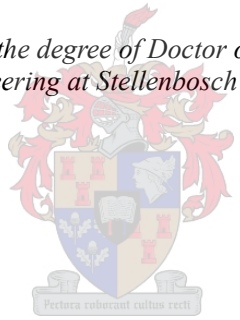
# PROSODY MODELLING FOR A SESOTHO TEXT-TO-SPEECH SYSTEM USING THE FUJISAKI MODEL

by

LEHLOHONOLO MOHASI

*Dissertation presented for the degree of Doctor of Philosophy in the Faculty of Engineering at Stellenbosch University*




Supervisors:

Prof. Thomas Niesler  Prof. Dr. Hansjörg Mixdorff

March 2015

# Declaration

By submitting this dissertation electronically, I declare that the entirety of the work contained therein is my own original work, that I am the owner of the copyright thereof (unless to the extent explicitly otherwise stated) and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: ….....

# Abstract

Sesotho, a Southern Bantu languages spoken nationally in Lesotho and as one of the official languages in South Africa, uses tone to convey grammatical intonation. Accurate prosodic modelling is crucial for the production of intelligible and natural-sounding speech in text-to-speech (TTS) systems, and can be achieved by the correct modelling of pitch in tonal languages. However, the interrelationship between tone and general intonation in Sesotho is currently poorly understood and technologically not addressed. This thesis considers prosodic modelling for Sesotho using the Fujisaki model with a view to the development of a text-to-speech system.

Fujisaki analysis can be used to analyse the tone associated with a syllable using only the acoustics. However, it is often at odds with the tone derived by surface tonal transcription. Preliminary experiments were performed in which minimal pairs in Sesotho were examined. Tonal alignments and magnitudes of the fundamental frequency (F0) excursions were analysed using the Fujisaki model. These experiments established that high surface tones in Sesotho are associated with Fujisaki tone commands of positive polarity while low tones are associated with the absence of a tone command. It was found that adjacent high surface tone syllables and adjacent syllables of alternating surface tone form different types of prosodic groups, which are Fujisaki tone commands spanning multiple syllables.

In subsequent experiments, the duration and amplitude of the Fujisaki tone commands matching high surface tones were modified. In cases where the tones mismatched, remedial modifications were performed by either setting the tone command amplitude to zero for low surface tone syllables, or inserting a tone command for high surface tone syllables. Perceptual evaluations indicated that, for high-tone prosodic groups, best results are obtained when simply aligning Fujisaki tone commands with syllable boundaries, and when using conservative tone command amplitudes. The perceived quality of the resynthesised speech was high, although it was a little lower on average for cases in which remedial modifications were necessary. Furthermore, it was found that the discrepancies between surface and Fujisaki tones were frequently due to an error in the dictionary providing the lexical tones, or due to errors and omissions in the current surface tonal transcription algorithm.

In further experiments, the Fujisaki parameters were determined not by heuristics, but by a machine learning algorithm. Three regression trees, one for each Fujisaki tone command parameter, were trained using sixteen syllable attributes as input to develop a data-driven prosodic model for the Sesotho language. In comparing the speech resynthesised using this model with that from resynthesis using simpler heuristics, it was found to offer improved performance. Speech resynthesised in this way was judged to have an "almost mother tongue" naturalness, and demonstrates the potential for synthesising Sesotho speech with natural prosody from tone-marked text using data-driven techniques.

Finally, the prosody generation capability of the Fujisaki model and that commonly used by HMM-based speech synthesis (HTS) systems were compared for the Sesotho and Serbian languages. Results showed that the Fujisaki model has a better F0 modelling capability for Sesotho whereas HTS showed a higher performance for Serbian.

# Opsomming

Sesotho, 'n Suid-Bantoetaal wat nasionaal in Lesotho gepraat word en ook een van Suid-Afrika se amptelike tale is, maak van toonhoogte gebruik om grammatiese intonasie oor te dra. Akkurate prosodiese modellering is noodsaaklik om verstaanbare teks-na-spraakstelsels te ontwikkel wat ook natuurlik klink, en dit kan bereik word deur die toonhoogte van tonale tale korrek te modelleer. Die interverwantskap tussen toonhoogte en gewone intonasie in Sesotho is egter nog nie deeglik bestudeer nie, en ook nog nie deur tegnologie aangespreek nie. Hierdie proefskrif beskou prosodiese modellering van Sesotho deur van die Fujisaki-model gebruik te maak, met die oog op die ontwikkeling van 'n teks-na-spraakstelsel.

Fujisaki-analise kan gebruik word om 'n enkele lettergreep se toonhoogte te analiseer, deur slegs die klanke te gebruik. Dit weerspreek egter dikwels die toonhoogte wat met behulp van tonale transskripsie onttrek is. Voorlopige eksperimente is uitgevoer waarin minimum-pare in Sesotho ondersoek is. Die toonooreenstemming en amplitudes van die fundamentele frekwensie (F0)-afwyking is met behulp van die Fujisaki-model geanaliseer. Hierdie eksperimente het aangetoon dat hoë oppervlaktoonhoogtes in Sesotho ooreenstem met Fujisaki-toonbevele van positiewe polariteit, terwyl lae toonhoogtes ooreenstem met die afwesigheid van 'n toonbevel. Dis bevind dat naasliggende hoë oppervlaktoonhoogtes en naasliggende lettergrepe met alternerende oppervlaktoonhoogtes, verskillende tipes prosodiese groepe vorm, wat ooreenstem met Fujisaki-toonbevele wat oor verskeie lettergrepe strek.

In opeenvolgende eksperimente is die Fujisaki-parameters nie deur heuristieke bepaal nie, maar deur 'n masjienleeralgoritme. Drie regressiebome, een vir elke Fujisaki-toonbevelparameter, is afgerig deur van sestien lettergreepeienskappe as intree gebruik te maak, om sodoende 'n datagedrewe prosodiese model vir die Sesotho-taal te ontwikkel. Wanneer die hersintetiseerde spraak wat van hierdie model gebruik maak vergelyk word met hersintese vanuit eenvoudiger heuristieke, is die verbeterde werksverrigting duidelik. Spraak wat op hierdie wyse hersintetiseer word, is as "byna-moedertaalkwaliteit" beoordeel in terme van natuurlikheid. Dit demonstreer die potensiaal om Sesotho-spraak met natuurlike prosodie te sintetiseer vanuit teks met toonaanduidings, deur van datagedrewe tegnieke gebruik te maak.

Uiteindelik word die prosodie-generasievermoë van die Fujisaki-model vergelyk met die VMM-gebaseerde spraaksintesemodel, vir beide Sesotho en Serbies. Die resultate toon dat die Fujisaki-model 'n beter vermoë het om F0 vir Sesotho te modelleer, terwyl die VMM-gebaseerde model beter werksverrigting vir Serbies toon.

# Published Output

The following publications have flowed from the work presented in this thesis. Conference papers 2 and 5 included collaborative research.

## Peer Reviewed Conference Papers

1. **Mohasi, L.**, Mixdorff, H. and Niesler, T., "An Acoustic Analysis of Tone in Sesotho", In: Proceedings of International Congress of Phonetic Sciences (ICPhS), Hong Kong, pp. 1402-1405, 17-21 August 2011.

2. Mixdorff, H., **Mohasi, L.**, Machobane, M. and Niesler, T., "A Study on the Perception of Tone and Intonation in Sesotho", In: Proceedings of Interspeech, Florence, Italy, pp. 3181-3184, 27-31 August, 2011.

3. **Mohasi, L.**, Mixdorff, H., Niesler, T. and Zerbian, S., "Analysis of Sesotho tone using the Fujisaki model", In: Proceedings of the 3rd International Symposium on Tonal Aspects of Languages (TAL), Nanjing, China, 26-29 May, 2012.

4. **Mohasi, L.**, Mixdorff, H. and Niesler, T., "Perceptual evaluation of the effect of mismatched Fujisaki model commands and surface tone in Sesotho", In: Proceedings of the 7th International Conference on Speech Prosody, Eds. N. Campbell, D. Gibbon and D. Hirst, Trinity College Dublin, Ireland, pp. 582-586, 20-23 May, 2014.

5. **Mohasi, L.**, Sečujski, M., Mak, R. and Niesler, T., "A comparison of two prosody modelling approaches for Sesotho and Serbian", In: Proceedings of the 16th International Conference on Speech and Computer (SPECOM), Eds. A. Ronzhin, R. Potapova and V. Delic, Novi Sad, Serbia, pp. 34-41, 5-9 October, 2014.

## Journal Articles

1. **Mohasi, L.**, Mixdorff, H. and Niesler, T. "Characterization of prosodic groups in Sesotho using the Fujisaki model", Journal of Chinese Linguistics (JCL) Monograph, Accepted for publication, 2013.

2. **Mohasi, L.**, Mixdorff, H. and Niesler, T. "Perceptual evaluation of Fujisaki high-tone prosodic groups from surface tone prediction in Sesotho", Language Resource and Evaluation, Submitted, March 2014.

# Acknowledgements

My profound gratitude goes to my supervisor, Prof. Thomas Niesler and my co-supervisor, Prof. Dr. Hansjörg Mixdorff. Thank you for believing in me, urging me on to do my best, and your unwavering support throughout the course of my research.

I am mostly grateful to the following people for their support and assistance in various ways, professional linguists whom I consulted and were very willing to assist me in every way they could: Prof. 'Malillo 'Mats'epo Machobane (NUL) as a native speaker and professional linguist in Sesotho, in the contribution she made in some of the experiments and her guidance in tonal marking for Sesotho language and other aspects; Prof. Dr. Sabine Zerbian (PU) for her assistance with transcription of the data; Prof. Lutz Marten (LU), Prof. Tom Guildemann (HU), Prof. Marianne Visser (SU), Dr Laura Downing (ZAS), Chris Newmann (HU), Dr Febe de Wet (CSIR), 'Mabati Makara and Lesotho Meteorological Services (LMS) for the weather data, Dr. Milan Secujski and Robert Mak for the collaboration with University of Novi Sad, Dr Rose Richards for assistance with the writing of this thesis.

Thanks go to Basotho students from the following institutions: University of Stellenbosch (LESA-SU), University of Cape Town (LESA-UCT), University of the Western Cape (UWC) and Cape Peninsula University of Technology (CPUT). My research wouldn't have been successful without your participation and support. Ke lebohile ho menahane. I appreciate the selfless support I received from Mrs Alison Wileman and Lerato Lerato for their assistance in the execution of the different experiments at Stellenbosch University and the National University of Lesotho respectively. My DSP colleagues, thank you for the encouragement and the fun throughout my journey.

I am mostly grateful the lovely family (Mum, Zee, Neuza, Kenny, Mphoza, Lehakoe) I have been blessed with, for their moral support and prayers during this challenging time. Thank you for your love, patience, consideration, and encouragement throughout the years of my study. Last, but not least, I would like to thank my sponsors, the DFG International grant for the financial assistance towards cost of this research and the travels undertaken.

This work was supported in part by the National Research Foundation of South Africa (grant UID 71926), by a DFG International collaboration grant (Mi 625/16-1), and by Telkom South Africa. Opinions expressed and conclusions arrived at are those of the authors and are not necessarily to be attributed to the sponsors.

# Dedications

In loving memory of my dearest uncle, Sechaba Francis Mohasi. Thank you for always being there for me and for your selfless support.

# Contents

# List of Figures

# List of Tables

# List of Abbreviations

| | |
|---|---|
| AAD | Average Absolute Delta |
| AM | autosegmental-metric |
| ASR | Automatic Speech Recognition |
| CART | classification and regression trees |
| CC | correlation coefficient |
| CI | confidence interval |
| CSIR | Council for Scientific and Industrial Research |
| CTexT | Centre for Text Technology |
| FHSL | Fujisaki (tone command) high, surface (tone) low |
| FLSH | Fujisaki (tone command) low, surface (tone) high |
| FR | Finality Restriction |
| GPS | Global Positioning System |
| GTI | Grammatical Tone Insertion |
| HFC | high frequency contour |
| HLTs | Human Language Technologies |
| HMM | hidden Markov model |
| HTS | HMM-based speech synthesis |
| HTS1 | High Tone Spread1 |
| HTS2 | High Tone Spread2 |
| ICC | intra-class correlation coefficient |
| IHTS | Iterative High Tone Spread |
| IPO | Institute of Perception Research |
| KIM | Kiel intonation model |
| LBD | Left Branch Delinking |

| | |
|---|---|
| LFC | low frequency contour |
| LMS | Lesotho Meteorological Services |
| LS | Lesotho |
| LTV | Lesotho TV |
| MAE | mean absolute error |
| NUL | National University of Lesotho |
| POS | part of speech |
| PSOLA | Pitch Synchronous OverLap and Add |
| RBD | Right Branch Delinking |
| RMSE | root mean squared error |
| RNN | recurrent neural network |
| SA | South Africa |
| SAMPA | Speech Assessment Methods Phonetic Alphabet |
| SFC | superposition of functional contours |
| SHTD | Specifier High Tone Delinking |
| SNR | signal-to-noise ratio |
| STEM-ML | soft template markup language |
| SU | University of Stellenbosch |
| SUS | Semantically Unpredictable Sentences |
| TBU | Tone Bearing Unit |
| ToBI | tones and break indices |
| TTS | text-to-speech speech |
| TV | Television |

# List of Symbols

| | |
|---|---|
| $\alpha$ | natural angular frequency of the phrase control mechanism |
| $\beta$ | natural angular frequency of the phrase control mechanism |
| $\gamma$ | relative ceiling level of tone components |
| $\mu$ | the mean |
| $\rho$ | statistical correlation coeffecient |
| $\sigma$ | standard deviation |
| ˆ | high surface tone (in Fujisaki waveform figures) |
| ‘ | high surface tone |
| At | amplitude of a Fujisaki tone command |
| Ap | magnitude of a Fujisaki phrase command |
| F0 | fundamental frequency |
| F1 | first formant |
| F2 | second formant |
| Fb | baseline fundamental frequency |
| Hz | hertz |
| H | high tone |
| L | low tone |
| ms | millisecond(s) |
| [o] | SAMPA representation of a closed vowel (example of closed o) |
| [O] | SAMPA representation of an open vowel (example of open o) |
| p | probability |
| s | seconds |
| s.d. | standard deviation |
| T0 | onset time of a phrase command |
| T1 | onset time of a Fujisaki tone command |
| T2 | offset time of a Fujisaki tone command |
| t1 | onset time of a tone command |
| t2 | offset time of a tone command |
| T1rel | relative timing, mean distance between the tone command onset time and T1 |

T2rel relative timing, mean distance between the offset time T2 and the segmental onset of the syllable

T1(mod) value of T1 after modification

T2(mod) value of T2 after modification

Ton syllable segment onset

Toff syllable segment offset

# Chapter 1

# Introduction

Speech is the most natural and most prevalent form of communication for humans. Text processing tools, electronic dictionaries, and advanced speech processing systems such as text-to-speech (TTS) and automatic speech recognition (ASR) have become available for several languages. However, languages from developing countries have received little attention and remain technologically under-resourced. Under-resourced languages have some or all of the following characteristics [59, 10]: lack of a unique writing system or stable orthography, limited presence on the web, lack of linguistic expertise, lack of electronic resources for speech and language processing, such as transcribed speech data, pronunciation dictionaries, and vocabulary lists. Sesotho, a tonal Southern Bantu language, is an example of an under-resourced language which has so far received extremely little attention by the speech community.

In order for text-to-speech systems to produce intelligible and natural-sounding speech, accurate prosodic modelling is crucial. Prosodic features include the fundamental frequency (F0) contour, duration, pause and amplitude. Tone is a linguistic property marked by prosodic features such as F0 and intensity. Sesotho uses tone to convey grammatical information, while English, for example, does not. Accurate prosodic modelling can be achieved by the correct modelling of pitch in tonal languages, whereby the use of tone marking contributes significantly to the quality of synthetic speech [25]. However, the interrelationship between tone and general intonation in Sesotho is still poorly understood and technologically not addressed. This is complicated by the fact that tonal information is not indicated in the orthography [49, 87]. Prosodic modelling therefore is still a challenge for tonal Bantu languages such as Sesotho [124].

This thesis considers prosodic modelling for Sesotho using the Fujisaki model with a view to the development of a text-to-speech system. Fujisaki analysis can be used to indicate the tone associated with a syllable using only the acoustics, but often differs from the surface tone that would be available for TTS synthesis.

To the author's knowledge and prior to her related work, the Fujisaki model has not been tested for any African or indigenous language, be it tonal or non-tonal (accent). The ultimate goal is to establish automatic prosody modelling rules for Sesotho text using the Fujisaki model, and integrate these rules into a Sesotho TTS system for a natural-sounding and intelligible speech output.

## 1.1 Motivation

In the speech technology field, there has been significant progress and development for text-to-speech

synthesis. This includes formant, rule-based, unit selection and the currently popular statistical parametric HMM-based (HTS) speech synthesis approaches. Although commercial TTS systems have become available, investigation is ongoing in this discipline in a bid to achieve natural-sounding and intelligible synthetic speech. For tonal languages such as Sesotho, tone, a linguistic phenomenon, is important for the synthesis of natural-sounding speech. Van Niekerk and Barnard [110] indicate that for the development of TTS and ASR systems of tonal languages, knowledge in two areas is crucial, that of 1) surface tone transcription from text, i.e. tone assignment of syllables in target context after linguistic processes (e.g. sandhi) have been applied and 2) understanding the relationship between acoustic parameters (such as pitch) and these surface tones. The motivation behind this research is to bring this under-resourced language to the fore in terms of tone modelling for Sesotho TTS synthesis. Extensive research has been dedicated to tone modelling in East-Asian languages such as Chinese [114, 64, 65], but very little to African languages. In this sense, African languages and Bantu in particular, are under-resourced in terms of available data for tone modelling and the implementation of TTS systems.

## 1.2 Literature Synopsis

Sesotho is a Southern Bantu language spoken in Lesotho as a national language and in South Africa as one of the eleven official languages. It has two tone levels: a high tone (H) and a low tone (L), of which the high tone is active. The tone of a syllable is carried by the vowel, by the nasal (if the nasal is syllabic) or by the lateral *l* [44, 21]. Sesotho is classified as a grammatical tone language, which means that words may be pronounced with varying tonal patterns depending on their particular function in a sentence. In order to create certain grammatical constructs, tonal rules may modify the underlying tones of the word and thus lead to differing surface tones. The underlying tone, also known as the lexical tone, is the tonal pattern of the word in isolation. The surface tone, on the other hand, is a 'spoken' tone, i.e. the tone given to a word when spoken as part of a sentence. The surface tone can be derived from the underlying tone using a pronunciation dictionary, morphological analysis, and a combination of tonal rules [124]. All three components are crucial elements for speech processing systems.

In terms of text-to-speech technology advancement for Sesotho, the Meraka Institute at the Council for Scientific and Industrial Research (CSIR) in Pretoria, South Africa, undertook a project (Lwazi Project [48]) which was aimed at developing various basic speech technologies for the South African languages as the first phase from 2006 to 2009. According to the final Lwazi Project report [15], the second phase, which ran from 2010 to 2012, involved further research and development work on speech technologies for resource-scarce languages, Sesotho being one of them. During this phase, a Sesotho TTS speech corpus was developed in 2011 which contained 231 utterances (23 minutes of speech). The corpus is obtainable from the Language Resource Management Agency of the affiliated institution, the North-West University [1]. The orthographic format used for text data is that of the South African convention.

A TTS speech corpus needs to encapsulate prosody for a natural-sounding synthesis. The Fujisaki model [30] is a manageable and powerful model for prosody manipulation and has shown a remarkable effectiveness in modelling the F0 contours. It is reliant on the acoustics of the uttered speech. Its validity has been tested for several languages, including tonal languages such as Mandarin [72], Thai [80], and Vietnamese [23]. The model, which is formulated in the log F0 domain, analyses the F0 contour of a natural utterance and decomposes it into a set of basic components which, together, lead to the F0 contour that closely resembles the original. These components are: a base frequency, a phrase component, which captures slower changes in the F0 contour as associated with intonation phrases, and a tone component that reflects faster changes in F0 associated with tones. The tone commands of the Fujisaki analysis are therefore an indicator of tones in the utterance.

The model was first proposed by Fujisaki and his co-workers in the 70s and 80s [30] as an analytical method which describes fundamental frequency variations in human speech. By design, it captures the essential mechanisms involved in speech production that are responsible for prosodic structure. A chief attraction of the Fujisaki model lies in its ability to offer physiological interpretation that connects F0 movements with the dynamics of the larynx, a viewpoint not inherent in other currently-used intonation models which mainly aim to break down a given F0 contour into a sequence of 'shapes' [99]. The Fujisaki model has been integrated into a German TTS system and proved to produce high naturalness compared with other approaches [79]. The inverse model, automated by Mixdorff [74], determines the Fujisaki parameters which best model the F0 contour. However, the representation of the F0 contour is not unique. In fact, the F0 contour can be approximated by the output of the model with arbitrary accuracy if an arbitrary number of commands is allowed [2]. Therefore, there is always a trade-off between minimising the approximation error and obtaining a set of linguistically meaningful commands.

