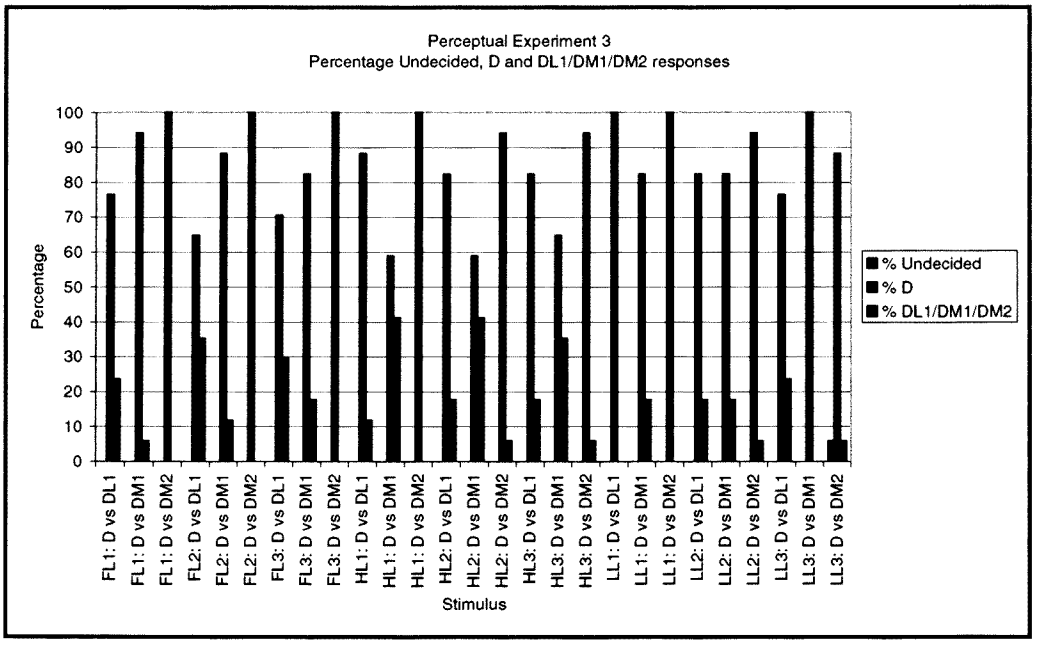


Figure G.4 Graph of the percentage Undecided, $\tilde{D}$ and *DL1/DM1/DM2* responses for a given set of stimuli.

Graphs showing results for mean response time

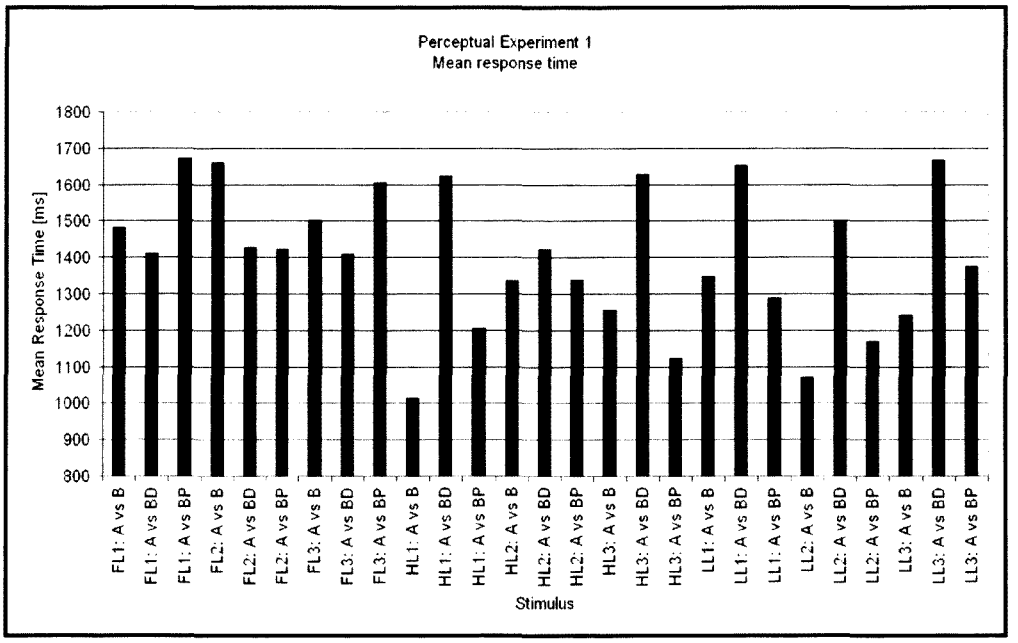


Figure G.5 Mean response time for Perception Test 1.