## 1.3 Research Objectives

The main objective of this research is to derive a means of automatic prosodic modelling for Sesotho using the Fujisaki model and the surface tonal transcription. This will be achieved by the following tasks:

- Prosodic marking of Sesotho text by surface tonal transcription (corpus development).
- Establishing the relationship between surface tonal transcription and the Fujisaki model.
- Determining the correct Fujisaki parameters (phrase and tone commands), their position, magnitude (for phrase commands) and amplitude (for tone commands) based on the prosodically-marked Sesotho text.
- Performing perceptual tests to establish the naturalness and intelligibility of the synthesised speech at appropriate stages of the research from the Sesotho tone-marked text.
- Developing a statistical prosodic model which will automatically determine Fujisaki tone command parameters.
- Synthesising speech using a Sesotho HMM-based speech synthesis system and comparing the output with resynthesized speech from a Fujisaki model-based analysis.

Other objectives include:

- Exploring linguistic factors such as tone sandhi, Obligatory Contour Principle (OCP), peak delay, and downstep in order to incorporate them into our prosodic modelling rules.
- Verifying and/or correcting the lexical tone of some Sesotho words, and to establish tone for proper nouns and other words not found in current dictionaries, based on our database.

## 1.4 Project Scope and Contributions

The scope of this research is limited to the application of the Fujisaki model as a prosody modelling tool for the Sesotho language. Other prosody modelling tools will be presented and compared with the Fujisaki model, but will not be implemented. Bantu languages other than Sesotho will not be explored.

Experimentation will be limited to the corpus compiled as part of this research. The orthography of the corpus is that of the Lesotho convention format.

Contribution by this research is relevant to both the scientific and non-scientific fields.

- **Adding new knowledge in terms of an innovative approach in the development of a natural-sounding and intelligible Sesotho TTS system**. New knowledge is added in two ways: 1) This is a first application of prosody modelling using the Fujisaki model to an African language or Bantu language. 2) To the author's knowledge, it is also the first application whereby the Fujisaki parameters are established heuristically. This new method can be extended to other tonal Bantu languages, including those which have no prosodic marking.

- **Increase of speech database for Sesotho**. Sesotho is under-resourced in terms of available data and human resources for speech technology purposes. It is a national language in Lesotho but its orthographic format has not been explored for speech processing. Currently, there is no existing TTS speech corpus in the country. The contribution of this research therefore, is in terms of transcribed data, pronunciation dictionary and a vocabulary list in the Lesotho orthographic convention. The system will be an effective tool in synthesising any Sesotho text written in this format, with the assistance of a mapping tool from the Lesotho to the South African conventional writing when necessary.

- **Standardising tone marking in Sesotho**. The establishment of correct tone marking will be supported by the application of the Fujisaki model. This will result in a more accurate tone-marked dictionary (for lexical tone) and subsequently a more accurate surface tonal transcription.

- **Cultural identity and empowerment**. Spoken language is the primary means of human communication. Language is not only a communication tool, but also fundamental to cultural identity and empowerment [8, 41]. The existence and accessibility of a TTS system that can synthesise Sesotho text written in the Lesotho orthographic format will both empower Lesotho citizens and make the language more attractive to them. This system will also help overcome barriers related to language, illiteracy and disability as TTS can be used in applications such as screen readers for the blind and for interactive response systems [14, 71].

## 1.5 Thesis Overview

The structure of this thesis is as follows: In Chapter 2, a Southern Bantu language, Sesotho, is introduced for familiarity to the reader. A brief background is given, followed by different orthographical formats used in the two countries where it is mostly spoken. The tonal system used for the language is also described in detail. Chapter 3 gives a detailed account of the Fujisaki model, and its deployment for various languages. Other prosody modelling tools, which have been used in the past and those in current use, are also presented. In the chapter, it is also elaborated on why the Fujisaki model was chosen for this study over other intonation modelling tools. Chapter 4 details preliminary experiments conducted as the initial step in the application of the Fujisaki model for the Sesotho language. Through these experiments, the tonal pattern of Sesotho utterances from the Fujisaki analysis was determined as portrayed by the model's parameters. Intonation of statements and questions, where stimuli were modified, was also explored. Further experiments which were performed are explained in Chapter 5, where a larger corpus was developed. These experiments were essential as a guide on the appropriate and consecutive steps to follow from the subsequent results obtained. These results were important because there was no reference

on use of the Fujisaki model in any Bantu or African language, except for tonal Asian languages whose tonal system differs from that of Sesotho. In Chapter 6, a statistical prosody model is developed using regression trees. Perceptual experiments are also performed whereby the naturalness of the synthesised speech from this model and that from the heuristic approach in Chapter 5 is evaluated. In Chapter 7, the prosody generation ability of the Fujisaki model and that of an HMM-based speech synthesis (HTS) system is explored for the Sesotho and Serbian languages. The thesis ends with the overall summary and conclusions in Chapter 8.

# Chapter 2

# The Sesotho Language

## 2.1 Introduction

The chapter introduces the Sesotho language, the focal language on which the prosodic modelling research is based. A background of the language is given, followed by an explanation of the different orthographies the language has. Prosodic modelling entails manipulation of tone, and in Sesotho, the tone is carried by a syllable. The two aspects are discussed in detail with reference to their role in this study for text-to-speech (TTS) development. Other essential linguistic phenomena, related to tone and relevant to our work, are also touched on.

## 2.2 Background

Sesotho is the national language of Lesotho, a country with approximately two million inhabitants. It is also spoken in South Africa, where it is one of the 11 official languages and a mother tongue for approximately 3.8 million people (7.6% of the population) [97]. It is one of the first African languages to be reduced to writing by the European missionaries who arrived in Lesotho in 1833. In South Africa, Sesotho has two related dialects, Setswana and Sesotho sa Leboa (Northern Sotho), with which Sesotho is largely mutually intelligible [20]. The three languages are classified as belonging to the Sotho-Tswana group of languages, and more generally, to the Southern Bantu language family. Like all other Bantu languages, Sesotho is an agglutinative language spoken conjunctively. Suffixes and prefixes are used extensively in sentence construction and can cause sound changes. However, in contrast with some Bantu languages such as the South African Nguni languages, e.g. isiZulu, isiXhosa, Sesotho is written disjunctively, with a characteristically European word division used for writing the language.

## 2.3 Orthography

Although Sesotho is spoken in both Lesotho and South Africa, the two countries have different orthographic conventions. The Lesotho orthography is older and uses diacritics on some vowels to distinguish ambiguous spellings [20]. For instance, the vowels *o* and *e* have variants written as *ò*, *ō* and *è* and *ē* respectively. (Other scholars like Kock [55] use the circumflex ˆ to distinguish between these vowels, e.g. *o* and *ô*, *e* and *ê*. This form has however largely fallen into disuse.) These variants can be distinguished in phrases such as the following:

**Table 2.1:** Differences between the Lesotho and South African orthographic conventions.

| Lesotho | South African | Examples |
|---|---|---|
| li, lu | di, du | likeleli - dikeledi (tears), lumela - dumela (hello) |
| kh | kg | khathatso - kgathatso (trouble) |
| ts' | tsh | Mots'eanong - Motsheanong (May month) |
| ch | tjh | sechaba - setjhaba (nation) |
| e | y | moea - moya (wind) |
| o | w | ho utloisisa - ho utlwisisa (to understand) |
| fsh | fj | ho bofshoa - ho bofjwa (to be tied) |
| psh | pjh | mpshe - mpjhe (ostrich) |
| 'm | mm | 'm'e - mme (mum, mother) |
| 'n | nn | 'na - nna (I, me) |

*ho tsêla* → "to pour" *ho tsēla* → "to cross"
*ho ròka* → "to sing a praise poem" *ho rōka* → "to sew"

These examples also have differing tone patterns and are considered to be minimal pairs.

One other change in the Lesotho written form is the use of the symbol *ŝ*, which represents the aspirated alveolar affricative *tŝ*. Because the symbol is not standard on all keyboards, Lesotho has resorted to writing *ts'* instead.

The orthographies also differ in the choice of letters, marking of initial syllabic nasals, and (to a much lesser extent) in written word division. Examples of these differences are shown in Table 2.1.

Other differences in orthography between the two languages include:

- In the Lesotho orthographic convention, the lateral *l*, when followed by *i* or *u*, gives a *d* sound. The letter *d* is not part of the Lesotho orthography, though it is used in the South African orthography. Doke [21] mentions that the Sesotho orthography uses the symbol *d* in cases of foreign acquisitions. For example, *Davida* (David), *dinare* (dinner), *daemane* (diamond), *Diabolose* (Devil). However, except in the case of *Diabolose*, the pronunciation of *d* in these words is a Sesotho *t* (ejected), e.g. *Tafita*, *tinare*, *taemane*, and *t* should have been used in the orthography.

- When a word begins with a syllabic nasal (*m*, *n*) and is followed by the same nasal, it is written differently in Lesotho; the initial syllable is replaced by an apostrophe in the Lesotho orthography, as the last two examples in Table 2.1 show. When not word initial, the Lesotho orthography uses the same one as the South African one, e.g. *bonngoe* (unity).

- The Lesotho orthography also uses apostrophes to indicate the missing sounds due to omitted consonants or vowels. This is applicable in contractions of words. The South African orthography does not use this format. Examples are given in Table 2.2.

- The prosodic penultimate *e-* that is sometimes affixed to monosyllabic verbs is written with a dash in Lesotho, e.g. *eba!* = *e-ba!* → "and then".

- The class 2a prefix is usually simply attached to the class 1a noun in South African form but Lesotho orthography uses a dash, e.g. *ntate* "father" → *bontate* = *bo-ntate*.

- The 'focus marker' /-a-/ is inserted between the subject concord and the verb stem in different ways in the two orthographies. This is probably the most commonly encountered difference between the word divisions of the two orthographies. E.g.

**Table 2.2:** The Lesotho (LS) and South African (SA) writing on contractions.

| Sentence | Contractions (LS) | South African (SA) writing | Meaning |
|---|---|---|---|
| Ha a tle. | H'a tle. | Ha tle. | He/she is not coming. |
| Mora oa ka. | Mor'a ka. | Mora ka. | My son. |

*Likhomo lia fula.* (LS) - *Dikgomo di a fula.* (SA) → "The cows are grazing."

- In order to distinguish between concords of class 1a and the $2^{nd}$ person singular, Lesotho orthography uses *u* to represent phonetic *o* and *w* for the $2^{nd}$ person, while for the $3^{rd}$ person, *e* is represented by *y*. (*y* and *w* are considered as semi-vowels. A detailed explanation can be found in [87], [44] and [21]). E.g.

  *Ke uena ea tlang.* (LS) - *Ke wena ya tlang.* (SA) → "It is you who is coming."
  *Ke eena ea tlang.* (LS) - *Ke yena ya tlang.* (SA) → "It is him/her who is coming."

## 2.4 The syllable

The syllable is the basic feature on which this study is based. For the Sesotho language, the syllable is the bearer of tone as it carries the prosodic information. In speech processing, the syllable plays a crucial role in the production as well as perception of speech. According to Yip [116], it is not always clear whether tones associate to segments, moras or syllables. However, her study on different languages revealed that the tone bearing unit (TBU) can be either a mora or a syllable, and not a segment, as tone always associates with prosodic entities. Languages can differ in whether the syllable or the mora is the TBU.

### 2.4.1 Syllable Type

Sesotho has three types of syllable [44]:

1. The type that consists of a vowel only, i.e. V syllable.
2. The type that consists of a consonant plus a vowel, i.e. CV syllable.
3. The type that consists of a syllabic consonant, i.e. C syllable.

Of the three types of syllables two end in a vowel. For this reason, syllables in Sesotho as well as in other Bantu languages are said to be open. These syllable types are distributed as follows:

- V syllables
  - Initial vowel syllable, e.g. *a-ba* (divide), *e-ma* (stand).
  - Medial vowel syllable, e.g. *le-e-ba* (dove), *pho-o-fo-lo* (animal).
  - Final vowel syllable: e.g. *bu-a* (speak), *mo-ru-i* (a rich person).

  Some vowel syllables in medial and final position tend to be pronounced with a preceding semi-vowel glide, e.g. *le-ee-ba*, *bu-oa*.

- CV syllables

  Examples are *bo-na* (see), *bo-tsa* (ask), *ta-fo-le* (table). In the case of CV syllables, it should be noted that one syllabic margin may, in spelling, be represented by more than one letter. However, these segments of two, three or sometimes more letters of the alphabet represent a single sound. Hence, we have CV CV CV examples such as: *pje-tle-tsa* (emit, spit); *pshe-mo-ha* (become loose).

- C syllables

  The four nasal consonants, *m*, *n*, *ng*, *ny*, and the lateral *l*, may each occur syllabically, i.e. as a syllable. Examples include *m-mu-so* written *'muso* (government), *po-m-po-ng* (sweet), *n-na* written *'na* (I, me).

### 2.4.2 Syllable Length

In normal speech, three lengths of syllables occur in Sesotho [44]:

1. Syllables with normal or short length, which are not marked. Normal length occurs in final syllables, non-penultimate syllables and in monosyllabic words: e.g. *motse* (village), *ja* (eat), *mamela* (listen).

2. Syllables with half length, which in this thesis are indicated by /•/ after the vowel of the syllable or following a syllabic consonant. Half length occurs in syllables which are in a non-final position: e.g. *Ba•tho ba ra•ta ho phe•la ka kho:tso.* (People like to live in peace.)

3. Syllables with full length indicated by /:/ in this thesis, after the syllable or syllabic consonant. Full length occurs in the penultimate syllables of words pronounced in isolation or at the end of a sentence: e.g.

   *bapa:la* - "play"
   *Bana ba ka bapa:la.* - "Children may play."

   Full length always occurs with the future infix /-tla-/ or /-ea-/, even if the predicate is not at the end of a sentence. e.g.

   *Re tla: bo:na.* - "We shall see."
   *Re ea: te:ng.* - "We are going there."

   Other infixes with regular full length are /-n'o-/ (to do as a rule) and /-nt'o/ (to do after), e.g.

   *U n'o: bu:a han•tle le ba:tho.* - "You should speak properly to people."
   *Jaa: u nt'o: sebe:tsa.* - "Eat and work thereafter."

   Full length also occurs regularly with the infix /-a/ of the present indicative positive long form and the perfect indicative negative: e.g.

   *Kea: bo:na.* - "I see."
   *Ha kea: bo:na.* - "I did not see."

   Similarly, monosyllabic radicals in the present participial are always preceded by prefixal morphemes bearing full length: e.g.

*Ha a: e-ja, u tho:le.* - "When he/she is eating, you should keep quiet."
*Ha a sa: je, oa: ku:la.* - "If/when he/she is not eating, he/she is sick."

Further, prolonged or abnormal length occurs in emotional and dramatic speech including ideophones and interjectives. When such length occurs in the final syllable of the word, it is best indicated by doubling the final vowel or consonant that represents the lengthened sound, e.g.

*Joo!* - "Alas!"
*Ahee!* - (of greeting, thanks)

Prolonged vowel length also occurs in non-final syllables. In such cases, it may be indicated by a repeated colon as above or by dots after the lengthened vowel, e.g.

*Thaba e::la* or *Thaba e...la.* - "Yonder mountain."
*Ho::le* or *Ho...le.* - "Very far."

## 2.5 Tone

Sesotho is a grammatical tone language and like most Bantu languages, it uses a register tone system where the distinguishing feature is the relative difference between a high and a low pitch, also denoted as high tone (H) and low tone (L). In a grammatical tone language, words may be pronounced with varying tonal patterns depending on their particular function in a sentence. In order to create certain grammatical constructs, tonal rules may modify the lexical tones of the word and thus lead to differing surface tones.

According to Khoali [51], the tonal rules of Sesotho can be classified into two groups: assimilatory and dissimilatory. Assimilatory rules involve the spreading of high tones. In Sesotho, there are two types of high tones: lexical high tones and grammatical high tones. Lexical high tones assimilate by spreading one syllable to the right whereas grammatical high tones spread all the way to the end of the verb. Dissimilatory processes, on the other hand, involve various kinds of high tone deletion and delinking.

### 2.5.1 Lexical Tone

Lexical tone, also known as the underlying tone, is the tonal pattern of a word when spoken in isolation. These tonal patterns need to be compiled in a pronuciation dictionary. Pronunciation dictionaries specify a mapping between the orthographic representation of a word and its pronunciation, and are fundamental to most modern speech processing systems. According to Zerbian [124], such a dictionary should provide both lexical and grammatical tone – lexical tone for nouns, verbs and adjective stems, and grammatical tone to indicate tense, mood, and aspect for verbs.

Schadeberg [93] points out that only two dictionaries with tone markings are available for Sesotho, namely that compiled by Mabille *et al.* [68] and that compiled by Du Plessis *et al.* [22]. However, a closer comparison of the two reveals considerable variation in the tonal patterns they specify [93]. Furthermore, not all Sesotho words considered in this work appear in both dictionaries. In their introduction, Mabille *et al.* [68] mention that their dictionary is incomplete, and also that some tone markings might not be accurate. In addition, entries in this dictionary only showed tone marks for the 'main' word, but not for its relatives. Du Plessis *et al.* [22], on the other hand, provide tonal pattern also for these relatives.

**Table 2.3:** Tonal pattern match of 122 random words in the three tone-marked dictionaries by Mabille *et al.* [68], Du Plessis *et al.* [22] and Kriel *et al.* [60].

| Dictionary comparison | Number of matching tonal patterns |
|---|---|
| All three dictionaries | 40 |
| Mabille *et al.* *vs* Du Plessis *et al.* | 49 |
| Mabille *et al.* *vs* Kriel *et al.* | 50 |
| Du Plessis *et al.* *vs* Kriel *et al.* | 64 |

**Table 2.4:** Comparison of the tone assigments for a selection of dictionary entries. The acronyms are: K = the Northern Sotho dictionary; DP = the Sesotho dictionary, K_rel/DP_rel = the relative of the word in the Northern Sotho/Sesotho dictionary. A relative is a word similar in origin and related in meaning to the 'main' word. For example, *batla* is the relative (stem) of *batlang*. For the tonal pattern, H denotes a high tone and L a low tone.

| Word | In which dictionary? | Tonal pattern | Stem in dictionary | Do dictionaries agree on tonal pattern? |
|---|---|---|---|---|
| atleha | K, DP | LLL | - | Yes |
| batlang | K_rel, DP_rel | - | batla (LL) | Yes |
| bokella | DP | HHLL | - | Word not in K |
| bontsha | K, DP | HHL/HLL | - | No |
| kgemang | K_rel, DP_rel | - | kgema (HL/LL) | No |

Although there are more entries in [68] than in [22], most words in the former are not tone-marked. Therefore, [22] was used in our work as the primary source of tone-marked pronunciations.

A tone-marked Northern Sotho dictionary by Kriel *et al.* [60] was used as an additional source to obtain the lexical tones for words which were not found in [22]. Northern Sotho belongs to the same language family as Sesotho, and uses many of the same tonal patterns. It was found, however, that both [22] and [60] contained many entries for which the lexical tone pattern differed from [68].

In order to demonstrate the differences and similarities in the tonal patterns provided by the three dictionaries described above, a sample of 122 random tone-marked words was selected from [68]. Of these 122 words, 102 were also present in [22] while 90 were also present in [60]. As shown in Table 2.3, of the total entries, approximately one third of the words had matching tonal patterns in all three dictionaries. The best agreement was found between [22] and [60], for which 64 of 102 words present in both dictionaries were marked with the same tones.

Table 2.4 shows a few selected examples and their tonal patterns as provided in [22] and [60]. The tonal pattern is presented per syllable in a word. For instance, *batlang* is a 3-syllable word, *ba-tla-ng*, and all syllables are low-toned, thus LLL. Where there is disagreement, the tonal pattern of the word from each dictionary is given. For instance, the word *bontsha* appears in both dictionaries but has a different tonal pattern – HHL in the Northern Sotho dictionary and HLL in the Sesotho dictionary. In the case for *batlang*, its stem, *batla*, appears in both dictionaries with tonal pattern LL, yet the word *batlang* is not in any of the dictionaries. The entry for *bokella*, on the other hand, appears in the Sesotho dictionary only, and not in the Northern Sotho dictionary. Note that the word *bontsha* is given in the South African orthographic format. In the Lesotho orthography, it would be *bonts'a* or *bontŝa*. However, the tonal pattern of words is not affected by the different orthographies.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| re | tla | qála | pé̲l̲e̲ | ká̲ | hó | sheba | pú̲la | é̲ | atléh-íle-ńg |
| SC2PL | FUT | start | first | with | NP15 | look | NP9.rain | REL.9 | manage-PERF-REL < atleha |

| | | | | |
|---|---|---|---|---|
| ho | bó̲kéll-oa | le-tsatsí-ng | lá̲ | máobane |
| NP15 | collect-PASS | NP5-day-LOC | POSS.5 | yesterday |

**Figure 2.1:** An example of a morphological analysis of a Sesotho sentence: *Re tla qala pele ka ho sheba pula e atlehileng ho bokelloa letsatsing la maobane* - "We will first start by checking the rain that was measured yesterday."

### 2.5.2 Morphological Analysis

In order to motivate tonal transcriptions, a morphological analysis is necessary. Morphological analysis deals with identification and description of the structure of a given language's morphemes and other linguistic units such as root words, affixes, parts of speech, intonation and stress, or implied context. Morphological structure influences the behaviour of tones. In Bantu languages, for example, verbs have a complex morphology, and tonal processes may apply to sub-constituents of the verb only [116]. Zerbian [124] further adds that only a morphological analysis of the verb word in its context can lead to the derivation of the correct tonal pattern.

Figure 2.1 shows an example of a morphological analysis of a Sesotho sentence. The morphemic glosses used follow the Leipzig glossing rules [27] wherever possible. Abbreviations used are: SC2PL = subject concord 2nd person plural, FUT = future (tense), NP15 = noun prefix Class 15, NP9 = noun prefix Class 9, REL = relative, PERF-REL = perfect (tense) relative, PASS = passive, LOC = locative, and POSS = possessive. In the example, the morphological analysis identifies the word *atlehileng* as a verb in the perfect tense form (*atleha* is the original form). The tonal rule to be applied in this context is the Grammatical Tone Insertion (GTI) rule, which associates the second syllable of the verb stem with a high tone. A detailed description of all tonal rules follows in the next section.

### 2.5.3 Surface Tone

For tone modeling in African tonal languages, in which tone is not indicated by the orthography, a process that can provide the tonal labels of syllables in a sentence is a prerequisite [67, 124]. This process is known as surface tonal transcription.

A surface tonal transcription is deduced from the lexical tone provided by the pronuciation dictionary and from the results of a morphological analysis by means of a set of tonal rules. These rules describe the steps which must be taken to determine the surface tones. In Sesotho, such rules have been described in the literature, although scholars differ on their specific nature. Much work has been carried out by Kunene [61], Doke and Mofokeng [21], and by Khoali [51].

Khoali, whose work on Sesotho tone is the most recent, points out possible deficiencies in previous studies. He maintains that the previous studies were 'observational' and had no theoretical basis to sustain their discussions. The rules developed by Khoali were followed in the surface tonal transcription of our corpus. These rules were chosen mainly because they are the most detailed and that some have already been used in the analysis of other Sotho-Tswana corpora [90]. However, the development and refinement of tonal rules in Sesotho remains an active area of research in the linguistic literature.

In the following, seven tonal rules as proposed by Khoali [51] are described. Of these, only five were used for our corpus as they were the most applicable. Words of focal interest in the examples are written in bold. The underlying high tones are indicated by underlining and surface high tones by acute accents. For example, in the phrase

*Matsatsí á̲ mab̲éli á̲ látélańg.*

"The next two days."

*tsí*, *á*, *bé*, and the second *á* have both underlying and surface high tones, while *lá*, *té*, and *ńg* have high surface tone only.

1. High Tone Spread (HTS1)

   This rule spreads an underlying high tone to the immediate right syllable and is applicable across word boundaries. Care must be taken though since "a lexical high tone does not spread to toneless tone bearing units (TBUs) that belong to different words. The high tone particle, however, spreads across to toneless class prefixes across word boundaries but not onto a stem." [51]. For instance, in the example below, HTS1 should not be applied to the words ***hó naha***.

   *... káharé* ***hó naha*** *eá ronná.*
   "... in our country."

   The underlying high tone on *hó* does not spread to *na* of *naha* because these are different words: *hó* is a locative and *naha* is a low-tone noun. However, it is applicable in the following example because *sá* is a high-tone particle, and the high tone targets the low-tone class prefix *mo* of *mongobo*.

   *... sets'oáńts'o* ***sá móngobo*** *...*
   "... the humidity picture ..."

2. Iterative High Tone Spread (IHTS or HTS2)

   This involves the iterative spreading of a grammatical high tone all the way to the end of the verb. In the following example, the rule shows how a grammatical high tone spreads iteratively within the verb *lebelletsoe*. The high tone on the second stem syllable *bé* in the word spreads iteratively all the way to the end of the verb.

   *Maémo á ntsé á* ***lébéĺlétsoé*** *ho tsoela pele ...*
   "The situation is still expected to continue ..."

3. Grammatical Tone Insertion (GTI)

   This rule associates the second syllable of a verb stem with a high tone. This occurs in the following contexts: the imperative mood, the negative verb form, the participial narrative past tense, the habitual verb form, the perfect tense, and the subjunctive mood. The following example, which is in perfect tense form, illustrates the application of this rule.

   *Butha-Buthe é* ***atléhílé*** *ho bókéĺla ...*
   "Butha-Buthe has succeeded to accumulate ..."

   The low-tone verb, *atlehile*, is in perfect tense and hence has a grammatical high tone on the second syllable. This high tone spreads all the way to the end of the verb.

4. Right Branch Delinking (RBD)

   This rule dissociates the immediate right branch of a multiply-linked high tone syllable if, and only if, there is a high tone syllable immediately after the target of the HTS rule.

   *... lintlheng tsé le****shóme lé*** *métso ...*
   "... at ten points and ... units ..."

According to the HTS1 rule, the high tone from *shó* of *leshóme* spreads to *me*, but then the following article, *lé*, is also high-toned. Therefore, the high tone on *me*, which was linked from *shó*, will be delinked, following the RBD rule.

5. Left Branch Delinking (LBD)

   This rule delinks the immediate preceding left branch of a multiply-linked high tone syllable if it is preceded by a high tone syllable. An example below, taken from Khoali [51] shows the high tone on the second syllable of the verb stem *kgurukgurumetse* surfaces with a low tone. This syllable is associated with a high tone by GTI since the verb stem is in the negative form. This high tone then spreads by HTS2 until the penultimate syllable. The high tone on the second syllable of the verb stem is delinked by the LBD rule because the first syllable is high-toned.

   *Ha ké* ***kgúrukgú****rúmétsé* ***...***
   "I don't cover a bit ..."

   LBD is, in essence, a mirror image of RBD. Instead of delinking the right branch of a multiply-linked structure, the left branch is delinked if it is preceded by another high tone. The delinked syllable acquires a low tone by default rules.

6. Specifier High Tone Delinking (SHTD)

   This rule delinks a high tone on an object concord or reflexive prefix if they are not within the same phonological word as the stem. An example from Khoali [51] shows an underlying high-toned object prefix *mo* surfacing with a low tone. The high tone on the object prefix spreads by HTS1 to the first syllable of the verb stem *kgaramedítse*. The high tone on the object prefix then delinks since it is not in the same phonological word domain as the verb stem.

   ***Ó mo kgá****rámédítse.*
   "He pushed him/her."

7. Finality Restriction (FR)

   This rule exempts syllables at the end of a phonological phrase from the application of tonal rules. For example,

   *... mashóme á mabéli á métso é me****ráro****.*
   " ... twenty-three."

   The HTS1 rule indicates that the high tone on *rá* of *meráro* should spread to the *ro* that follows it. The finality restriction rule will not allow this, since *ro* is at the end of a sentence or a phonological phrase. This rule is, however, not applicable to relative verbs, whose last syllable by default has an underlying high tone, e.g. *Matsatsí á mabéli á látéla****ńg***. → "The next two days." According to Zerbian [119], if *ng* is a relative suffix, then it is always high-toned.

### 2.5.4 Tone Sandhi, Obligatory Contour Principle and Other Tone Phenomena

There are several other linguistic phenomena associated with tone which must also be taken into consideration in prosodic modelling for a natural sounding TTS system. This section will briefly describe those most relevant to our work.

### Tone sandhi

The high tone is the active tone in Sesotho and other Bantu languages. It participates in tone spread and is characterisitic of the so-called tone sandhi, a phonological change occurring in tonal languages, that can be observed if a high tone occurs together with other high tones. Tone sandhi is present in Sesotho as indicated by Demuth [19], and is modelled by the surface tonal rules HTS, which Demuth refers to as HTD, and IHTS. Other tonal rules, however, do not model tone sandhi. Ekpenyong and Udoh [26], in their study of Ibibio, a tone language spoken in the south-east region of Nigeria, emphasise the importance of tone sandhi and its effect on the overall fundamental frequency (F0) contour in tone languages. In their study, they proposed a fuzzy logic framework for modelling prosody in tone language systems. The proposed model was found to be suitable for the precise prediction of average F0 contour patterns in human speech.

### Obligatory Contour Principle

Tone sandhi leads our discussion to the related Obligatory Contour Principle (OCP). OCP is a phenomenon where adjacent identical tone elements are prohibited. According to Yip [116] , OCP violations can be avoided in a variety of ways, such as:

- tone deletion, where one of the adjacent high tones is deleted;
- blocking of high tone spread if it leads to adjacency; or
- fusion between tones, where two high tones are fused into one high tone.

Zerbian and Barnard [121] mention other repair strategies which are:

- retraction, where both high tone syllables are preserved but one is retracted away from the other, thus allowing intervening low tones.
- downstep, whose full description follows in the next sub-section.

Lastly, high tones underlie positional restrictions [120], depending on which syllable in a phrase or clause they originate from. For instance, the finality restriction, as explained in the previous section, does not generally prohibit high tones on phrase-final syllables. Underlying high tones are allowed to surface in a phrase-final position [66, 44, 118, 119], thus sometimes violating the OCP rule. In surface tonal transcription of Sesotho, OCP is observed via the RBD, LBD and FR rules. The resolution of OCP violations in Bantu languages has not yet been studied acoustically [121].

### Downstep

Downstep is a well-known phenomenon in the Bantu languages [51]. Downstep is a phenomenon where, if two adjacent high tones are realised, the second might be realised with a lower pitch in certain syntactic contexts. However, this is not an absolute resolution strategy as the two high tones are still adjacent, and the pitch plateau of the second is significantly lower than the level induced by the declination alone. Downstep is phonologically determined, most often, by an intervening low tone [121].

A preliminary study of downstep in Sesotho was carried out by Kunene [62] where he refers to realised downstep, denoted by the symbol **v**, and potential but unrealised downstep, denoted by **X**. **X** marks a position within the speech sequence where a downstep may reasonably be expected, but does not occur in the particular situation. According to this study, the observations made are:

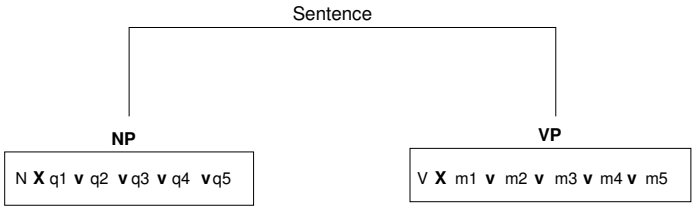



**Figure 2.2:** Downstep representation on the noun phrase (NP) and the verb phrase (VP). The symbol representations are: N = noun, q = qualifier, V = verb, and m = modifier.

1. Downstep occurs consistently between a noun phrase (NP) and a verb phrase (VP) in a subject-predicate relationship to each other, regardless of the length and complexity of either phrase.
2. The internal structure of each phrase is unaffected by the proximity of the other phrase.
3. The noun phrase and the verb phrase share two important intonational features, and these are:
   (a) the first qualifier after the noun has an unrealised downstep (**X**) between it and the noun. Similarly, the first modifier after the verb has an unrealised downstep (**X**) between it and the verb; and
   (b) each pair of qualifiers (if more than one is used) has a realised downstep (**v**) between them. Similarly, each pair of modifiers (if more than one is used) has a realised downstep (**v**) between them.

   These observations have been illustrated graphically in Figure 2.2.

On the other hand, when the verb is followed by an object instead of a modifier, a realised downstep (**v**) occurs between the verb and the object. For example, [*ba bina*] **v** [*pina*] (they sing a song) where *pina* is an object.

Kunene further explains that mid-word downstep occurs under the following circumstances:

1. the emphatic form of the second position demonstrative, e.g. *sée* **v** *nó* → "that one (near you)".
2. the emphatic form of the third position demonstrative, e.g. *sáa* **v** *né*, *saa* **v** *néné* → "that one over there".
3. the conjunctions *hója* **v** *né* (if, if only, while), *hóla* **v** *né* (if, if only, while), and *há* **v** *éba* **v** *né* (if).

The author of this thesis is not aware of any follow-up acoustic studies available to verify these rules in Sesotho or any other acoustic study of downstep in other Bantu languages. However, in their study on predicting utterance pitch for Yoruba, an African language spoken in Nigeria, van Niekerk and Barnard observed significant evidence for downstep in the H-LH context in speech recorded by all speakers, and that downstep seems to be directly dependent on tonal context [111]. The aforementioned rules are not modelled by the tonal rules described in Section 2.5.3. Their verification and implementation are a subject that requires further research.

### Peak delay and anticipation

Other linguistic factors associated with tone in a tonal language such as Sesotho are peak delay and anticipation. Peak delay is observed when the F0 peak corresponding to a high-toned syllable occurs at the end of that syllable or in the following syllable [84, 116]. This phenomenon is quite common

cross-linguistically [121]. Anticipation is observed when a high tone is realised on the preceding syllable [47]. Peak delay is modelled in the tonal rules via the HTS1, IHTS and GTI rules. Anticipation, on the other hand, is not modelled in surface tone transcription.

## 2.6 Summary

This chapter has given a brief background on the Sesotho language, a Southern Bantu language spoken in Lesotho and in South Africa. The different orthographic conventions used by the two countries have been highlighted. The properties of the syllable, which is a carrier of tone in Sesotho, have been described. Subsequently, the tone, its different forms, and the process of surface tonal transcription were explained. The chapters which follow will apply the processes described in this chapter to develop a computational model for Sesotho tone that can be used in text-to-speech (TTS) synthesis.

# Chapter 3

# The Fujisaki Model

## 3.1 Introduction

The Fujisaki model was introduced by Hiroya Fujisaki and his co-workers in the 70s and 80s [30]. The model relies on the extracted F0 contour of a natural utterance which is a superposition of a set of basic components which, together, constitute an F0 contour that closely approximates the original. This model was first proposed as an analytical model which describes fundamental frequency variations in human speech. By design, it captures the essential mechanisms involved in speech production that are responsible for prosodic structure. The best quality about the Fujisaki model is that it offers a physiological interpretation that connects F0 movements with the dynamics of the larynx, a viewpoint not inherent in other currently-used intonation models which generally attempt to break down a given F0 contour into a sequence of 'shapes' [99]. This form of modelling approach is also considered to be the most appropriate for high-quality synthesis.

The model was initially developed for the Japanese language [31] and it has shown its effectiveness in the production of a highly natural speech for languages such as Basque [85], Chinese [33], English [32, 39], German [76, 79], Greek [37], Korean [29], Portugese [102], Spanish [38], and Swedish [35], to name a few. Other tonal languages that have been investigated using this technique include Mandarin [72], Thai [80], and Vietnamese [23]. The model is capable of producing extremely close approximation to the F0 contours of natural utterances from a set of commands that are closely related to their underlying linguistic information [30].

## 3.2 Generation of the F0 Contour

The F0 contour is the consequence of control of the vocal fold vibration by a set of neuromotor commands carrying linguistic information; therefore, the process of F0 contour generation can be modelled by the input commands and the mechanism that generates the F0 contour as the response [36]. The model, which is formulated in the log F0 domain, decomposes the F0 contour extracted from the audio samples into three components: a base frequency (Fb) relative to which the F0 contour will be modelled, phrase components which capture slower changes in the F0 contour as associated with intonation phrases, and components which reflect faster changes in the F0 associated with high tones. The baseline fundamental frequency is constant in each utterance. In tonal languages such as Sesotho and Standard Chinese [34], the accent components are referred to as tone components and will henceforth be referred in this thesis. The phrase and tone components can be interpreted as responses of the model to impulse-wise

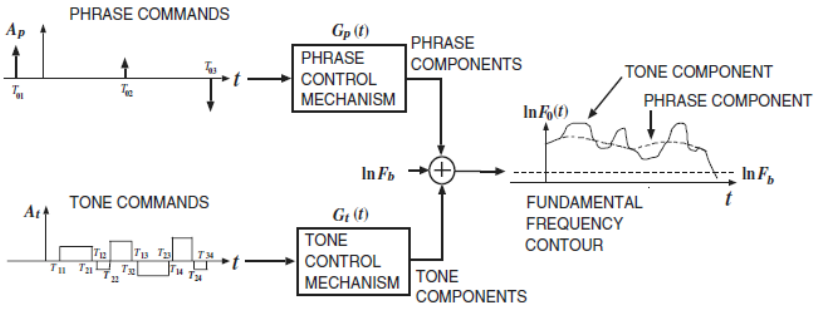



**Figure 3.1:** A command–response model for the process of F0 contour generation for Standard Chinese. [Adopted from [40]]. Ap represents the phrase command magnitude and At tone command amplitude; t is in seconds.

phrase commands and step-wise tone commands as shown in Figure 3.1 [40], which shows the command-response model for the process of F0 contour generation for Standard Chinese. The Chinese language, unlike Sesotho, has four tones, and these are of positive and negative polarity in the Fujisaki model, as depicted in the figure.

In this model, the phrase commands are assumed to be impulses applied to the phrase control mechanism to generate the phrase components, while the tone commands are assumed to be positive stepwise functions applied to the accent control mechanism to generate the tone components. Both mechanisms are assumed to be critically damped second-order linear systems, and the sum of their outputs, i.e., the phrase components and tone components, is superposed on a baseline value (ln Fb) to form an F0 contour, as given by the following equation:

$$\ln F_0(t) = \ln F_\mathrm{b} + \sum_{i=1}^{I} A_{\mathrm{p}i} G_{\mathrm{p}i}(t - T_{0i}) + \sum_{j=1}^{J} A_{\mathrm{t}1j} \left\{ G_{\mathrm{t}j}(t - T_{1j}) - G_{\mathrm{tj}}(t - T_{2j}) \right\} + A_{\mathrm{t}2j} \left\{ G_{\mathrm{tj}}(t - T_{2j}) - G_{\mathrm{t}j}(t - T_{3j}) \right\}, \tag{3.1}$$

where

$$G_{\mathrm{p}i}(t) = \left\{ {\alpha_i}^2 t \exp(-\alpha_i t), \right\}, \qquad t \geq 0, \tag{3.2}$$

$$G_{\mathrm{p}i}(t) = 0, \qquad t < 0, \tag{3.3}$$

and

$$G_{\mathrm{t}j}(t) = \{\min[1 - (1 + \beta_j t) \exp(-\beta_j t), \gamma], \}, \qquad t \geq 0, \tag{3.4}$$

$$G_{tj}(t) = 0, \qquad t < 0, \tag{3.5}$$

where $G_{\mathrm{p}i}(t)$ represents the impulse response function of the phrase control mechanism and $G_{\mathrm{t}j}(t)$ represents the step response function of the tone control mechanism. The symbols in these equations indicate:

$F_\mathrm{b}$ : baseline value of fundamental frequency upon which all the phrase and tone components are superposed to form an F0 contour,

I : number of phrase commands,

J : number of syllables,

$A_{\mathrm{p}i}$ : magnitude of the $i$th phrase command,

$A_{\mathrm{t}1j}$ : amplitude of the first tone command in the $j$th syllable,

$A_{\mathrm{t}2j}$ : amplitude of the second tone command in the $j$th syllable,

$T_{0i}$ : instant occurrence of the $i$th phrase command,

$T_{1j}$: onset of the first tone command in the $j$th syllable,

$T_{2j}$: end of the first tone command (and onset of the second tone command iff the second tone command exists) in the $j$th syllable,

$T_{3j}$: end of the second tone command (and onset of the third tone command iff the third tone command exists) in the $j$th syllable,

$\alpha_i$ : natural angular frequency of the phrase control mechanism to the $i$th phrase command, set empirically at $3/2\Pi$ Hz,

$\beta_j$: natural angular frequency of the tone control mechanism to the $j$th tone command, set empirically at $20/2\Pi$ Hz, and

$\gamma$: relative ceiling level of the tone components, set empirically at 0.9.

## 3.3 Fujisaki Model-based Analysis

Analysis of the Fujisaki model-based analysis considers extraction of the model's parameters. This process, which is the inverse of F0 generation, was automated by Mixdorff [74] to determine the Fujisaki parameters which best model the F0 contour. However, the representation of the F0 contour by phrase and tone commands is not unique. In fact, the F0 contour can be approximated with arbitrary accuracy if an arbitrary number of commands is allowed [2]. Therefore, there is always a trade-off between minimising the approximation error and obtaining a set of linguistically meaningful commands. This process begins with the estimation of the model's parameters and ends with the extraction and editing of the F0 contours. A detailed description follows.