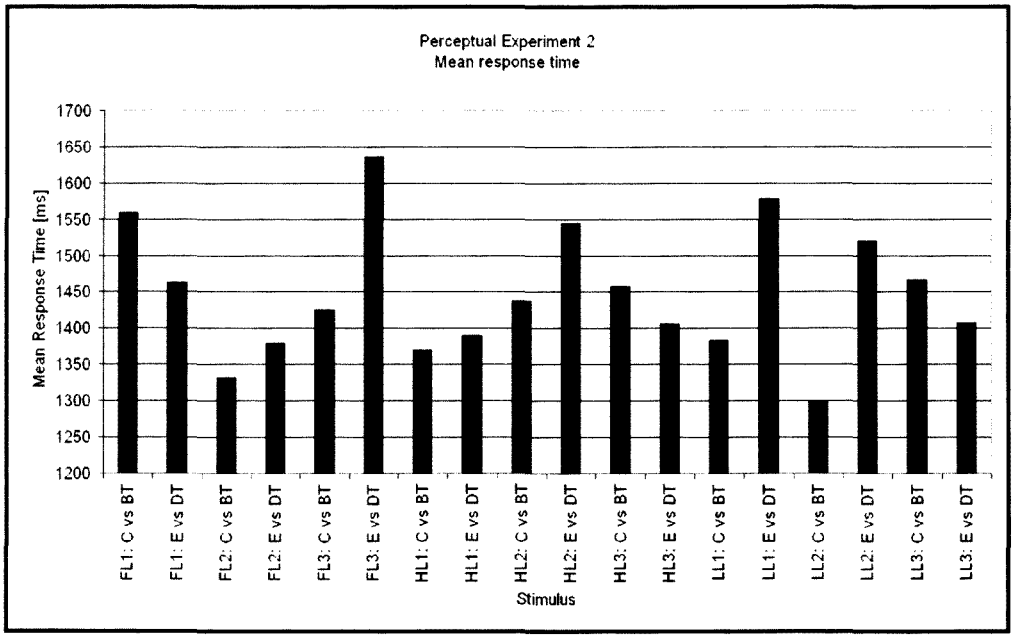


Figure G.6 Mean response time for Perception Test 2.

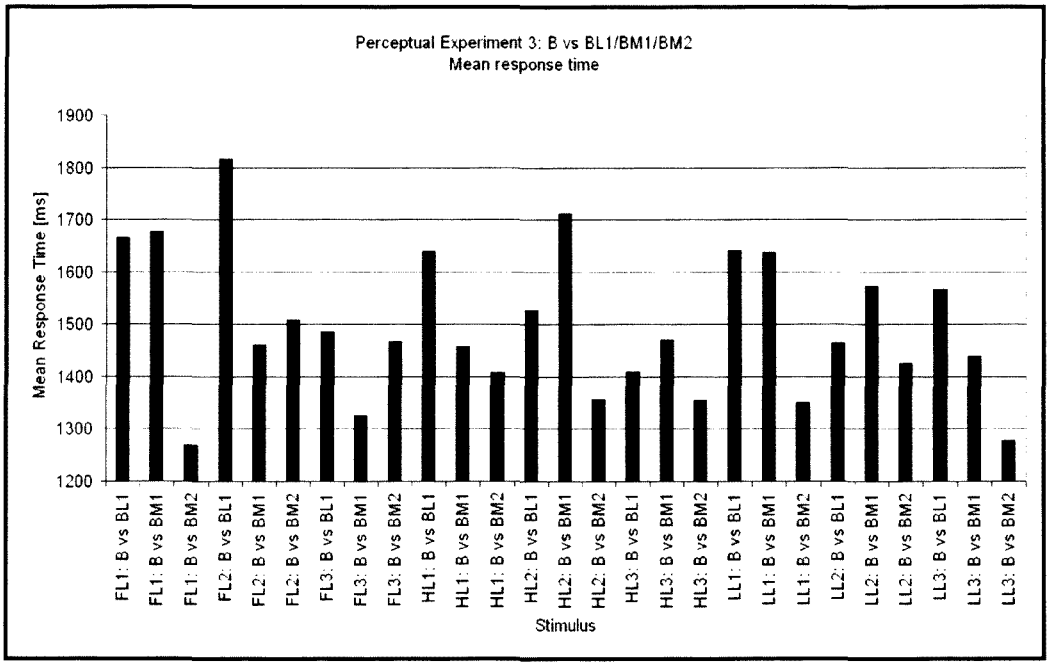


Figure G.7 Mean response time for Perception Test 3: *B* vs. *BL1/BM1/BM2*.