### 3.3.1 Fujisaki Model Parameter Estimation

The Fujisaki model parameter estimation is a multi-stage approach consisting of quadratic spline smoothing, contour filtering, tone command initialisation and a three-pass Analysis-by-Synthesis procedure [73].

#### Quadratic Spline Stylisation

Prior to modelling a given F0 contour, two tasks are performed: 1) Intermediate F0 values for unvoiced speech segments and short pauses are interpolated from the extracted F0 contour, 2) Microprosodic variations caused by the influence of individual speech sounds (explosion, frication, etc.) are smoothed out, as the Fujisaki model explicitly deals with macroprosody only [73]. One method successfully applied to the two tasks is the MOMEL model [45] which converts a given F0 contour into a sequence of target points used as a reference for performing a spline interpolation of the contour. It has been shown that MOMEL can be applied regardless of the particular language [73].

### High-Pass Filtering and Component Separation

In order to separate the tone component from the phrase component and Fb, the spline contour is passed through a high-pass filter with a stop frequency at 0.5 Hz. The output of the high-pass (henceforth called 'high frequency contour' or HFC) is subtracted from the spline contour yielding a 'low frequency contour' (LFC) which contains the sum of phrase component and Fb. The latter is initially set to the overall minimum of the LFC. Hence, partial contours roughly corresponding to phrase and tone components are determined.

### Command Initialisation

The initialisation procedure makes use of the characteristics of phrase and tone command responses making up phrase and tone components, respectively. In a sequence of phrase commands, the onset of a new command is characterised by a local minimum in the phrase component. Consequently, the LFC is searched for local minima, applying a minimum distance threshold of 1 s between consecutive phrase commands. To initialise the magnitude value Ap assigned to each phrase command the part of the LFC after the potential onset time T0 of a phrase command is searched for the closest local maximum. Ap is then calculated in proportion to the frequency value found at this point. As responses of several phrase commands may add up in the phrase component, contributions of preceding commands must be taken into account when calculating Ap, which is reduced accordingly. A full phrase command reset occurs at inter-phrase boundaries accompanied by a longer pause ($> 500$ ms). The time constant $\alpha$ was set to $0.90/s$ for our Sesotho corpus, a value found appropriate after a series of preliminary trials.

To initialise the appropriate number, onset times T1 and offset times T2 of tone commands, the HFC is searched for local minima, whose vicinity ($\pm 100$ ms) is scanned for even lower F0 values in order to avoid picking saddle points. Two subsequent local minima each are associated with a new tone command. Since the tone command response requires some time to decay to zero after T2, T2 is set back to 200 ms before the local minimum. The tone command time constant $\beta$ is set to an initial value of $20/s$. To initialise the tone command amplitude At, the maximum in the HFC between T1 and T2 is determined, and At is set in proportion to the frequency value found at this point. Tone commands are not continued across major pauses in the speech signal.

### Analysis-by-Synthesis

The analyis by synthesis procedure is performed in three steps, in the course of which the initial parameter configuration is subsequently optimised by applying a hill-climb search for reducing the overall mean-square error in the log F0 domain. Each step terminates when the improvement between subsequent iterations drops below a set threshold. At the first step, phrase and tone components are optimised separately, taking the LFC and HFC, respectively, as the targets. Next, phrase component, tone component and Fb are optimised jointly, taking the spline contour itself as the target. In the final step, the parameter configuration is further fine-tuned by making use of a weighted representation of the extracted original F0 contour. The weighting factor applied is the product of degree of voicing and frame intensity for every F0 value, which favours 'reliable' portions of the contour. Before longer inter-phrase pauses, the stylisation algorithm occasionally levels short rising sections of the F0 contour belonging to boundary tones, an error from which the analysis procedure cannot recover. Similarly, the procedure cannot recover if too few commands were set up in the initialisation phase. In some rare cases, a missing phrase command was hence compensated by a very small but long accent command. Superfluous commands

can usually be identified by their rather short durations ($< 50$ ms) and small amplitudes ($< 0.1$) and are removed between analysis steps.

The database was then examined manually for tone commands which could not be justified by the vicinity of high-tone syllables or boundary tones. These were deleted and the analysis procedure was resumed with the last stage which follows.

### 3.3.2 Extracting and Editing F0 Contours

For our corpus, the F0 values were extracted using Praat [12] at a step of 10 ms and inspected for errors. The F0 tracks were subsequently decomposed into their Fujisaki components. Adopting this rationale, automatically calculated parameters were viewed in the FujiParaEditor [77] and corrected when necessary. Manual editing was performed in two cases: (1) F0 contour extraction errors/F0 perturbations (creaky voice) that lead to additional and incorrect tone commands, and (2) minor phrase commands undetected by the automatic analysis resulting in prolonged sequences of low-amplitude tone commands.

## 3.4 Other Prosodic Modelling Tools

The Fujisaki model is considered to be the best of the superpositional F0 contour modelling tools [9]. As mentioned earlier, this model gives a physiological interpretation of the larynx movement, thus resulting in the production of a human-like pitch (F0 contour). This aspect has not been realised in other intonation modelling tools [99]. The modelling approach by the Fujisaki tool is therefore the most appropriate for high-quality speech synthesis. Other modelling tools are presented below [9].

#### Pierrehumbert's Theory of Intonation

The Pierrehumbert theory of intonation [89], developed by Janet Pierrehumbert in 1980 as part of her doctoral dissertation, is a phonological model of intonation. It is based on autosegmental-metrical (AM) phonology [42, 63]. In keeping with the AM theory, the Pierrehumbert model considers intonation to be a sequence of high (H) and low (L) tones. The H and L are in phonological opposition, i.e. the difference in sound between them serves to distinguish intonational meaning. The two types of tones never interact with each other, rather they follow each other sequentially in an utterance.

The H and L tones are the building blocks of three larger tone units: pitch accents, phrase accents, and boundary tones. Pitch accents mark prominence. They are either single tones (H*, L*), or pairs of tones (L+H*, L*+H, H+L*, H*+L); the symbol '*' donotes the alignment of the tone with a stressed syllable. One or more pitch accents comprise an intermediate phrase. One or more intermediate phrases comprise an intonational phrase, the largest prosodic unit posited by this theory. The edges of the intonational phrase are marked by boundary tones. The boundary tones are single tones (%H, %L, H%, L%); the symbol '%' denotes the alignment of the boundary tone with the pitch onset or offset of the intonational phrase. Pitch movement between a pitch accent and a boundary tone is indicated by a phrase accent *H-, L-, denoted by the discritic, -.

To ensure that the model renders well-formed intonational representations, Pierrehumbert defined a finite state grammar that specifies the combinations in which pitch accents, phrase accents, and boundary tones can occur. She also devised a set of phonetic realisation rules to produce F0 contours from the phonological model of intonation described above.

### ToBI-Based Approaches

ToBI stands for tones and break indices. Based on Pierrehumbert's theory of intonation, it was developed in four research meetings between 1991 and 1994 as a standard for describing American English intonation. It has since been extended to transcribe other languages and dialects [16, 92].

ToBI consists of three parallel labelling tiers. The first tier is the tone tier. The tones specified by Pierrehumbert's theory are labelled in the tone tier. The second tier is the break index tier. In the break index tier, break indices, ranging from 0 to 4, are marked. Break indices mark the boundary strength between adjacent words; 0 indicates no boundary, 3 indicates an intermediate phrase boundary (- in Pierrehumbert's model), and 4 indicates an intonational phrase boundary (% in Pierrehumbert's model). The third tier is a miscellaneous tier, where hesitations, disfluencies, laughter, non-speech sounds, etc., are labelled.

It is important to note that ToBI is a labelling system and it does not specify the means to produce quantitative intonation from the ToBI labels. However, there are both rule-based and statistically trained approaches that can be applied to the ToBI labels to generate F0 contours. An example of the rule-based approach is Jilka's handcrafted rule system for specifying the F0 contour of American English from ToBI labels [50]. Jilka's approach is similar to Pierrehumbert's phonetic realisation rules; rules specify the target F0 values associated with ToBI labels, depending on pitch range and the voiced part of the syllable. The target F0 values are calculated from left to right, taking into account only preceding ToBI labels, not subsequent ones. The F0 contour is produced by linear interpolation between target points.

An example of the statistically trained approach is Black and Hunt's linear-regression-based approach for generating F0 contours from ToBI labels [11]. This approach simply involves predicting three target F0 values for every syllable, one at the start of the syllable, one at mid-vowel position, and one at the end of the syllable, by means of linear regression. The prediction formula is defined as,

$$F0 = I + s_1 f_1 + w_2 f_2 + w_3 f_3 + ... + w_n f_n \tag{3.6}$$

The $f_i$ variables indicate the features that contribute to the F0 value of a syllable, such as ToBI label associated with the syllable, syllable position in the phrase, syllable stress, etc. The parameters $I$ and $w_i$ are estimated by linear regression.

### The RNN Intonation Model

The recurrent neural network (RNN) intonation model, developed by Traber in 1991 [107, 108], uses neural networks to predict the F0 contour. A neural network can be considered to be a nonlinear statistical model with many parameters. These parameters are estimated in the training phase so that they can return an optimal set of outputs from the corresponding input. The neural-network-based approach was motivated by the goal of using minimal human effort to obtain high-quality intonation, i.e. while humans would specify which phonological units were relevant for the phonetic realisation of intonation, clever machine learning techniques would figure out how the phonological units map to the F0 contour.

### The Tilt Intonation Model

The tilt intonation model, developed by Taylor and Black [101] at the Center for Speech Technology Research of the University of Edinburgh, considers intonation to be a sequence of intonational events, which are parameterised by tilt parameters. The model posits four basic types of intonational events:

pitch accents, boundary tones, connections (regions in the F0 contour between two pitch accents, two boundary tones, or a pitch accent and a boundary tone), and silence.

Pitch accents and boundary tones are each modelled by piecewise combinations of quadratic functions; these quadratic functions may be rising or falling. Connections are modelled by straight-line interpolations. The amplitude and duration of the rising and falling quadratic functions, the position of the associated intonational event in the time-F0 plane, together with a tilt value associated with each event constitutes the set of tilt parameters associated with the intonational events. The tilt parameter represents the amount of rise and fall of each accent. The tilt value is a difference in the amplitudes of the rise and fall functions, divided by their sum as shown below:

$$tilt = \frac{|Amp_{rise}| - |Amp_{fall}|}{|Amp_{rise}| + |Amp_{fall}|} \tag{3.7}$$

The tilt value ranges from -1 to 1, where -1 indicates a pure fall, 1 indicates a pure rise and 0 indicates a rise followed by a fall of equal magnitude. Thus, the tilt model uses continuous parameters rather than imposing categorical classification on the intonational events.

Dusterhoff and Black [24] have shown that the tilt model can be successfully used to predict F0 contours in a text-to-speech system. The tilt-based F0 generation process has two stages: a training stage and a testing stage. The training stage requires a training database. The database is labelled with tilt events, either automatically or by hand. For each syllable in the database marked with a tilt event, a set of linguistic-prosodic features are extracted. The features are grouped into separate training sets depending on event type. A CART [13] training algorithm is applied to each of the training sets to develop a decision tree for every tilt parameter. The decision trees thus describe the tilt parameters in terms of an optimal subset of the extracted features.

The training stage described above is performed offline. The tilt parameter descriptions obtained in the training stage are used in the testing stage for F0 contour generation of given text. The given text is labelled with tilt events, and the same set of linguistic-prosodic features (as in the training set) are extracted. The related tilt parameters are calculated from the extracted features using the descriptions obtained from training. The tilt parameters are then plugged into predetermined quadratic or linear functions to model the pitch accents and boundary tones, or connections respectively.

### The IPO Approach

The IPO approach [18, 98] was developed at the Institute of Perception Research (IPO) in Eindhoven, the Netherlands in the 1960s. It was originally used to model Dutch intonation. The IPO model is often classified as a perceptual intonation model because of the assumptions underlying the model. These assumptions are:

1. Not all changes in the F0 are perceived by the human ear.
2. Only F0 changes that are perceived by the human ear need to be modelled.
3. The human ear perceives tone variations (rise versus fall) and not tone intensities (high versus low).

Given these assumptions, the IPO approach models the raw F0 contour as a piecewise linear approximation of the original contour, known as a close copy contour. It is called a close copy contour because, upon resynthesis, it is perceptually indistinguishable from the original F0 contour. Generating the close

copy contour also includes specifying the declination line (a line that represents the overall downward trend of the F0 contour).

Close copy contours are classified into discrete, phonetically defined types of F0 rises and falls. The classification parameters describe the deviation of the close copy from the declination line, and include descriptive factors such as its height and slope relative to the declination line, its timing in relation to the span of the declination line, its timing in relation to associated syllable(s) duration, its rate of change, etc. The particular parameters used for classification differ from language to language.

Once an inventory of F0 rises and falls covering the entire combinatorial space of the classification parameters has been collected, a grammar specifying the possible and permissible combinations of the F0 rises and falls is written in terms of the parameters. When the IPO intonation model is used in speech synthesis systems, this grammar is used for F0 contour generation. F0 contours predicted by this grammar must be perceptually equivalent to, and as acceptable as natural F0 contours. The IPO model has been implemented in speech synthesis systems for Dutch [103], English [115], and German [109].

### The Linear Alignment Model

The linear alignment model is another example of a superpositional model. It was developed by van Santen and Mobius [112] at Bell Laboratories. The distinguishing characteristic of this model is that it pays particular attention to the alignment between the pitch contour and the segmental stream underlying it.

Its concern with alignment is most effectively expressed in its modelling of the accent curve component. The accent component represents the same aspects of the F0 contour as the accent component in the Fujisaki model, though it is modelled differently in this model. An accent curve is modelled by parameterised time warps of an accent curve template. The template can be defined as a sequence of anchor values, $Tp = \{P_1, P_2, ..., P_n\}$, that describe the archetypical shape of the associated accent curve type, P. Also associated with P is an alignment parameter matrix, an ensembler of regression weights that describe the alignment of P to the segmental region underlying it. All accent curves of type P have the same template and the same alignment parameter matrix. They differ from each other only in terms of their duration.

Besides the accent curve, two other components of additive F0 contour are specified by the linear alignment model: the phrase curve and the segmental perturbation curve. As in the Fujisaki model, the phrase curve illustrates the long-term shape of the F0 contour. The phrase curve is modelled by a piecewise linear function. The segmental perturbation curve described the segmental influences on the pitch contour such as pitch increase in vowels following voiceless plosives, and pitch lowering in nasals and glides. The segmental perturbation curves are modelled by exponential decay functions.

The linear alignment model has been used for generating intonation in the Bell Labs multilingual text-to-speech system [113]. To synthesise speech from text, each of the three components of F0 specified by the linear alignment model have to be related to linguistic entities. The phrase curve is anchored at three points: the start of the utterance, the start of the syllable that carries the nuclear pitch accent, and the end of the utterance. The accent curve is tied to a left-headed foot. A left-headed foot is defined as a sequence consisting of an accented syllable followed by all unaccented syllables that precede the next accented syllable or a phrase boundary. The degree of emphasis at a particular foot is obtained by multiplying the accent curve by a height factor. The phrase curve is tied to the intonational phrase. Minor and major phrase breaks are distinguished where major phrase breaks occur at sentence ends. Segmental perturbation curves are anchored at vowel onset. The amplitude of this function is determined by the

broad class of the onset consonant; it has a maximal value of voiceless consonants, a smaller value for voiced obstruents, and a zero value for sonorants.

### The SFC Model

The superposition of functional contours (SFC) model of intonation was developed at the Institute for Speech Communication. It was proposed by Auberge [4] and implemented by Bailly and Holm [46]. Like other superpositional models, the principal assumption of the SFC model is that the pitch contour is obtained by a superposition of simpler contours. In case of the SFC model, the simpler contours are multiparametric contours called functional contours (FCs). Functional contours form the core of the distinguishing assumption of this model. They are assumed to directly encode specific metalinguistic functions tied to various discourse units, without any intermediate representation [7]. Metalinguistic functions refer to intonation functions that delimit phonological units and convey propositional and interactional information about these units within the discourse. Examples of metalinguistic functions are hierarchy, segmentation, emphasis, and speaker attitude.

Every functional contour has the following three properties:

1. It is function-specific, i.e., tied to a particular metalinguistic function.
2. It spans the extent of the unit(s) tied to the function it encodes. This extent is called the scope or domain of the FC.
3. The FC shape is a function of the metalinguistic function it encodes and its scope. However, it is important to note that the FC shape is not specified a priori in this model, rather it emerges in the training phase of the model's implementation.

The SFC model has been implemented for pitch prediction in TTS systems for German [6], Spanish [5], and Mandarin Chinese [17]. The metalinguistic functions that will be encoded by the functional contours are defined. One contour generator per function is implemented as a neural network. Each contour generator generates a family of functional contours that encode the same metalinguistic function and hence have the same shape, differing only in terms of their time domains. The input to each contour generator is information relating to the scope of the associated function, and the position of each syllable within the scope. The output is four output parameters per syllable (three F0 values and a lengthening factor).

Training the contour generators to generate a particular pattern of functional contour is not a straightforward process as recovering the unique contributions of the contour generators to their sum (i.e. the F0 contour) is an ill-posed problem. To determine the individual contributions of each contour generator and the particular pattern of the FC, an analysis-by-synthesis loop [46] is used.

### The Kiel Intonation Model (KIM)

The Kiel intonation model was developed by Kohler and his colleagues [56, 58] to model intonation patterns in German. In KIM, the F0 contour is modelled as a sequence of global intonational units, each linked to one emphasized word. The global intonational units are considered to be produced and perceived holistically, and cannot be split. These global units are either peaks or valleys or peak-valley combinations, and differ from each other in terms of their pragmatic, semantic, syntactic, and meaning functions. They were determined in KIM by means of function-oriented phonetic experiments. KIM postulates that there is a protypical intonational unit associated with a particular pragmatic-semantic-syntactic function combination. However, KIM does not ignore the microprosodic phenomena (e.g., F0

shifts at obstruent-vowel boundaries, F0 change in nasals and glides; effect of intrinsic pitch are also included in this category) observed in the F0 contour; it is also incorporated in the model.

Since KIM was developed with a focus on TTS synthesis, the F0 prediction rules are well specified. KIM applies two sets of rules, namely symbolic feature rules and parametric rules, for pitch prediction in TTS systems. The symbolic feature rules are applied to phonological units which have been annotated with syntactic, pragmatic and semantic markers. The phonological units are either segmental (vowels and consonants) or nonsegmental (morphological and phrase boundaries). The symbolic feature rules output the global intonational units associated with the phonological units, encoded as binary features (such as +/-terminal, +/-valley, +/-quest, +/-early, +/-late). These feature values are then used by the parametric rules to generate the F0 contour of the target utterance. The parametric rules include rules for aligning the global intonational units with the segmental structure of the target utterance, downstepping of accent peaks, speech rate, prosodic boundaries and, finally, articulation-induced microprosody [57].

### The STEM-ML Intonation Model

Soft template markup language (STEM-ML) is a physiological model of intonation. It was developed by Shih and Kochanski in 2000 [53] to investigate the deviation of Mandarin Chinese tones from their expected canonical shape when occurring in natural sentences. However, this model has been designed to be language independent, and thus can be applied to non-tone languages such as English.

STEM-ML is founded on three key assumptions.

1. Human speech is pre-planned several syllables in advance.

2. Humans produce speech that optimally balances physiological effort required to speak against unambiguity of the spoken message. The speaker expends maximal effort to produce correct prosody at prosodically crucial events because the cost of ambiguity is high at these points. However, he minimises effort between such events because the cost of ambiguity is low.

3. Speech prosody is continuous and smooth over short time periods.

STEM-ML includes a tagging system (see [54] for a complete description of the tag set) for intonation markup and specification, and a quantitative model to generate the F0 contour. Two important building blocks of the STEM-ML model are parameters and soft templates. The parameters are associated with the tags in the tagging system. The soft templates are a part of the quantitative intonation generation model. The parameters and the soft templates together generate the F0 contour.

In this model, the F0 contour is considered a concatenation of the local accents. The local accents are represented by the soft templates. The soft templates are soft in the sense that the accent templates allow substantial distortion caused by neighbouring accents. The concept of soft templates arises from the previously stated pre-planning assumption. An accent template is affected by past as well as future templates. The degree of distortion is controlled by a parameter called strength. The strength parameter reflects the cost of ambiguity in the previously stated assumption regarding optimally balanced speech. So, if strength (hence cost of ambiguity) is large, the template shape remains unchanged to reflect maximal articulatory effort, whereas if it is low, the accent shape is compromised to reflect minimal articulatory effort.

Besides local tags that control local accent shapes, there are global tags that control speaker-specific information. Thus, a STEM-ML model is built on a particular speech corpus. The implementation of the STEM-ML involves two phases: the learning phase and the generation phase. In the learning phase, the values of the parameters are determined iteratively by minimising the differences between the actual F0

of every STEM-ML tagged utterance in the corpus and the F0 predicted by the model. In the generation phase, when faced with a target utterance, the model first tags the text underlying the utterance, then uses the predetermined values of the parameters associated with the tags to modify the soft templates, and finally, concatenates the modified accent templates to produce the F0 contour.

## 3.5 Summary

This chapter has presented a detailed description of the Fujisaki model and its analysis process. It was also shown that the model, though initially developed for the Japanese language, has proven its efficiency for other languages, tonal and non-tonal. Other F0 contour modelling tools were also highlighted and their different intonation processing methods described. The chapters which follow (Chapters 4 - 7) are experiments in which the Fujisaki model was deployed for the Sesotho language.

# Chapter 4

# Preliminary Experiments

## 4.1 Introduction

As mentioned in Chapter 2, Sesotho is a tonal language with two contrasting tones, high (H) and low (L). In order to quantify the tonal alignments and magnitudes of excursions in Sesotho, the fundamental frequency (F0) contours are parameterised using the Fujisaki model [30]. In earlier studies, the model was applied to F0 contours of Asian tonal languages such as Chinese [72], Vietnamese [23] and Thai [80]. These and other tonal languages examined in other research required tone commands of negative polarity in order to model low, falling and rising tones. The two experiments that follow were performed to examine: 1) whether this was the case for Sesotho, and 2) the intonation of sentences and questions, and the effect of their prosodic manipulations using the Fujisaki model.

## 4.2 Realisation of Sesotho Tone Using the Fujisaki Model

### 4.2.1 Aim

The aim of the current experiment was to investigate an acoustic realisation of Sesotho tone using the Fujisaki model. The main acoustic correlate of tone is pitch, measured as F0 [111]. F0 contours of a set of recorded minimal pairs were thus analysed using the Fujisaki model.

### 4.2.2 Speech Material and Method of Analysis

A corpus of tonal minimal pairs was created. Due to the small number of examples provided in the literature (in [21] for instance), the author, a native speaker of Sesotho, augmented these with her own minimal pairs, yielding a total of 14. The corpus was recorded in a professional recording studio using two female and four male native speakers of Sesotho from Lesotho. All subjects had a university education, two at undergraduate, three at postgraduate, and one at post-doctoral level. The recording was performed using two strategies: reading and repeating.

The first strategy involved subjects reading Sesotho text in random order, from slides presented on a computer screen. Each slide displayed Sesotho text and its English translation which was intended to guide subjects at choosing the right tone for the critical words in the Sesotho text. The repeating strategy involved the same subjects speaking the Sesotho utterances after the author, where the author's recordings were used as a reference. Since there was also interest in examining the interaction between

**Table 4.1:** List of minimal pairs showing the critical words, their positions in the utterance, their respective English translations, expected tones (*exp. tone*), observed tones (*obs. tone*), and vowel differences (if any) in the nucleus of the first syllable. The means and standard deviations of the amplitude (*At*) and time of the underlying tone commands are displayed (*T1 rel* and *T2 rel*), the latter given relative to the beginning of the first syllable of most critical words (underlined), and, with respect to the second for *lehata* and third syllable for *lehare*, *bobatsi*, respectively. The (timing) differences that set apart the tonal assignments of the two partners in a pair are set in boldface.

| word | position | translation | exp. tone | obs. tone | vowel | At | T1 rel [ms] | T2 rel [ms] |
|---|---|---|---|---|---|---|---|---|
| hlola | final | created | LL | HL | [O] | 0.21/0.08 | -16/32 | 194/43 |
| | | conquered | HH | HL | [o] | 0.19/0.08 | 27/56 | 269/69 |
| seba | final | gossip | HH | HL | [e] | 0.15/0.09 | -195/263 | **208/115** |
| | | do mischief | LL | LL | [e] | 0.21/0.06 | -451/173 | **21/65** |
| pota | medial | coming over | LL | HH | [O] | 0.18/0.08 | -268/211 | 300/67 |
| | | talking crap | HH | HH | [o] | 0.14/0.04 | -232/163 | 222/110 |
| tena | final | getting dressed | HH | HL | [e] | 0.21/0.14 | **-134/111** | **56/93** |
| | | annoying | LL | LL | [e] | 0.28/0.14 | **-303/53** | **-1/34** |
| bolla | medial | was circumcised | HHH | HHH | [o] | 0.22/0.06 | -155/212 | **309/84** |
| | | decayed | LLL | LLL | [O] | 0.25/0.07 | -521/131 | **8/107** |
| hlopha | medial | prepares | LL | HL | [O] | 0.27/0.07 | -126/36 | 273/47 |
| | | torments | HH | HL | [o] | 0.31/0.11 | -66/50 | 230/51 |
| bona | initial | them | LL | LH | [o] | 0.32/0.10 | **216/61** | **486/208** |
| | | see | HH | HL | [o] | 0.24/0.13 | **-22/25** | **175/71** |
| bopa | medial | sulked | LL | LL | [O] | 0.20/0.07 | -515/181 | **-177/117** |
| | | molded | HH | HH | [o] | 0.16/0.04 | -271/282 | **328/221** |
| lapa | final | patched (it) | HH | HL | [a] | 0.13/0.04 | **-155/196** | **261/98** |
| | | became hungry | LL | LL | [a] | 0.28/0.08 | **-362/78** | **-73/22** |
| ts'ela | final | poured | LL | HL | [E] | 0.19/0.07 | -17/41 | 197/58 |
| | | crossed | HH | HL | [e] | 0.18/0.06 | -12/45 | 281/71 |
| hloma | medial | acted | HH | HH | [o] | 0.20/0.06 | **-13/19** | 394/109 |
| | | planted | LL | LH | [o] | 0.29/0.02 | **229/48** | 481/246 |
| lehare | final | razor | LHH | LLL | [e] | **all low tone command** | | |
| | | middle | LLL | LLH | [e] | 0.32/0.16 | **-8/60** | **191/33** |
| bobatsi | final | nettle | LHL | LLL | [o] | **all low tone command** | | |
| | | width | LLH | LLH | [o] | 0.40/0.15 | **-2/44** | **231/87** |
| lehata | medial | liar | LLL | HLL | [e] | 0.16/0.05 | **287/72** | 745/108 |
| | | skull | LHH | HHH | [e] | 0.16/0.05 | **-6/54** | 624/286 |

syllabic tones and the intonation of questions, one of the partners in five of the minimal pairs was elicited in the interrogative mode. All minimal pairs are listed in Table 4.1.

Initial auditory analysis of the utterances revealed considerable mismatches between the intended tones and those actually produced. The error rate for reading amounted to 23.6%, whereas that for the repeating task was 7.4%. Therefore, only utterances from the repeating task were admitted to the acoustic analysis, together with the author's sample utterances, yielding a total of 200 tokens. F0 values were extracted using the standard method of Praat [12] at a step of 10 ms and inspected for errors. The F0 tracks were subsequently decomposed into Fujisaki commands applying an automatic method originally developed for German [74]. The decomposition revealed that the low tones in the critical words could be modelled with sufficient accuracy without employing negative tone commands. As a consequence, only high tones were associated with tone commands. Adopting this rationale, automatically calculated parameters were viewed in the FujiParaEditor [77] and corrected as necessary. The utterances were segmented at the word and syllable levels, using the Praat TextgridEditor [12].

**Table 4.2:** List of subjects indicating sex, means ($\mu$) and standard deviations ($\sigma$) of syllabic durations, Fb, and means and standard deviations of At and Ap. HL is the author of the study.

| subject | sex | Fb [Hz] | syll.dur $\mu/\sigma$ [s] | At $\mu/\sigma$ | Ap $\mu/\sigma$ |
|---|---|---|---|---|---|
| HL | F | 140 | 144/88 | 0.19/0.08 | 0.25/0.11 |
| MM | F | 130 | 137/61 | 0.26/0.11 | 0.34/0.11 |
| NR | F | 150 | 162/82 | 0.21/0.09 | 0.28/0.11 |
| BL | M | 85 | 176/88 | 0.29/0.11 | 0.40/0.18 |
| SR | M | 90 | 154/75 | 0.23/0.10 | 0.31/0.13 |
| MS | M | 100 | 149/68 | 0.25/0.09 | 0.50/0.13 |
| KK | M | 95 | 132/61 | 0.17/0.11 | 0.30/0.10 |

### 4.2.3 Results and Analysis

Table 4.2 lists the subjects whose data were considered for this study, their respective values of base frequency (Fb), which was kept constant for all utterances by the same speaker, as well as means and standard deviations of syllabic durations, tone command amplitudes (At), and phrase command magnitudes (Ap). The latter two parameters reflect the pitch range that subjects employ, with At capturing their interval of local tonal transitions, and Ap the amount of declination reset at the onset of a phrase. Since the Fujisaki model is defined in the log F0 domain, values for male and female subjects are in the same range. Figure 4.1 shows examples of the sentence *'O ile a bolla thabeng'* (which has dual meaning - "He was circumcised on the mountain." and "His body decayed on the mountain.") produced by subject MS. Each panel displays from the top to the bottom: the speech waveform, the F0 contour (extracted, represented by '+' signs, and modelled, represented by a solid line), as well as the underlying phrase and tone commands. The syllable boundaries are indicated by the dotted vertical lines. The top and centre panels show *bòlla* and *bólla*, respectively, embedded in statements, and the bottom panel *bólla* embedded in a question. (The acute accent above the vowel or on a syllabic consonant indicates a high tone while the grave accent represents a low tone.) As can be seen, the tone commands extend across several syllables which are hence associated with high tones. The main distinction between *bòlla* and *bólla* is that the tone command underlying *'o ile a'* ends before the segmental onset for *bòlla* whereas it extends to the last syllable for *bólla*.

The succeeding low tones follow the phrase contour to the end of the utterance. It should be noted that in most instances of utterance-final low tones, female subjects produced creaky voices whereas the male subjects exhibited modal vocal fold vibrations. In the case of the question, the underlying phrase command has a much higher amplitude (0.74) compared to that of the statement (0.30) which raises the F0 pattern without changing the tonal assignment for *bólla*. However, the long tone command that was found in the statement is split into two separate commands, one for *'o ile a*' and the other for *bólla*. A comparison with other utterances suggests that this prosodic chunking seems to be an acceptable choice, even for the statement (see Figure 4.2). The three right-most columns of Table 4.1 list averaged Fujisaki model parameters for all critical words. In addition to averaged tone command amplitudes At, the relative timing is given, expressed as the mean distance between the tone command onset time T1 and the offset time T2 and the segmental onset of the syllable marked by underlining in the word in milliseconds (ms), henceforth T1rel and T2rel. Negative timing values therefore indicate tone command onsets and offsets occurring before the onset of the syllable, whereas positive values indicate tone command onsets or offsets after the syllable onset. The tonal transitions distinguishing the two partners in a minimal pair are indicated in boldface. Since high tones are aligned with tone commands, the onset of a tone command at time T1 indicated the transition between a low and a high tone, whereas the offset at time T2 marks

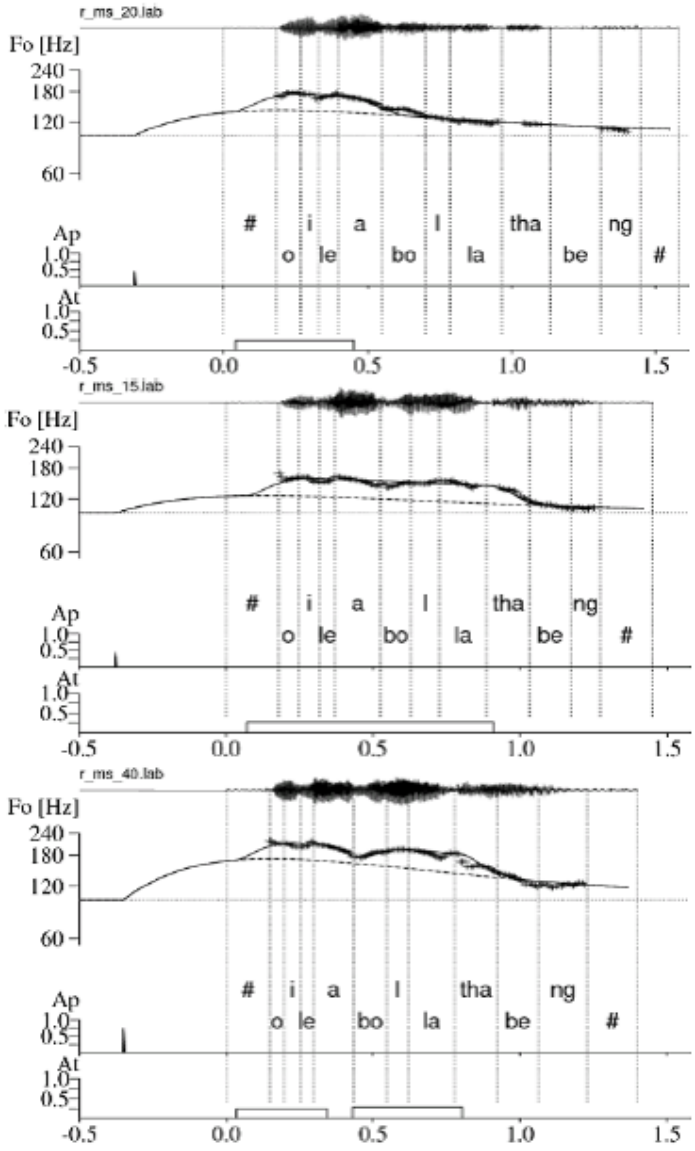



**Figure 4.1:** Examples of analysis of the sentence *'O ile a bolla thabeng'* - "He was circumcised/his body decayed on the mountain", uttered by speaker MS. Panels from the top to the bottom: low tone statement (*bòlla*), high tone statement (*bólla*), high tone question (*bólla*).

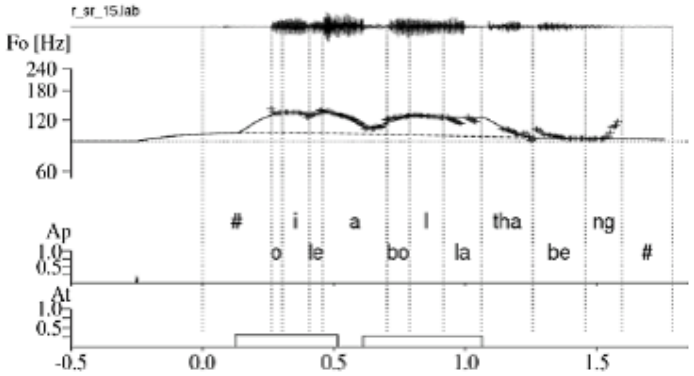



**Figure 4.2:** Analysis of the sentence *'O ile a bólla thabeng'* - "He was circumcised on the mountain", uttered by speaker SR as a statement.

the transition from a high to a low tone. In the case of *bolla*, T1rel is negative, -151 ms for *bólla* and -521 ms for *bòlla*, respectively, indicating that the command starts before the beginning of the word. Since, as described, prosodic chunking might occur, the onset of the tone command can vary considerably in the case of *bólla*, hence also T1rel will vary. This is also reflected by the high standard deviation of 212 ms in contrast to only 131 ms for *bòlla*.

T2 is the time at which the tone command ends. In Figure 4.1, this happens before the onset of the critical word for *bòlla* (T2rel = 8 ms) and afterwards for *bólla* (T2rel = 309 ms). With respect to *bolla*, T2rel is the important timing difference. For the sake of conciseness, other cases will not be discussed in detail, but Table 4.1 referred to instead. The fourth column lists the expected tone for all critical words, and the fifth column the tones that were actually observed. It should be noted that by rule, utterance-final high tones on final syllables of verbs are converted to low tones (see, for instance, *hlola* and *lapa*). In the cases of *hlola*, *pota*, *hlopha*, and *ts'ela*, only vowel differences could be detected. In all of these minimal pairs, the presumed high tone partner exhibits a closed vowel and the low tone one an open vowel in the first syllable. These perceptual observations were also confirmed by formant measurements carried out with Praat [12]. For example, for the high *ts'ela* mean F1/F2 of 364 Hz/217 Hz and for the low *ts'ela* mean F1/F2 of 481 Hz/211 Hz were measured. This suggests that tonal and vowel differences might interact in the perceptual assessment of tones.

Next, the Fujisaki model parameters for the questions and their corresponding statements were compared. At and especially Ap for questions were higher for questions (means/s.d.: 0.24/0.10, 0.42/0.16) than for the statements (0.23/0.09, 0.32/0.15). The literature [21, 121] indicates that questions exhibit compressed penultimate syllables. This was confirmed by the measurements: The mean duration of penultimate syllables was 168 ms for questions as compared to 254 ms for corresponding statements.

Finally, the effect of the position of a tonal transition in an utterance on its interval expressed by At was examined. The correlation between the index of the syllable and At is highly significant ($\rho = -0.258$; $p < 0.01$). This indicates that the F0 range narrows slightly towards the end of the phrase.

### 4.2.4 Discussion

The material examined in this experiment presented a first snapshot of tonal realisations in Sesotho. Since orthography does not reflect tones, the data collection was problematic. The repeating task that yielded a lower error rate can also be questioned as it remains unclear whether subjects identified intended meanings

or simply imitated what they heard. The tonal organisation of Sesotho is different from that found in Asian tone languages where every syllable is assigned a specific tonal target. In Sesotho, only high tone syllables are associated with tone commands and other syllables are either transient or their F0 follows the phrasal contour. In other words: low tones can be interpreted as the absence of tone since they do not require tone commands. This interpretation has also been suggested in [52], for instance. Furthermore, tone commands may extend over several syllables, facilitating prosodic grouping of utterances.

## 4.3 The Effect of Prosody Modification on Minimal Pairs

### 4.3.1 Aim

In the previous experiment it was revealed that perceived tone also appeared to depend on the underlying vowel. The aim of the current experiment therefore, was to investigate to what extent tonal perception interacted with vowel perception to facilitate word identification. Prosodic features which facilitate the perceptual differentiation between statements and questions were examined.

### 4.3.2 Stimulus Design

All stimuli of the minimal pairs from the previous experiment were produced using the resynthesis capability of the FujiParaEditor [77] which replaced the original F0 contour of an utterance by one generated with the Fujisaki model. The actual acoustic resynthesis employs the Praat ManipulationEditor [12] which is based on Pitch Synchronous OverLap and Add (PSOLA). Original utterances had been produced by three male native speakers of Sesotho from Lesotho.

#### 4.3.2.1 Lexical Identification

Table 4.3 lists all minimal pairs of words that were used in the lexical identification task. As shown in the table, *seba*, *tena* and *lehata* only vary with respect to tone, *bolla* varies with respect to tone and vowel and *ts'ela* only varies with respect to the vowel. Interest was in the following research questions for this experiment:

1. What is the minimum duration of a tone command associated with a high tone syllable?
2. What is the minimum amplitude At or amplitude ratio compared to neighbouring high tones required to signal a high tone?
3. Does reducing the tone command amplitude on a high tone syllable lead to perception of the low tone - even without modifying the vowel quality?
4. Does raising the tone command amplitude on a low tone syllable lead to the perception of the high tone – even without modifying the vowel quality?

To answer these questions, the manipulations listed in Table 4.3 were introduced. Manipulated utterances are indicated by normal font and the nature of the manipulation is described. The intended target meanings are written in the bold and italic font and the way in which the source has been manipulated to obtain the target is listed. Note that *bolla* as well as *ts'ela* exhibit vowel differences. Vowel quality was not modified in this study.