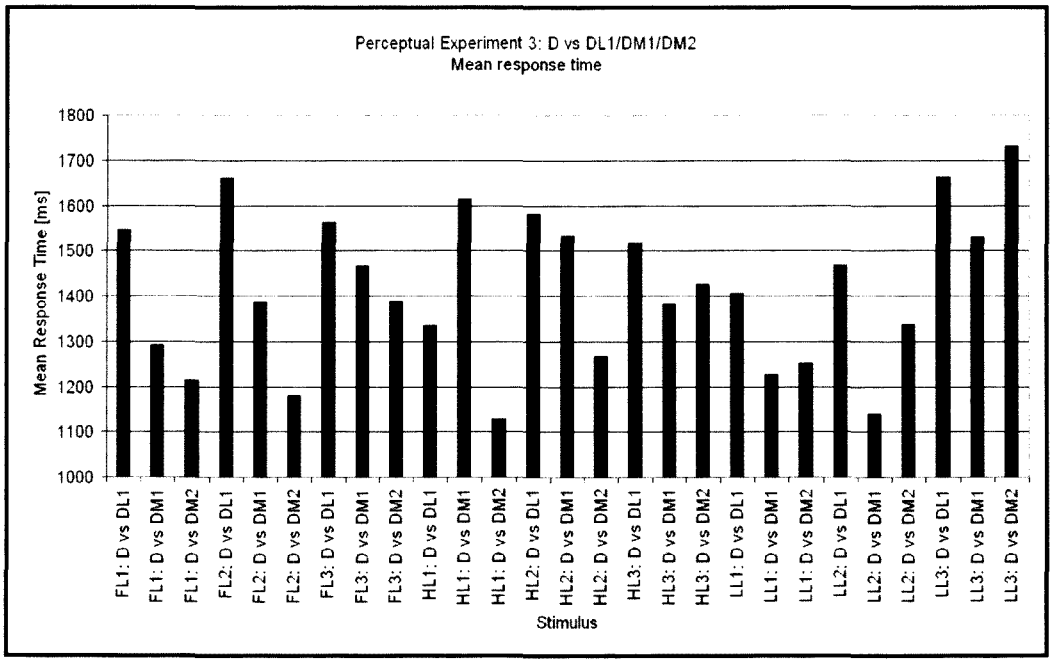


Figure G.8 Mean response time for Perception Test 3: *D* vs. *DL1/DM1/DM2*.

References

Allen, J. 1991. Overview of text-to-speech systems. In Sadaoki, F., Sondhi, M.M. (ed.) *Advances in speech signal processing*. New York: Marcel Dekker, Inc.

Allen, J., Hunnicut, S., Klatt, D. 1987. *From text to speech, The MITALK system*. Cambridge: Cambridge University Press.

Bell Labs (Ed.). (1997, April 21) Bell Labs Text-to-Speech Synthesis. [Homepage of Bell Laboratories], [Online]. Available: http://www.bell-labs.com/project/tts/tts-overview.html [1999, December 10].

Bloothooft, G., Hazan, V., Llisterri, J. & Huber, D. (ed.). 1995. European studies in phonetics and speech communication. Utrecht: OTS Publications. [Homepage of Essex University Speech Group], [Online]. Available: http://www.essex.ac.uk/eras-speech/syllabus/synthesis.html [1999, December 10].

Carlson, R., Granström, B., & Hunnicut, S. 1982. A multi-language text-to-speech module, ICASSP, 3:1604-1607.

Claughton, J.S. 1983. The tones of Xhosa inflections. In Fivaz, D. (ed.) *Communications.* 13

Cruttenden, A. 1986. *Intonation*. Cambridge: Cambridge University Press.

Crystal, D. (ed.) 1997. A Dictionary of linguistics and phonetics. 4th edn. Oxford: Blackwell Publishers Ltd.

D'Alessandro, C. & Liénard, J. 1997. Synthetic speech generation. In Cole, R.A., Mariani, J., Uszkoreit, H., Zaenen, A. & Zue, V. (eds.) *Survey of the state of the art in human language technology*. Cambridge: Cambridge University Press, Also Available: http://cslu.cse.ogi.edu/HLTsurvey/ch5node4.html#SECTION52 [1999, December 10].

Davey, A. 1973. *The moods and tenses of the verb in Xhosa*. M.A. dissertation. Pretoria: University of South Africa.

De Wet, Febe. 1999. *Isolated word speech recognition in Xhosa*. M.Eng. dissertation. Pretoria:University of Pretoria.

De Wet, Febe & Botha, Elizabeth C. 1998. Indigenous speech technology - A speech recognition front-end for computer assisted Xhosa language learning. *South African Journal of Linguistics*. Supplement 36:137-139).

Dirksen, A. et al. 1995. Representation and phonetic realization of prosodic structure. In Collier, R., Dirksen, A. & De Pijper, J.R. (eds.). (1995) Institute for Perception Research : Annual Research Overview [Homepage of Institute for Perception Research], [Online]. Available: http://www.tue.nl/ipo/hearing/aro95/terken1.htm#terken1000 [1999, December 10].

Du Plessis, J.A. 1978. isiXhosa 4. Goodwood: Oudiovista.

Du Plessis, J.A & Visser, M. 1992. Xhosa syntax. Pretoria: Via Afrika.

Dutoit, T. & Leich, H. 1993. MBR-PSOLA: Text-to-speech synthesis based on an MBE re-synthesis of the segments database. *Speech Communication*, 13:435-440. Also Available: http://tcts.fpms.ac.be/publications.html [1999, December 10].

Dutoit, T. (1997a, April 25-Last update). A short introduction to text-to-speech synthesis. [Homepage of The Circuit Theory and Signal Processing Lab (TCTS Lab) of the Faculté Polytechnique de Mons], [Online]. Available: http://tcts.fpms.ac.be/synthesis/introtts.html [1999, December 10].

Dutoit, T. (1997b, April 25-Last update). New MBROLA databases. [Homepage of The Circuit Theory and Signal Processing Lab (TCTS Lab) of the Faculté Polytechnique de Mons], [Online]. Available: http://tcts.fpms.ac.be/synthesis/mbrjoin.html [1999, December 10.

Dutoit, T. 1993. High quality text-to-speech synthesis of the French language. Ph.D.dissertation: Faculté Polytechnique de Mons, Belgium. [Homepage of The Circuit Theory and Signal Processing Lab (TCTS Lab) of the Faculté Polytechnique de Mons],

[Online]. Available: http://tcts.fpms.ac.be/publications/phds/dutoit/tdphd.html [1999, December 10].

Fallside, F. 1994. Speech synthesis. In Asher, R.E. (ed.) *The encyclopedia of language and linguistics,* 8:4264-4272. London: Pergamon.

Hertz, S. (1999, January 12-Last update), The technology of text-to-speech. [Homepage of Eloquent Technologies, Inc.], [Online]. Available: http://www.eloq.com/SuePap.htm [1999, December 10].

Hirschman, L. & Thompson, H.T. 1997. Overview of evaluation in speech and natural language processing. In Cole, R.A., Mariani, J., Uszkoreit, H., Zaenen, A. & Zue, V. (eds.) *Survey of the state of the art in human language technology.* Cambridge: Cambridge University Press. Also Available: http://cslu.cse.ogi.edu/HLTsurvey/ch13node3.html#SECTION131 [1999, December 10].

Holmes, J., Mattingly, I. & Shearme, J. 1964. Speech synthesis by rule. In *Language and Speech*, 7:127-143.

House, D. 1990. *Tonal perception in speech.* Sweden: Lund University Press.

Institute for Perception Research (IPO): Annual Research Overview (ARO) 1995, Available: http://www.tue.nl/ipo/hearing/aro95/pros.htm. [1999, December 10].

Johns-Lewis, C. (ed.) 1986. *Intonation in discourse.* London: Croom Helm Ltd.

Jones, J., Louw, J. & Roux, J.C. 1998a. Queclaratives in Xhosa: An acoustic analysis. *South African Journal of Linguistics*, Supplement 36:3-18.

Jones, J., Louw, J. & Roux, J.C. 1998b. Perceptual experiments on queclaratives in Xhosa. *South African Journal of Linguistics*, Supplement 36:19-32.

Kamm, Candace, Walker, Marilyn & Rabiner, L. 1997. The role of speech processing in human-computer intelligent communication. *Speech Communication,* 23:263-278.