Figure 4.3 shows an example of how the stimuli were created. The figure displays, from the top to the bottom: the speech waveform, the extracted (+) and modelled (-) F0 contour, and the underlying

**Table 4.3:** List of manipulations used in the word identification task. The alternative meanings targeted are indicated in italics and boldface. In the case of *bolla*, the modification was applied both ways between the contrasting words.

| word | translation | vowel | tone | modification |
|---|---|---|---|---|
| lehata | skull | [e] | HHH | increase of T1 |
| | ***liar*** | ***[e]*** | ***LLL*** | ***T1 later*** |
| seba | gossip | [e] | HL | variation of T2 |
| | ***do mischief*** | ***[e]*** | ***LL*** | ***T2 earlier*** |
| tena | is getting dressed | [e] | HL | variation of tone command location (both T1 and T2) |
| | ***is annoying*** | ***[e]*** | ***LL*** | ***T1, T2 earlier*** |
| bolla | was circumcised | [o] | HHH | reduction of T2, reduction of At |
| | decayed | [O] | LLL | increase of At |
| ts'ela | crossed | [e] | HL | reduction of At |
| | ***poured*** | ***[E]*** | ***HL*** | ***only vowel difference*** |

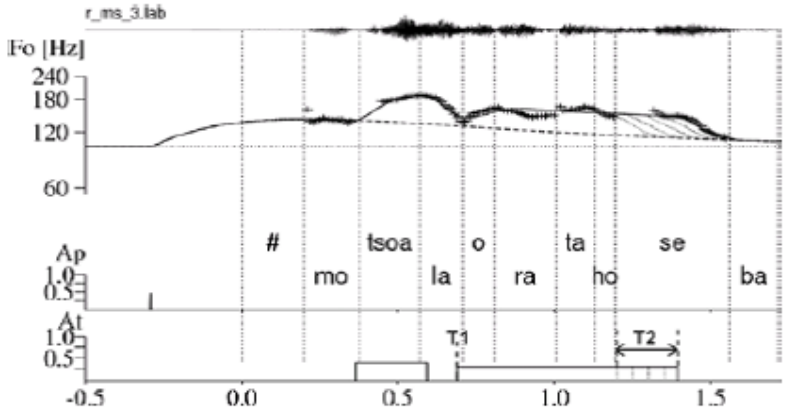



**Figure 4.3:** Illustration of stimulus variation for the utterance '*Motsoala o rata ho seba.*' - "My cousin likes to gossip.". The high tone on *seba* in the original is slowly transformed into a low tone by reducing the tone command offset time T2. The tone command onset time T1 is left constant in this example.

phrase (magnitude Ap) and tone (amplitude At) commands. The syllable boundaries are indicated by the dotted vertical lines. In the utterance, *seba* exhibits an HL pattern. The parameter modified in this example was the offset time T2 of the second tone command. As indicated in the figure, T2 was reduced in four equal steps of 50 ms, yielding earlier falls of the F0 contour on *seba*. With an LL pattern this word means "do mischief". This is the modification in meaning expected to be observed as T2 was reduced.

All ranges of parameter variations to which the stimuli were subjected were based on the previous measurements on natural utterances of the minimal pairs. The total numbers of stimuli for the word identification task was 38.

#### 4.3.2.2 Question vs. Statement

For this task, two utterances of statements were selected and modified with respect to the following parameters:

**Ap:** The phrase command magnitude indicates the amount of F0 reset at the beginning of an utterance. In the natural productions, it was observed that Ap in questions was considerably higher than in

statements, hence raising the onset value of F0 as well as increasing the falling F0 slope across the whole utterance. Starting from the original Ap of the statement, its value was increased in three steps of 0.15 yielding four different onset values and slopes of F0.

**Speech rate:** When comparing the overall speech rate of statements and questions, it was observed that questions were generally spoken faster than statements. Therefore, Praat ManipulationEditor was used to increase the speech rate by 20%.

**Shortening of the penultimate syllable:** Measurements from the previous experiment had shown that the penultimate syllable in questions was considerably shorter than in statements. Therefore, stimuli were created in which the penultimate was shortened to 70% of its original duration. In the utterances with increased speech rate, the penultimate was compressed even further, maintaining the 70% ratio.

All combinations of feature modifications yielded 16 stimuli per sentence, and hence a total of 32 stimuli for the question/statement task. (four levels of Ap, two speech rates, normal and shortened penultimate syllable).

### 4.3.3 Perceptual Evaluation

The experiment was performed at the University of Stellenbosch (SU) as well as the National University of Lesotho (NUL). Stimuli were grouped in randomized sequences for the two sub-experiments and played back on a laptop computer over loudspeakers. The subjects consisted of 15 students of engineering at SU, nine 3rd-year students of linguistics at NUL, and four staff members at NUL. In total there were 17 male and 11 female participants. Each trial in the stimulus sequences consisted of (1) the number of the stimulus, (2) a one-second pause, (3) the stimulus itself, (4) a five-second pause. The judgments on the stimuli were noted down by the subjects on questionnaires. Each sub-experiment was preceded by a warm-up session in which the subjects heard natural stimuli, and contrasts with respect to word difference and statement/question distinction were presented. The correct answers to the warm-up trials were already listed in the questionnaires. Of the re-synthesized stimuli, 35 were presented twice in the experiment in order to test the consistency of the judgments.

### 4.3.4 Results and Analysis

At first, the results from the two groups at SU and NUL were evaluated separately. However, since the correlation of group result means was found to be 0.96, they were pooled for subsequent analysis. The correlation between results of the first and second presentation was 0.90, hence their average was taken for the repeated stimuli.

**Lexical Identification**. Figure 4.4 displays word identification rate (y-axis) in percent as a function of the parameter modified (x-axis) for six of the stimulus sequences. The stimuli are indicated by circles, and the word meaning is that of the stimulus with unmodified F0. This stimulus is located where the parameter modified equals 0.0 on the x-axis. The solid line indicates a third order polynomial approximation of the data. The word meaning gradually changes with the degree of parameter modification. However, in some cases, *bolla* - "circumcise", for instance, the alternative meaning is only identified with a maximum ratio of about 80%. A special case is *seba* - "gossip". As T2 decreases the identification rate of "gossip" drops to 20%, but the two stimuli with the earliest T2 – indicated by filled circles – which were also expected to be associated with the low tone meaning

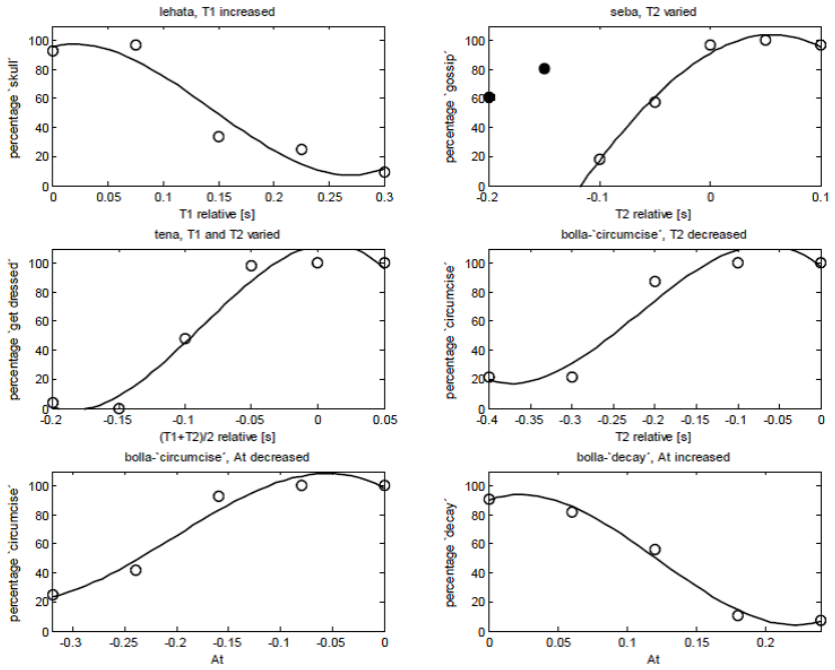



**Figure 4.4:** Results of the lexical identification experiment. Each panel displays the percentage of stimuli (circles) in which words are identified as having unmodified meaning as a function of the parameter being modified. The word meaning is that of the stimulus with unmodified F0. The parameter being modified is marked on the x-axis, with a parameter value of 0.0 indicating an unmodified F0. The solid line indicates a third order polynomial approximation of the data.

of "do mischief" are in contrast and are identified as "gossip" by 61% and 81% of the subjects, respectively. Careful auditory examination of the two stimuli did not reveal anything conspicuous. In fact, the author and a Sesotho linguist, a lecturer at NUL, had classified them as "do mischief" as expected.

When At is decreased in the high tone of the word *bolla* - "circumcise" , the hypothetical 50% boundary between the categories is located at At=-0.24. In contrast, when increasing At on the low of tone *bolla* - "decay", this boundary lies at At=+0.12. These values roughly correspond to -4 and +2 semitones respectively. With respect to temporal alignment, the 90% and 50% thresholds vary with each individual stimulus sequence. For single syllables, these correspond to reductions of T2 by approximately 50 ms and 100 ms, respectively, turning a high tone into a low tone, whereas in the word *bolla* with two high tone syllables these values are around 150 and 240 ms. As expected, lowering At on *ts'ela* - "crossed" does not lead to the perception of "poured", since the two words only differ with respect to the vowel. Even for values of At at or close to 0.0, identification rates remain at 98%.

**Question vs. Statement.** The percentage of stimuli judged as statements as well as the correlation between this value and the factors Ap, speech rate and presence/absence of penultimate shortening was calculated. These correlations are -0.25, -0.36, and -0.83, respectively, but only the two latter values are significant ($p < 0.05$) and highly significant ($p < 0.01$). The stimulus exhibiting the combination of highest Ap, increased speech rate and presence of penultimate lengthening was

identified as a question by 89.3% of subjects. In contrast, the stimulus which was unmodified with respect to F0, speech rate and penultimate shortening was judged as a statement by 97.6% of subjects. Shortening of the penultimate in the aforementioned stimulus reduced this figure to 66.1%. When in turn Ap was increased to the highest level, while all other features were left unchanged, the percentage of stimuli judged to be statements only reduced to 84.8%. When only the speech rate was increased, this figure dropped to 86.9%. This suggests that although all three factors contribute to the identification of an utterance as a question, the shortening of the penultimate is by far the most important.

### 4.3.5 Discussion

This experiment presented a first study of the perception of tone and intonation in Sesotho. All stimuli were produced by resynthesis with the Fujisaki model-based F0 contours. Modified stimuli were obtained by changing the Fujisaki model parameters, as well as the speech rate and the duration of penultimate syllables in the case of the statement/question distinction. With few exceptions, results regarding word identification are in line with our expectations. Reduction of At as well as reduction of T2 for high tone stimuli led to their perception as their low tone counterparts. Increasing At for a low tone word had the opposite effect. F0 modifications even seemed to override vowel differences between words, as was shown for *bolla*, when both affect the word's meaning. In the case of *seba,* the intended low tone stimuli were associated with the high tone meaning, although they were correctly classified by the author and the Sesotho linguist at NUL. It can only be speculated whether some resynthesis artifact or a positional effect (the stimulus with the highest rate was the second in the sequence) led to these observations.

With respect to the statement/question distinction, shortening of the penultimate syllable, higher speech rate, and also increased phrase command magnitude Ap, increase the probability of an utterance being perceived as a question. Of these three modifications, the shortening of the penultimate syllable had the strongest impact on the judgments.

## 4.4 Conclusions

Preliminary experiments were performed in which minimal pairs of words embedded in identical carrier sentences were examined, and tonal alignments and magnitudes of F0 excursions were analysed using the Fujisaki model. Results showed that high tones in Sesotho are associated with the Fujisaki tone commands of positive polarity and low tones with the absence of a tone command, and not negative polarity as in many other tonal languages. With the low tone syllables, the F0 contour either follows the phrase component or vocal fry occurs. It was also observed that some of the contrasting words in the minimal pairs examined differed with respect to the vowel quality of their first syllable. Some showed only vowel differences, and no tonal differences. The vowel differences observed were systematic in that high tones were associated with closed vowels, such as [o], and low tones with open vowels, such as [O] (represented by SAMPA transcription).

In the second set of experiments, it was investigated to what extent tonal perception interacted with vowel perception to facilitate word identification. It was verified that minimal pairs which were distinguished by vowel quality remain distinct once the pitch has been monotonised. Tonal and vowel differences interact in the perceptual assessment of tones. In several instances, tonal minimal pairs (according to dictionary) actually appear to differ in vowel quality and not tone, but this difference is perceived as a tonal difference. Reducing or increasing tone command duration and/or amplitude leads

to a perceived change in meaning in the minimal pairs. With regard to intonation, these modifications can also lead to the distinction between a statement and a question. Questions are distinguished by a higher speech rate, a shorter penultimate syllable, and an increased phrase command magnitude.

# Chapter 5

# The Fujisaki Model and Surface Tone Transcription

## 5.1 Introduction

The preliminary experiments from the previous chapter were essential in the establishment of tonal alignments, F0 excursions and tonal realisations (including intonation for sentences and questions) in Sesotho using the Fujisaki model. The model presented an appropriate mechanism for manipulation of Sesotho speech. The subsequent step was to carry out experiments which would allow us to determine a strategy for prosody modelling for a Sesotho TTS system using the Fujisaki model.

In order to achieve accurate prosodic modelling, prosodic marking of text is crucial. Since orthography is not marked in Sesotho text, the initial step was to compile a speech corpus and find means of marking the text by surface tone transcription. This procedure is detailed in Section 5.2. The Fujisaki model and surface tone transcription are two methods by means of which the tonal pattern of sentences in Sesotho can be determined. It is therefore important to find the relationship between the two. This relationship was investigated and established in Section 5.3. The relationship between the two approaches revealed patterns of prosodic groups which were studied and categorised in detail based on their characteristics in Section 5.4. The duration and amplitude of these prosodic groups were then modified by means of the Fujisaki model in Section 5.5. The modification was performed in a way that would allow easy synthesis from the orthography. The modified stimuli were then subjected to perceptual testing where the naturalness of the output speech was evaluated.

## 5.2 Developing a Corpus for Prosodic Modelling in Sesotho

### 5.2.1 Aim

As a basis for subsequent larger-scale experiments, a tone-marked Sesotho corpus intended for the development of a prosody modelling algorithm for a Sesotho TTS system was compiled.

### 5.2.2 Corpus Compilation

A set of weather forecast bulletins obtained from the weather bureau in Lesotho, Lesotho Meteorological Services (LMS) was used for our corpus. The corpus used is a forecast for 16 consecutive days. The

orthographic transcriptions were compiled using the Lesotho orthographic conventions. The corpus consists of 406 sentences, with an average of 18 words per sentence, which amount to 91 minutes of speech. The data was chosen because, with the goal of TTS in mind, it is efficient to initially focus on a limited domain.

The original data was compiled and broadcast for Lesotho TV but the audio was not of high quality, containing considerable background noise, as well as a large variability in speaking rate. The poor signal-to-noise ratio (SNR) in particular made this data unsuitable for eventual use in TTS development. For this reason, the sentences were re-recorded. This was done by the author, who is a female native speaker of Sesotho. The recording was performed in a quiet studio environment using a large membrane SHURE KSM32SL microphone. All recordings were made at a sampling rate of 48 kHz.

A presentation of a related Sesotho corpus [91, 1] is noted. The sentences of the corpus in [91] were chosen specifically to cover the contexts necessary for the application of the tonal transcription rules due to Khoali [51], whereas those in [1] included English words, and English is not our focal interest. Our somewhat larger corpus, in contrast, had been drawn from broadcast speech and therefore constituted a semantically meaningful passage in a targeted domain, suitable for later use in the development of TTS systems.

### 5.2.3 Surface Tone Transcription

Studies [69, 88] have shown that the pronunciation of lexical items can differ from one syntactic context to another. Words in isolation are pronounced differently from words in context in many languages [3], and therefore, there is a need to perform a surface tone transcription as part of the compilation corpus. The corpus was annotated using surface tonal transcription as described in Section 2.5.3.

In determining the surface tones of the sentences in the corpus, some challenges were faced and these include:

- In order to motivate tone transcriptions, a morphological analysis is necessary and the lexical tone for each word needs to be known. The morphological analysis was carried out by a linguist with a strong tonal background in Tswana[1]. (The Sesotho linguist from NUL was not available when this experiment was conducted.) While Tswana is strongly related to Sesotho, on occasion differences between these two languages complicate the analysis. This highlights that the varieties of the Sotho-Tswana languages (i.e. Sesotho, Northern Sotho, and Tswana) do differ in some aspects of their tonology.

- For some words in our corpus, the lexical tone pattern could not be determined because there was no corresponding tone-marked dictionary entry, or because the word was a proper noun. Since lexical tones for proper nouns and other words were not found in any of our reference dictionaries, [22] and [60], these words were assumed to have an underlying low tone pattern. These words were then recorded individually in Praat [12], afterwhich the tones were determined from the pitch (F0) contour formed by the author and the Tswana linguist.

- According to Zerbian [124], a tone-marked pronunciation dictionary should provide both lexical and grammatical tone. Our pronunciation dictionary lacked the grammatical tone.

The surface tone predicted on the basis of the lexical tones retrieved from tone-marked dictionaries, the grammatical tones, and tonal rules as described in the linguistic literature should eventually be translated

[1]We profoundly thank Prof. Sabine Zerbian.

**Table 5.1:** A summary of the tone-marked sub-corpus from the Sesotho weather forecast.

| | Manually annotated tones |
|---|---|
| Number of sentences | 24 |
| Number of words | 490 |
| Number of syllables | 1096 |
| Number of high surface tones | 478 |
| Number of low surface tones | 618 |
| Number of high perceived tones | 512 |
| Number of low perceived tones | 584 |

into specific tone targets for a TTS system. The expectation is that the resulting pitch contours generated by the TTS should then match a naturally produced utterance in perception and/or acoustic realisation.

In a bid to determine how far the predicted surface tones match the acoustic realisation and/or perception of a naturally produced utterance, a sub-corpus of 24 sentences was selected on which perceived tonal transcription was performed. While it is ongoing work to compare the predicted surface tone with the actually produced pitch by means of a computational model [74], a preliminary comparison between predicted surface tone and perceived tone was explored as described in the section that follows.

### 5.2.4 Perceived Tone Transcription

Tone transcription based on auditory impression is notoriously difficult (see for example [64] for criticism of the method in classical linguistic tone studies and [91] for information on inter-transcriber reliability). While it should be a long-term goal to develop an appropriate methodology for studies in tone perception, for the time being, tone transcription based on auditory impression is what can be relied on. The perceived tones for a subset of 24 sentences were transcribed by the author, accounting for 1096 syllables in total. In selecting the sentences, care was taken to avoid repetition, which is to some degree inherent to the weather-reporting domain. The transcription, in which the label H (high tone) or L (low tone) was assigned to every syllable, was based on auditory impression alone. Note that although the transcriber is a native speaker of Sesotho, native speakers of tone languages are not necessarily aware of their language's tonal system as it is neither taught at school nor indicated in orthography. Some experience with tone and the transcription system used is therefore necessary in order to transcribe tone consistently.

During the transcription of the perceived tones, syllables for which there was no doubt about the tone (H or L) in the transcriber's opinion, were marked as such. Syllables for which the tone was judged difficult to determine, were marked with their most likely tone but also tagged as questionable. For these tones, the F0 contour extracted using Praat [12] was used as an aid in making the H/L decision. After two passes through the data by the annotator, fewer than 0.5% of the syllables fell into the questionable class. These were omitted from the following analysis, although it can be noted that their inclusion would not lead to noticeable changes in the results.

Table 5.1 presents a summary of the transcribed tones, the differing numbers for surface tone transcription and perceived tone transcription. An inspection of the resulting perceived notes showed that these did not always correspond to the predicted surface tone. As an example, Figure 5.1 shows the lexical and surface tone transcription of the Sesotho phrase *'Moea o tla foka o bobebe ho hlaha ...'* which means "A light wind will blow from ...". The F0 contour extracted from the speech is also shown in the figure, and its inspection with increased pitch values. However, there is a disagreement for the second syllable of the word *bobebe*. On the first *be*, the predicted surface tone is high (as indicated by the symbol '^'), yet the F0 contour shows a downward excursion. For the second *be*, the observed F0 contour and

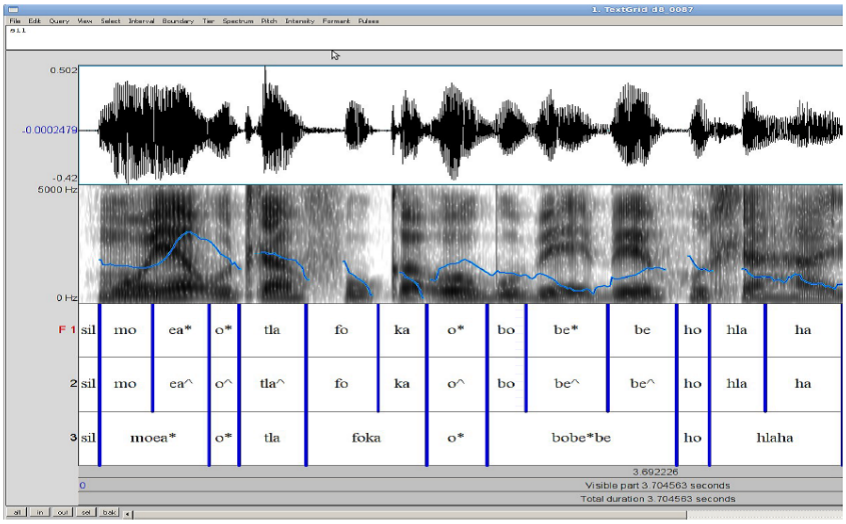



**Figure 5.1:** A sentence annotation using Praat [1] showing the waveform (panel 1), the F0 contour (panel 2), lexical tone (tier 1), surface tone (tier 2), and the orthographic transcription (tier 3). Tiers 1 and 2 are both syllable levels, while tier 3 is at word level. For Praat purposes, lexical high tones were marked with the symbol '*' and surface high tones with the symbol '^'.