Kent, R.D. & Read, C.R. 1992. *The acoustic analysis of speech.* San Diego: Singular Publishing Group,Inc.

Klatt, D.H. 1980. Software for a cascade/parallel formant synthesizer, *Journal of the Acoustic Society of America*, 67:971-995.

Klatt, D.H. 1987. Review of text-to-speech conversion for English. In *Journal of the Acoustic society of America*, 82 (3):773-793.

Kocatkaert, H. & Godwin, D. 1996. Voicing status of syllable-initial plosives in siSwati. *South African Journal of African Languages*, 17 (3): 100-104.

Kreyszig, E. 1988. *Advanced Engineering Mathematics.* Sixth Ed. New York: Wiley.

Lange, R.H. 1993. Speech synthesis and speech recognition: Tomorrow's human-computer interfaces? In Williams, M.E. (ed.) Annual Review of Information Science and Technology (ARIST), 28. Medford, N.J.: Learned Information, Inc. [Homepage of American Society for Information Science], [Online]. Available: http://www.asis.org/Publications/ARIST/arist-93/arist-93-section-2/arist-93-chapter4.html [1999, December 10].

Laureate. (1996, January 11- last modified). About Laureate. [Homepage of BT Laboratories], [Online]. Was Available: http://www.labs.bt.com/innovate/speech/laureate/about.htm [1997, June 27]. Now Available: http://innovate.bt.com/showcase/Laureate/whitepaper.htm [1999, December 10].

Lehiste, I. 1970. *Suprasegmentals.* Cambridge: The M.I.T. Press.

Lieberman, P. & Blumstein, S.E. 1988. *Speech physiology, speech perception, and acoustic phonetics.* Cambridge: Cambridge University Press.

Maddieson, I. 1978. Universals of tone. In Greenberg, J.H. (ed.) *Universals of human language 2*. Stanford: Stanford University Press.

Moulines, E. & Charpentier, F. 1990. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9: 453 – 467.

Ostendorf, M. (nd.) Evaluating the use of prosodic information in speech recognition and understanding, [Online]. Available: http://www.cs.tufts.edu/~jacob/isgw/Ostendorf.htm [1998, February 6].

Pinnock, Patricia.S. 1994. *Xhosa: A cultural grammar for beginners*. Cape Town: African Sun Press.

Pols, L.C.W. 1991 Quality assessment of text-to-speech synthesis by rule. In Sadaoki, F., Sondhi, M.M. (ed.) *Advances in speech signal processing*. New York: Marcel Dekker, Inc.

Pols, L.C.W. 1994. Speech technology systems: Performance and evaluation. In Asher, R.E. (ed.) *The encyclopedia of language and linguistics*, 8:4289-4296. London: Pergamon.

Pols, L.C.W. 1997. Speech synthesis evaluation. In Cole, R.A., Mariani, J., Uszkoreit, H., Zaenen, A. & Zue, V. (eds.) *Survey of the state of the art in human language technology*. Cambridge: Cambridge University Press, http://cslu.cse.ogi.edu/HLTsurvey/ch13node9.html#SECTION137 [1999, December 10].

Price, P. 1997. Spoken language understanding. In Cole, R.A., Mariani, J., Uszkoreit, H., Zaenen, A. & Zue, V. (eds.) *Survey of the state of the art in human language technology*. Cambridge: Cambridge University Press. Also Available: http://cslu.cse.ogi.edu/HLTsurvey/ch1node10.html#SECTION18 [1999, December 10].

Riordan, J. 1969. *Lumko Xhosa self instruction course*. Lady Frere, South Africa: Lumko Institute.

Roux, J.C. 1989. Grapheme-to-phoneme conversions in Xhosa. *South African Journal of African Languages*, 9 (2): 74-78.

Roux, J.C. 1995a. Prosodic data and phonological analyses in Zulu and Xhosa. *South African Journal of African Languages*, 15 (1): 19-28.

Roux, J.C. 1995b. On the perception and production of tone in Xhosa. *South African Journal of African Languages*, 15 (4): 196-204.

Roux, J.C. 1998. SASPEECH: Establishing speech resources for the indigenous languages of South Africa. *Proceedings of the 1st international conference on language resources and evaluation*, Granada, Spain, 351-355. ELRA.

Sagisaka, Y. 1997. Spoken language output technologies, Overview. In Cole, R.A., Mariani, J., Uszkoreit, H., Zaenen, A. & Zue, V. (eds.) *Survey of the state of the art in human language technology*. Cambridge: Cambridge University Press, Also Available: http://cslu.cse.ogi.edu/HLTsurvey/ch5node3.html#SECTION51 [1999, December 10].

Salza, P.L., Foti,E., Nebbia, L. & Oreglia, M. 1996. MOS and pair comparison combined methods for quality evaluation of text-to-speech systems. *Acustica - Acta acustica*, 82: 650-656.

Sanfilippo, A. 1997. Lexicons for constraint-based grammers. In Cole, R.A., Mariani, J., Uszkoreit, H., Zaenen, A. & Zue, V. (eds.) *Survey of the state of the art in human language technology*. Cambridge: Cambridge University Press. Also Available: http://cslu.cse.ogi.edu/HLTsurvey/ch3node6.html#SECTION34 [1999, December 10]

Solomon, R.C. 1996. *Statistics*. London: John Murray.

SPI Lab (Ed.).(1999, March 15-last update). Computational modeling of intonation for synthesis and recognition. [Homepage of The Signal Processing and Interpretation Lab], [Online]. Available: http://raven.bu.edu/projects/inton.htm [1998, February 11].

Sproat, R. 1997. Text interpretation for TTS synthesis. In Cole, R.A., Mariani, J., Uszkoreit, H., Zaenen, A. & Zue, V. (eds.) *Survey of the state of the art in human language technology*. Cambridge: Cambridge University Press. Also Available: http://cslu.cse.ogi.edu/HLTsurvey/ch5node5.html#SECTION53 [1999, December 10].

Theron, P.Z. 1999. *Automatic acquisition of two-level morphological rules*. D.Phil. dissertation. Stellenbosch: University of Stellenbosch.

Van Bezooijen, R. & Van Heuven, V. 1997a. What are speech output systems? In Gibbon, D., Moore, R. & Winski, R. (ed.). *Handbook of standards and resources for spoken language systems*. Berlin & New York: Walter de gruyter Publishers, [Homepage of Expert Advisory Groups on Language Engineering Standards Spoken Language Working Group (EAGLES)], [Online]. Was Available: http://www.degruyter.de/EAGLES/degruyt/node 385.htm [1998, February 6].

Van Bezooijen, R. & Van Heuven, V. 1997b. Text-to-speech synthesis. In Gibbon, D., Moore, R. & Winski, R. (ed.). *Handbook of standards and resources for spoken language systems*. Berlin & New York: Walter de gruyter Publishers, [Homepage of Expert Advisory Groups on Language Engineering Standards Spoken Language Working Group (EAGLES)], [Online]. Was Available: http://www.degruyter.de/EAGLES/degruyt/node 69.htm [1998, February 6].

Van Bezooijen, R. & Van Heuven, V. 1997c. Linguistic aspects. In Gibbon, D., Moore, R. & Winski, R. (ed.). *Handbook of standards and resources for spoken language systems*. Berlin & New York: Walter de gruyter Publishers, [Homepage of Expert Advisory Groups on Language Engineering Standards Spoken Language Working Group (EAGLES)], [Online]. Was Available: http://www.degruyter.de/EAGLES/degruyt/node 405.htm [1998, February 6].

Van Bezooijen, R. & Van Heuven, V. 1997d. Functions of prosody. In Gibbon, D., Moore, R. & Winski, R. (ed.). *Handbook of standards and resources for spoken language systems*. Berlin & New York: Walter de gruyter Publishers, [Homepage of Expert Advisory Groups on Language Engineering Standards Spoken Language Working Group (EAGLES)], [Online]. Was Available: http://www.degruyter.de/EAGLES/degruyt/node 406.htm [1998, February 6].

Wentzel, P.J., Botha, J.J. & Mzileni, P.M. 1972. Xhosa taalboek. Johannesburg: Perskor-uitgewery.

Westphal, E.O.J., Notshweleka, M. & Tindleni, S.M. 1967. Tonal profiles of Xhosa nominals and verbo-nominals. *Communications form the School of African Studies*, no.32. Cape Town: University of Cape Town.

Witten, I.H. 1986. Making computers talk. New Jersey: Prentce-Hall.

Zue, V. & Cole, R. 1997 Overview. In Cole, R.A., Mariani, J., Uszkoreit, H., Zaenen, A. & Zue, V. (eds.) *Survey of the state of the art in human language technology*. Cambridge: Cambridge University Press. Also Available: http://cslu.cse.ogi.edu/HLTsurvey/ch1node3.html#SECTION11 [1999: December 10].