**Table 5.2:** Lexical, surface and perceived tones for a selection of phrases drawn from the weather forecast corpus. Discrepancies between surface and perceived tone are indicated in bold.

| Sentence | Lexical tone | Surface tone | Perceived tone |
|---|---|---|---|
| Likhomo ka lapeng | Likhomo ka lapeng | Likhomó ká lapéng | Li**khó**mó ká lapéng |
| Atamelang le ipehe leseling | Atamelang le ipehe leseling | Atámelaŋ́ lé ípéhé léselíng | **Á**támé**lá**ŋ́ lé ípéhé léseé |
| Matsatsi a mabeli a latelang | Matsatsi a mabeli a latelang | Matsatsí á mabéli á látélaŋ́ | **Mátsátsi** á ma**be**li á **late**la**ng** |
| Re tla qala pele ka ho sheba pula | Re tla qala pele ka ho sheba pula | Re tla qála pélé ká hó sheba púla | Re **tlá** qá**lá** pé**le** ká hó sheba **pulá** |
| Eona ea lateloa ke Leribe | Eona ea lateloa ke Leribe | Eoná eá láteloa ké Léribe | Eoná eá láteloa ké Leribe |
| Ka ntlha tse ts'eletseng feeloane supa | Ka ntlha tse ts'eletseng feeloane supa | Ká ntlhá tsé ts'elétseŋ́ feeloane supá | Ká **ńtlha** tsé ts'elé**tsé**ŋ́ feeloane supá |

predicted surface tone agree.

### 5.2.5 Analysis of the Transcriptions

Since the surface tone should reflect the perceived tone, the relationship between the two, and that with lexical tone, was investigated. Table 5.2 depicts the relationship between the lexical, surface, and perceived tones for a selection of sentences, and illustrates cases for which the latter two do not coincide.

Table 5.3 shows a tone match between the surface and perceived tones as a percentage. Only the syllables for which the tones could be assigned with confidence were included in this analysis, although the inclusion of the (few) remaining tones did not lead to noticeably different results. The table shows that, of the syllables marked with a high surface tone, approximately 80% were perceived with a high

**Table 5.3:** The relationship between the surface and perceived tones.

| | % Perceived tone high | % Perceived tone low |
|---|---|---|
| Surface tone high | 79.8 | 19.1 |
| Surface tone low | 18.2 | 81.3 |

**Table 5.4:** The relationship between the lexical and perceived tones.

| | % Perceived tone high | % Perceived tone low |
|---|---|---|
| Lexical tone high | 65.2 | 38.4 |
| Lexical tone low | 38.4 | 61.4 |

tone in the speech data. Similarly, of the syllables marked with a low surface tone, approximately 81% were perceived as low-tone syllables. Table 5.4 illustrates the same comparison but between lexical and perceived tones. When considering the lexical tones, it is seen that only approximately 65% of the high tones are perceived as high tones, and 61% of the lexical low tones are perceived as low. A comparison with Table 5.3 confirms that the surface tones on average correspond more closely to the perceived tones than the lexical tones, as expected.

The source of the discrepancies between the predicted surface tones and the perceived tones might be threefold: it might lie in the predicted tones, in the produced tones and/or in the perceived tones.

For the predicted tones, Section 2.5.1 already brought to light differences in lexical tones provided by tone-marked dictionaries. Other reasons for divergences in predicted surface tone and perceived tone have been mentioned in Section 5.2.3, namely the uncertainty concerning the lexical tones of proper names, uncertainties in morphological analysis, and dialectal differences between Tswana and Sesotho in morphology and tone (cf. [51]:iii]).

For the produced tones, only one speaker was recorded. Given dialectal differences present in any language, there might be single idiosyncratic differences from the Sesotho described in the literature.

Finally, for the perceived tones, tonal transcription is hampered by the fact that there is no standard methodology for tone transcriptions based on auditory impressions. Guidelines are missing that illustrate tonal contrasts in different contexts and that offer exercises in tone marking in order to provide transcribers with relevant training (cf. [70]).

### 5.2.6 Discussion

Sesotho is a tonal language, and in order for future research to consider the modelling of tonal prosody, the corpus has been annotated at the tonal level. In order to achieve this, a set of 406 sentences was recorded from the weather report domain. Surface tones were determined for each word in the corpus, and then a subset of 24 sentences was hand-labelled based on perceived tones. In the subset, 79.8% of the high tones and 81.3% of the low tones were found to agree with the corresponding surface tones. This indicates that there is room for improvement in the tonal rules. The disagreement of syllables (the remaining 20%) could either be due to pronunciation errors in the actual tone during recording, surface tone prediction errors, or perceived tone which has no standard methodology. While the surface tone rules appear to be appropriate, they are not perfect. There appears to be a substantial number of tone phenomena in Sesotho which have not yet been captured by the tonal rules and hence not reflected in the surface tone transcription. Nevertheless, the corpus provides a good basis for subsequent experimentation in the modelling of tone in Sesotho to achieve intelligible and natural speech synthesis.

**Table 5.5:** Syllable match between surface tone and the Fujisaki tone commands.

| | % with tone commands | % without tone commands |
|---|---|---|
| Surface tone high | 71.1 | 27.5 |
| Surface tone low | 28.6 | 69.4 |

## 5.3 Comparing the Fujisaki Model-Based Analysis and Surface Tone Transcription

### 5.3.1 Aim

The previous experiment highlighted the relationship between the surface tone and the perceived tone. In the current experiment, the relationship between the surface tone transcription and the tone commands as determined by the Fujisaki model was investigated and the two approaches compared. These are the two methods by means of which the tonal pattern of sentences in Sesotho can be determined. The objective was to establish the relationship between the two on a broader scale and finer detail. In addition, the Fujisaki tone commands were compared with the perceived tone. Further on in the experiment, the mismatched syllables were explored in order to account for the discrepancies, and particular attention was given to the influences of the adjacent syllables on the tone of the target syllable.

### 5.3.2 Data Preparation

A selection of 53 sentences from the main corpus was used. After deduction of surface tone transcription, the sentences were annotated at word and syllable levels using Praat TextGrid editor [12].

In order to allow a better analysis of the correspondence between the surface tones and the tone commands from the Fujisaki model, each syllable was listened to individually by a human annotator, from within the FujiParaEditor [77], and the tone labels noted. Of the 53 sentences, 24 were annotated in this way, accounting for 1096 syllables in total. This was followed by a manual and visual inspection of the F0 contour produced by the Fujisaki model of each utterance. The F0 excursions were examined and compared with the surface tone prediction. Furthermore, each syllable was listened to individually and resynthesized with the FujiParaEditor [77] for perceptual verification. This was carried out by the author and a linguist trained in Sepedi (Northern Sotho) tone labelling. The aim was to identify all cases where the modelled F0 contour contradicted the predicted surface tone, and the perceived tone. These mismatches were then subjected to more careful analysis.

### 5.3.3 Analysis of the Findings

An overall comparison between the surface, perceived and tone commands is depicted in Tables 5.5 and 5.6. Table 5.5 illustrates the percentage match between the surface tone and tone commands derived from the Fujisaki model. Table 5.6 presents the percentage match between the Fujisaki tone commands and the perceived tones. These tables analyse each syllable in isolation, and do not take the possible influence of adjacent syllables into account. An analysis which considers the effect of neighbouring syllables follows in the next sub-section.

In this simplistic analysis, the presence of a tone command of any amplitude and duration within the syllable was interpreted as a Fujisaki tone command. Taking this view, approximately 71% of the syllables ascribed tone commands are also predicted as high surface tones in Table 5.5, despite the presence of tone commands. Table 5.6 shows a similar comparison between the perceived tone and the Fujisaki tone

**Table 5.6:** Syllable match between the perceived tone and the Fujisaki tone commands.

| | % with tone commands | % without tone commands |
|---|---|---|
| Perceived tone high | 78.2 | 21.3 |
| Perceived tone low | 29.5 | 68.7 |

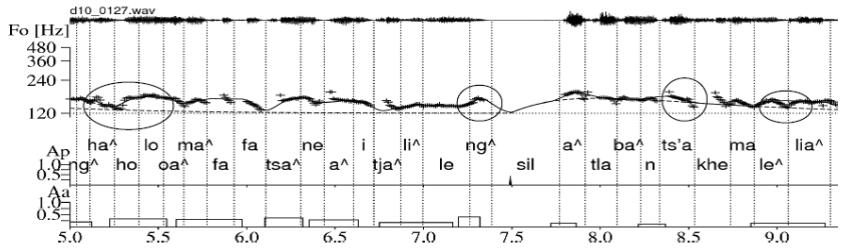



**Figure 5.2:** Sample from sentence d10_0127 illustrating some of the discrepancies between the Fujisaki-modelled F0 contour and the surface tone labels. The predicted high surface tones are indicated by the symbol '^', the modelled F0 contour by a solid line, and the extracted F0 contour by the crosses (+). The part-sentence reads '... *haholo oa mafafatsane a itjalileng, a tla ba nts'a khema le lialuma.*' - "... of scattered showers in particular, with a bit of thunder."

commands. This table shows that 78% of the tones perceived as high were also associated with tone commands by the Fujisaki model and that 69% were deemed low by both the model and the perceptual evaluation. It is worth noting that some of the tone commands that are associated with low tones in Tables 5.5 and 5.6 are a result of tone commands that start and end in low tone syllables in order to reach a high value for a neighbouring high tone. In these tables, 29% (of the 28%) and 19% (of the 29%) respectively fall into this category. Thus a more discerning analysis would give more optimistic percentages.

As a second step, the prosodic groups of high tone syllables, as depicted by the Fujisaki tone commands, were studied. The purpose was to understand how such sequences of high tones related at word and inter-word boundaries. The results of these analyses are given in the following two subsections.

### Tonal influences of neighbouring syllables

The tones of the syllables were studied with reference to the influence of their immediate left and right neighbours. It was suspected that the effect of these neighbours on the syllable tone might be different based on whether a syllable is within a word or at a word or phrase boundary.

The relationship between the predicted surface tone, the perceived tone, and the output of the Fujisaki model was also considered. This was investigated by comparing the tonal patterns of each syllable in the 53 sentences, and attention focused on the discrepancies that occur among the three 'predictions'.

The considered sentences contained some repetitions of (partial) phrases, and this was to check for consistency of both the Fujisaki model parameters and the perceived tone classifications. Where there were discrepancies, explanations were made based on our analysis. During the comparison with the Fujisaki model, it was often possible to identify prominent obvious mismatches by listening to utterances resynthesized with the model.

Figure 5.2 illustrates the three important aspects that were identified by our analysis, namely F0 extraction errors by the Fujisaki model, surface tone/perceived tone mismatches, and the surface tone prediction rule.

In the figure, in the word *haholo* (indicated by the first ellipse), F0 shows a low excursion for *ha*, yet the surface prediction is high. Although *haholo* has an underlying pattern of LLL, the high tone on *ha* has been spread from the syllable *n^g* preceding it by HTS1 rule. The question here is whether this is a surface tone prediction error. The suspicion is that the mistake originates from the lexical tone pattern, since *haholo* is ascribed two different tonal patterns by the two tone-marked dictionaries used as references – LHL and LLL. Based on the F0 contour, the more plausible tonal pattern appears to be LHL. From this pattern, one might conclude that perhaps there is no tone spread in this instance. There is a high tone spread from *oa* to *ma* of *mafafatsane* as per HTS1 rule, and the F0 excursion is high as expected. Application of the RBD rule is observed where the high tone of *tsa* of *mafafatsane* is spread to *ne* (by HTS1 rule) but because the following *a* is underlyingly high, the high tone targeted at *ne* is delinked. The F0 excursion is in agreement with the surface tone prediction.

The second ellipse in Figure 5.2 highlights the *n^g* of *itjalileng*, which is at a phrase-final position and shows a high F0 excursion. The question here is whether this is due to a continuation rise or not. The literature provides two possible answers to this question. The first would be that the tone should rise due to the continuation rise. Taylor [100] states that often F0 is low at a phrase boundary, but in many cases F0 is high. For instance, if another phrase follows the current one, a continuation rise may be observed. The second is that there should be a high F0 excursion because *itjalileng* is a relative verb. According to Zerbian [119], relative verbs always have a high tone on their last syllable, *ng*. Therefore, in this case, the finality restriction (FR) rule is not necessarily violated but an exception to the rule is highlighted.

The tone command on *ts'a* (third ellipse in Figure 5.2), which was deleted, might well be the influence of the *ts* sound. In other words, it might be a result of micro-prosody. The deep dip between *le^* and *lia^* (fourth ellipse in Figure 5.2) is the influence of the d-like sound.

The sub-corpus was dominated by sentences in future tense (51 out of 53 sentences). These sentences were marked by the future tense marker, *tla*, of which there were 22 instances. This marker is underlyingly a low tone syllable. The surface tone transcription, in 20 out of 22 appearances, also predicts a low tone for this syllable. (In the two instances where this was not so, it could be ascribed to a high tone spread from the preceding syllable, and the following syllable is again low-toned.) In these instances, the syllable is surrounded by high-toned neighbours but remains low-toned due to the right-branch delinking (RBD) rule, which delinks the high tone spread from its preceding neighbour, and due to the right-adjacent neighbour being underlyingly high. The expectation is that the F0 contour of the Fujisaki model will show a low excursion at this point, but instead a high excursion is observed. As a matter of fact, *tla* is perceptually high and its contour is higher than its neighbouring syllables. This is not expected by the surface tone rule. Furthermore, this is one case in which the F0 contour shows consistency with the tone commands in all instances, and in each instance the surface tone disagrees with the Fujisaki model. This discrepancy though, was also observed in cases where a low-tone syllable was surrounded by high tone neighbours. In these instances, the F0 contour stays high. The explanation for this kind of discrepancy is that since the F0 contour takes time to rise and fall, it may in such a case simply stay high. The high-toned effect of the preceding syllable is carried over to the next syllable, and this is observable perceptually and by the tone commands in the Fujisaki model, but is not accounted for by the surface tone prediction.

Another interesting case is the word *teng*. This word is underlyingly LL or HL, and in our corpus, the latter pattern was assumed. This word appears seven times in the corpus of 24 sentences and is located at phrase-final position in six of those instances, and mid-sentence once. While the predicted surface tone is also HL, in each case for the phrase-final position the F0 contour shows a LH tonal pattern. The perceived tone is LH in four instances and HL in two. In mid-sentence position, the observed Fujisaki

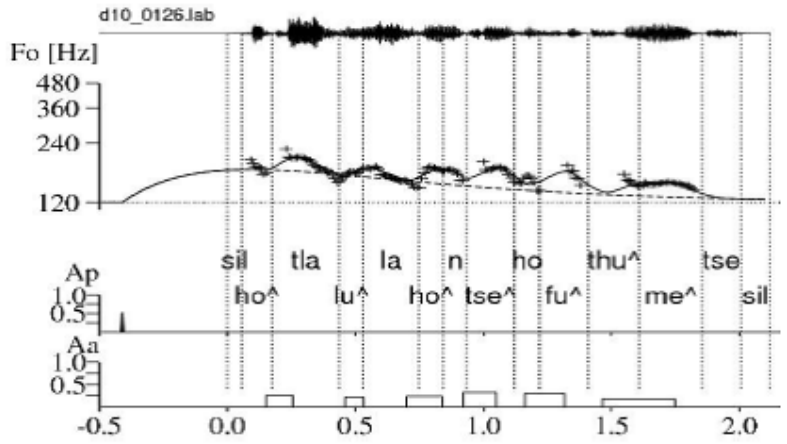



**Figure 5.3:** Sentence d10_0126 showing delays in the maximum of the F0 waveform. The sentence is: '*Ho tla lula ho futhumetse.*' - "It will remain warm."

tonal pattern (F0) is LL, and is also perceived as LL. Since the *ng* of *teng* is rising, the question is whether this is some kind of continuation rise even though the predicted tone is low. If so, it could be concluded that the continuation rise overrides the tone prediction rule.

A further observation is that the tone commands pertaining to a high tone can be delayed in some sentences. A high tone corresponds to a period of relatively high F0, and sentence d10_0126 in Figure 5.3 illustrates such a case. In this figure, the first *ho^* should be high but the following *tla* is higher than *ho^*. The same applies to *lu^la* where the F0 peak is delayed. In [121], it is stated that peak delay is quite common cross-linguistically. In a study for Northern Sotho [122], which belongs to the same family group as Sesotho, it was shown that the F0 peak associated with a high tone is not necessarily reached in the syllable it is associated with. This aspect is also corroborated by van Niekerk and Barnard [110] in their study on tone modelling for Yoruba. The study showed that peaks are generally realised late in the syllable possibly even in the following syllable, especially in the case of LHL sequences. In his study for Chichewa, another Bantu language, Myers [84] found that the timing of the F0 peak was dependent on the position of the syllable in a phrase: medial, penultimate, or final. According to Zerbian and Barnard [122], the peak shift inducing object concord has been confirmed for Sesotho, but no detailed data and accounts were available.

### Prosodic grouping of high tone syllables

In the Fujisaki model, sequences of consecutive high tones were sometimes observed. In these sequences, the words and syllables linked up so closely that they appeared to be one unit, indicated by a single prolonged tone command. Therefore, it was investigated how these prosodic groups are formed, and how the syllables or words within them are related. The findings for the high tone groupings are given below, and Figure 5.4 illustrates some of the issues raised.

- Most of the groupings (67 of 125 instances) were due to two or more adjacent high tone syllables, with the last high tone being carried over (possibly by some form of peak delay) into the immediate right neighbouring low-tone syllable. This was observed both within words and across word boundaries.

- Another frequent grouping consisted of alternating tone labels (37 examples), e.g. HLHLH. When

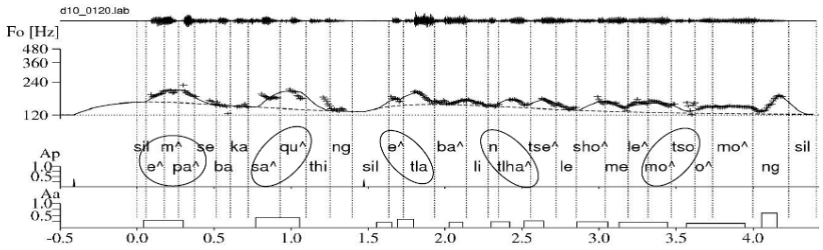



**Figure 5.4:** A sample sentence illustrating prosodic groupings of different types.

the F0 contour is not given enough time to decline and rise for the next high or low tone, it remains high. Also, in this instance, the final high tone label is carried over (third ellipse in Figure 5.4) to the next low tone syllable. This was also observed across word boundaries.

- There were also prosodic groups that begin with a low-tone syllable (31 examples), both across words and within words. From these examples, what stood out was that the low-tone syllable in question was preceded by a predicted high surface tone syllable, and this high tone syllable was not captured in the same tone command grouping (fifth ellipse in Figure 5.4 – *tso* preceded by *mo^*). Of these low-toned syllables at the start of the sequence, nine were *ng* and five were *n* (fourth ellipse in Figure 5.4).

- For infinitive verbs, where a prosodic sequence was formed, the low-toned class prefix, *ho*, was included in the prosodic group. This occurred irrespective of the tone label of the syllable preceding *ho*.

- In a few cases, the prosodic groupings did agree with surface tone predictions, but across word boundaries (*sa quthing*) and within words (*empa*) – (first and second ellipses in Figure 5.4). In these instances, there was no carrying over of the high tone. The high tone sequence for *mocheso* was consistent throughout, with the tone group consisting of *cheso*. This sequence agrees with the surface tone.

Prosodic groupings that do not share the same tone commands, and thus have differing tone command amplitudes, were also analysed. There were 39 such instances, in 34 of which the tone command overlap occurred across word boundaries, and in five it occurred within words. For instances across word boundaries, this was either at the end of a word, at the beginning of a word, or at a single-syllable concord. The prosodic effect of two or more adjacent high-tone syllables, the alternating tone labels, and the carrying-over of the high tone held here too.

As observed in our examples, and as confirmed by [123], the F0 does not drop rapidly within utterances after it has reached a peak. A gradual decline is observed, rather than a steep drop to a subsequent low tone [123]. Myers [83] shows that the pitch value of a low tone is determined by the tonal environment. Therefore, in the instances of alternating tone labels for syllables, the favourable environment seems to be that for a high tone, thus a sequence of high tones depicted by the tone commands. According to [121], the actual pitch value of a low-tone syllable thus depends on the presence and location of preceding high tones. This is confirmed by the examples considered in our analysis above.

### 5.3.4 Discussion

Since neither the surface tone transcription nor the perceived tone labels are completely reliable, care must be exercised in the interpolation of the Fujisaki model commands. F0 can be inaccurately estimated, surface tones are sometimes incorrectly specified, and there can be flaws in the Fujisaki analysis, like the insertion of a phrase command instead of a long tone command of small amplitude. Then there is also the influence of micro-prosody to be considered. In such cases, careful inspection of the Fujisaki analysis in conjunction with listening tests, is currently the only means to achieve consistency.

In order to reconcile the high surface tones with the tone commands from the Fujisaki model, the reasons for disagreement must be determined. In the case of surface tones, there is the possibility of incorrect prediction due to wrong or unknown underlying high tone labels, but also of incorrect or unknown tonal rules. For the perceived tones, the perceptual tests should be repeated by different subjects and the tone labels achieving the greatest consensus noted. This would allow greater confidence to be placed in the annotation of perceived tone.

Overall, the perceived tone more closely matches the results of the Fujisaki analysis than the surface tone. In many cases, however, the discrepancies between these alternatives could be accounted for by a visual inspection of the F0 contour (in the Fujisaki analysis) of each syllable, and by considering the influence of the neighbouring syllables on the tone label of the target syllable. The results obtained from this inspection revealed errors introduced by micro-prosody in the Fujisaki analysis, as well as cases in which the surface tone prediction rule fails. For high tone sequences, the tone commands reveal that groupings are mostly due to two or more adjacent syllables (either within a word or across words) with high surface tones. The other factor contributing to the prosodic groupings is the alternation of tone labels between adjacent syllables. In both cases, a delay in the F0 values is observed. This also means that the high tone can be carried over to a neighbouring low-tone syllable.

## 5.4 Characterisation of Prosodic Groups in Sesotho

### 5.4.1 Aim

In the previous experiment, it was revealed that prosodic groups formed by the Fujisaki tone commands correspond mainly to sequences of two or more adjacent syllables with high surface tone labels, and sometimes to adjacent syllables with alternating surface tone labels. In the current experiment, this work was extended by categorising and characterising these groups in order to understand how they are formed in relation to surface tone transcription. Since the sub-categories of these groups were found to depend on the predicted surface tone labels of their neighbours, the prosodic groups could be identified using this relationship. Repeated phrases were also analysed to check consistency.

### 5.4.2 Speech Data

A subset of 144 sentences from the main corpus, which equated to 45 minutes of speech, was used for this experiment. This corpus had a total of 7884 syllables, 3847 of which were found to have a high surface tone. Of the high surface tone syllables, approximately 89% also had associated Fujisaki tone commands as shown in Table 5.7. For low surface tone syllables, approximately 61% ascribed Fujisaki tone commands. In the table, each syllable has been analysed in isolation. The high percentage match between low-toned syllables from the surface tone prediction and Fujisaki tone commands can be ascribed largely to F0 peak delay associated with high-tone syllables. The delay leads to the presence of a Fujisaki

**Table 5.7:** Syllable match between the surface tone prediction and the Fujisaki tone commands, for a corpus of 144 sentences.

| | % with Fujisaki tone commands | % without Fujisaki tone commands |
|---|---|---|
| Surface tone high | 89.1 | 10.9 |
| Surface tone low | 61.4 | 38.6 |

tone command in the succeeding low surface tone syllable. Peak delay accounts for approximately 65% of the 61.4% mismatched low-tone syllables shown in Table 5.7. This table complements Table 5.5.

### 5.4.3 Categorising Prosodic Groups

The tone commands resulting from the Fujisaki analysis were studied and prosodic groups were identified by locating sequences of high tones. These were then classified into two categories: 1) Groups for which the surface tone transcription consisted of two or more adjacent high-tone syllables, e.g. HHHHHH. 2) Groups for which the surface tone transcription consisted of alternating tone labels, e.g. HLHLHL. In the analysis, the onset and the offset of a tone command or prosodic group can start anywhere – the beginning, middle or end of a word. The first group is referred to as the high-tone sequence group and the second as the alternating-tone sequence group. The sub-categorisation of each of these two groups is described in the following sections.

#### High-tone sequence group

This prosodic group was divided into four sub-categories: 1) sequences of all high tone syllables (e.g. HHHHH), 2) sequences of all high tone syllables plus one final low tone syllable (e.g. HHHHL), 3) all high tone syllables with an initial low tone syllable (e.g. LHHHH), and 4) all high tone syllables with one mid-sequence low tone syllable (e.g. HHLHHH). A mid-sequence low tone syllable was allowed at any position except the beginning or the end of the sequence. For the last three groupings, a sequence must contain a minimum of three syllables, the low tone syllable included. All 2-syllable sequences were strictly HH. The minimum number of syllables in a grouping was two and the maximum was seven, with only one instance of the latter.

#### Alternating-tone sequence group

This group was sub-categorised into three groupings: 1) alternating sequences starting with a low tone syllable (e.g LHLHLH), 2) alternating sequences with initial high tone syllable (e.g. HLHLHL), and 3) alternating sequences (any tone label at initial position) with an additional high tone syllable at the end (e.g. LHLHLHH). The last two sub-categories apply to groupings with five syllables or more, all syllables included. The sequences in the alternating-tone grouping contained a minimum of two syllables and a maximum of seven.

### 5.4.4 Analysis of Prosodic Group Categories

The two major prosodic groups were further analysed and some observations were made as indicated in Table 5.8. The table shows the number of occurrences of the prosodic groupings in our corpus. The table shows that, for both major groups, most sequences are short, and, as the number of syllables within a sequence increases, the number of occurrences of such sequences decreases. The maximum number of

**Table 5.8:** Number of occurrences of prosodic groups of various lengths.

| Group type | 2 syllables | 3 syllables | > 3 syllables | **Total** |
|---|---|---|---|---|
| High-tone | 296 | 258 | 113 | **667** |
| Alternating-tone | 296 | 169 | 101 | **566** |

syllables in a prosodic group for both group types observed in our data was seven, with one occurrence of a seven-syllable sequence in each group.

A general observation made for all sequences that include a low-tone syllable among surrounding high-tone syllables is that the Fujisaki model tends to indicate a high tone, in a form of tone command, even though the surface tone of the syllable is low. This can be interpreted as the carrying over of the previous high tone to the immediate right neighboring tone, by some form of peak delay. Our results support the concept of peak delay mentioned by [123], and that a gradual decline, rather than a steep drop of the F0, is always observed on a subsequent low tone.

### High-tone sequence group

The syllable sequences forming part of this group were associated with the following surface tone forms:

HHHHH: This is a perfect sequence of high tone syllables, and in this case the Fujisaki model agrees with the surface tone prediction.

HHHHL: In this case, the high-tone effect of the final high-tone syllable is carried over to the concluding low-tone syllable. This carrying over of a high tone is due to peak delay, where the higher pitch of the predecessor syllable associated with the high tone takes time to subside, thus resulting in the detection of a tone command by the Fujisaki analysis for a low-tone syllable.

LHHHH: In this case, a string of high-tone syllables is preceded by a low-tone syllable. The high tone realised on this low-tone syllable is due to anticipation, since the subsequent high tone can only be reached with a limited speed. Out of 156 LHHH groupings (Table 5.9), the majority was preceded by a high-tone syllable, which was not considered to be part of the same grouping by our analysis.

HHLHH: In this case, a low-tone syllable is surrounded by high-tone neighbours. Because the F0 contour takes time to rise and fall, it may in such a case stay high enough to lead to the detection of a tone command by the Fujisaki model. This peak delay results in all syllables being interpreted as high toned by the Fujisaki model.

LHHL: There were very few instances of this type of grouping with both leading and trailing low-tone syllables and, in all cases, the low tone syllables were apart by two high tone syllables, e.g. HLHHL. These had a minimum of four syllables and a maximum of seven.

Table 5.9 illustrates cases of all high-tone sequence groups described above. Examples of each case are illustrated in Figure 5.5. From the table it is seen that, for groupings with one low tone syllable, the low tone syllable occurs almost twice as often at the initial position as at the final position, and the occurrence of the low tone syllable at the final position is more than twice as frequent as its occurrence at a middle position.

In Figure 5.5, the three panels represent three different partial utterances and these show examples of the high-tone prosodic groups given in Table 5.9. The figure also illustrates that an utterance can contain more than one prosodic grouping.

**Table 5.9:** Types of high-tone sequence groups found in our data. Corresponding speech examples are illustrated in Figure 5.5.

| | Grouping type | Indicated in Figure 5.5 by ... | Number of cases | Percentate [%] |
|---|---|---|---|---|
| All H | e.g. HHHHHH | A | 385 | 57.7 |
| One L | L-initial (e.g. LHHHHH) | B | 156 | 23.4 |
| | L-mid (e.g. HHLHHH) | C | 36 | 5.5 |
| | L-final (e.g. HHHHHL) | D | 83 | 12.4 |
| Two L's | LHHL | E | 2 | 1.0 |
| | LHHLHH | F | 2 | |
| | LHHLHHH | F | 1 | |
| | HLHHL | G | 2 | |
| **Total** | | | **667** | 100 |

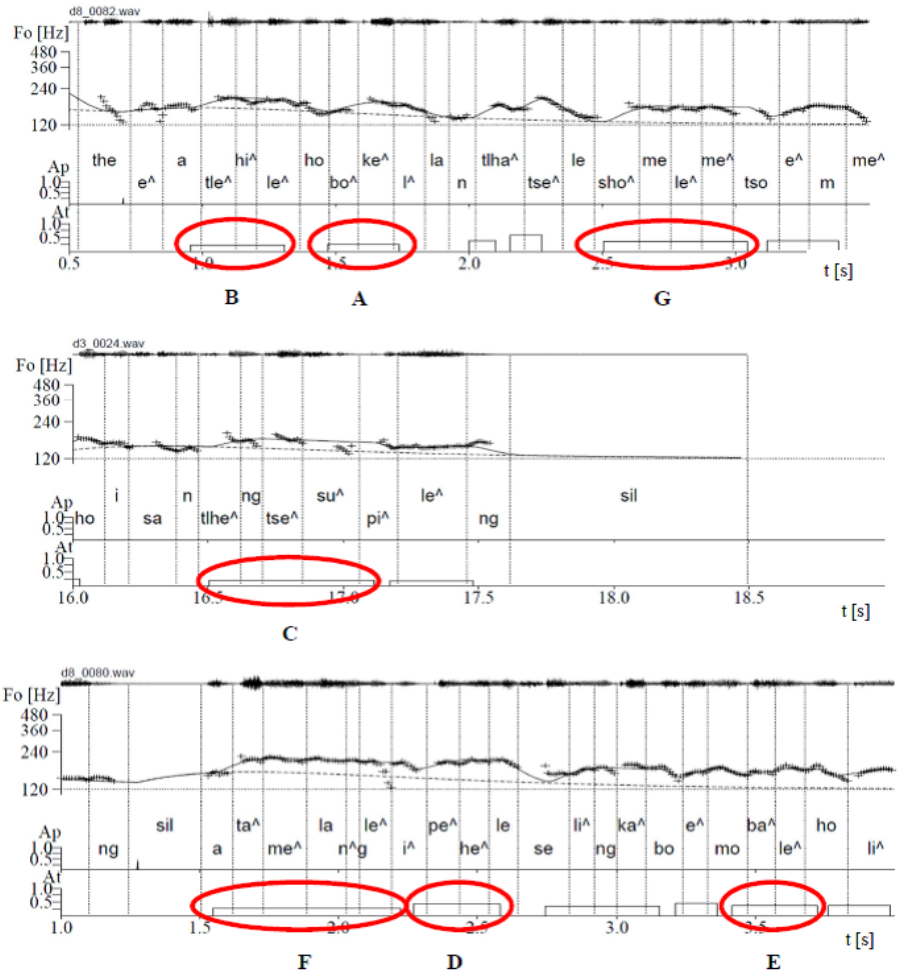



**Figure 5.5:** Examples of high-tone sequence groups as shown in Table 5.9.

### Alternating-tone sequence group

The members of this group consisted of a sequence of syllables with alternating high and low surface tones. There were 308 such instances beginning with a low tone syllable, and 158 starting with a high tone syllable. There were 16 instances ending with two consecutive high tone syllables, e.g. LHLHLHH. In these cases, the F0 contour behaved as expected by staying high. There were no cases of more than two adjacent low tone syllables in the final position.

For alternating-tone sequence groups in general, a peak delay is observed, leading to a sustained high-tone realisation by the Fujisaki model. This is confirmed by Myers [83] who indicates that the pitch value of a low tone is determined by the tonal environment. Therefore, in cases of alternating tone labels for syllables, the favourable environment for Sesotho seems to be that of a sustained high tone. The F0 contour of the alternating sequence also supports this view, as it stays high within the prosodic group environment.

### Repeated Phrases

Since the corpus is derived from weather forecast bulletins, certain sentences and/or phrases are repeated. This allows the consistency of the Fujisaki analysis and the prosodic groups to be checked. From the two major groups, thirty-five such repeated phrases were selected randomly and compared. It should be noted that in all these cases, the amplitude of the tone commands forming these prosodic groups might be different to those they are being compared with, although this difference was observed to be small. The reason for the difference in amplitude could almost always be attributed to a phrase contour with lower than usual magnitude, leading the tone commands to compensate for the subsequent loss in F0 amplitude. The relationship between the compared repeated phrases was classified as matching, expansion or split.

**Matching:** By matching prosodic groups it is meant that the prosodic groups captured the same syllables, and the same number, and were of approximately equal amplitude. Figure 5.6 shows instances of similar prosodic groups for the phrase *'Matsatsing a mabeli a latelang'* - "The next two days" for utterances d2_0018 and d1_007. The prosodic group of interest is *'tsatsing a'* (indicated by a rounded rectangle). The phrase occurred in different utterances but the alternating-tone sequence group, *tsatsing a* (LHLH), is identified for the same syllables in both cases.

**Expansion:** In the case where there was no exact match between prosodic groups, they would often differ by a single syllable. The additional syllable occurred either to the left (expansion to the left) or to the right (expansion to the right) of the prosodic group. Figure 5.7 illustrates expansion to the left while Figure 5.8 shows expansion to the right.

In Figure 5.7, the marked prosodic group (indicated by a rounded rectangle) of the phrase *'bosiung ba kajeno'* - "tonight" covers the syllables – *'ba kaje'* - in utterance d11_0148. In utterance d4_0030, there is expansion to the left of the same prosodic group where *ng* is also captured. This results in *'ba kaje'* (HHH) expanded to *'ng ba kaje'* (LHHH). The F0 contours of the phrase in both utterances look very similar, but the phrase commands have different magnitudes.

Figure 5.8 shows an expansion to the right of the prosodic group *bosiu* (LHH) in utterance d4_0030 to *bosiung* (LHHL) in utterance d11_0148 of the same phrase as in Figure 5.7, where *ng* is captured in the second utterance. The prosodic group has expanded to the right adding a low tone syllable.

Overall, expansion to the left occurred more often than that to the right. For both expansions to the left and to the right, the F0 contour approximation made by the Fujisaki model is very similar to the contour obtained when no expansion occurs.

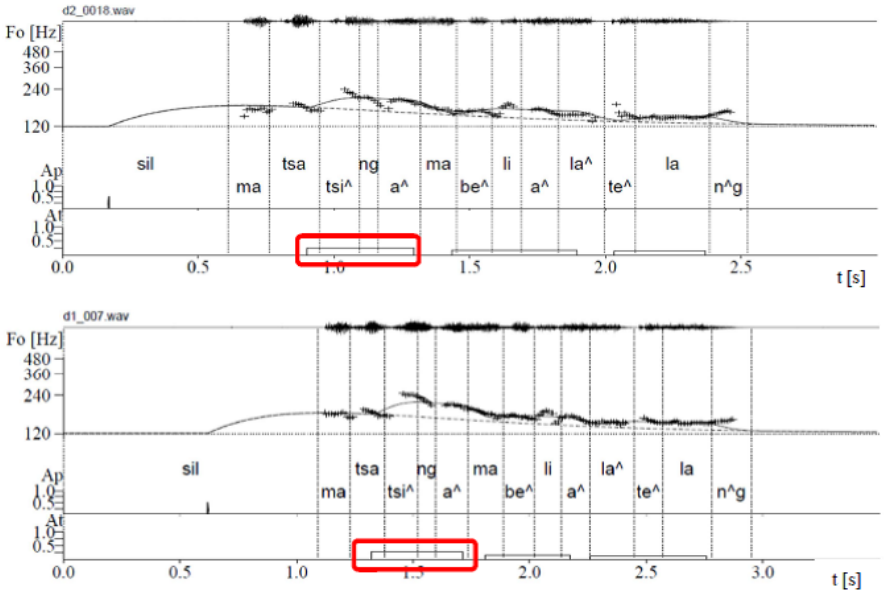



**Figure 5.6:** Matching prosodic groups from different utterances of a partial phrase *'tsatsing a'*. The main phrase is *'Matsatsing a mabeli a latelang.'* - "The next two days."

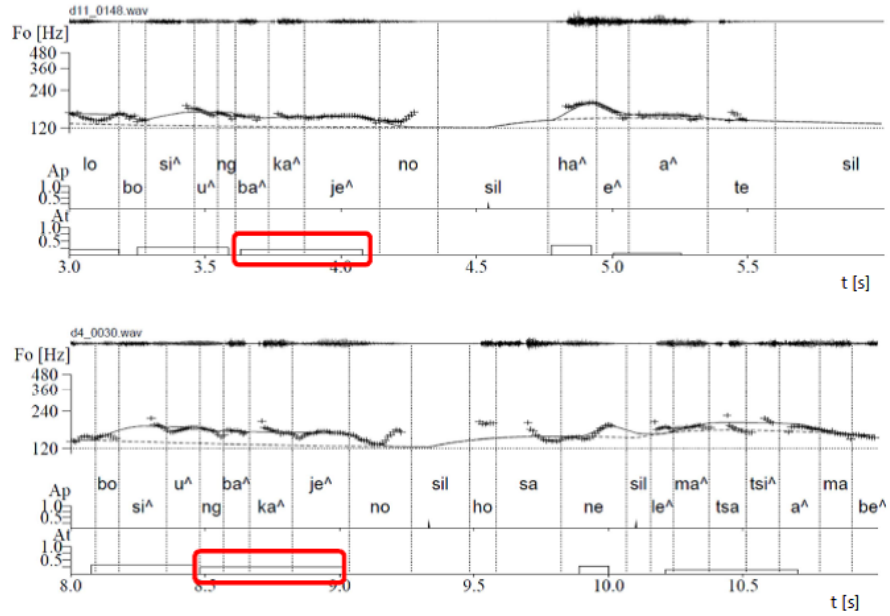



**Figure 5.7:** Expansion to the left of the prosodic group *'ba kaje'* to *'ng ba kaje'*.

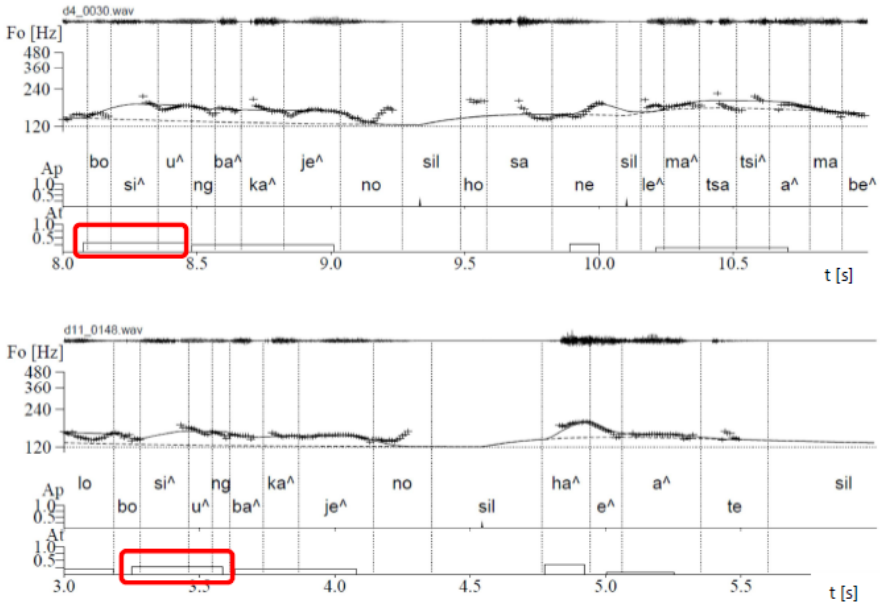



**Figure 5.8:** Expansion to the right of the prosodic group *'bosiu'* to *'bosiung'*.

The expansions observed in these examples do not necessarily indicate a degradation in the F0 approximation presented by the Fujisaki model, and consequently the naturalness of the speech output. They indicate that similar F0 contours can be achieved by means of differing tone command timing. The effect of expansion by a full syllable will be tested perceptually and verified in future work.

**Split:** Occasionally, for the same phrase, a prosodic group was seen to split into two. In this case, the same syllables would still be covered although now by two tone commands, as shown in Figure 5.9. The split would occur within a syllable segment. The splitting of prosodic groups was less common than expansion.

In Figure 5.9, the prosodic group *'mmohi oa he'* is split into two smaller groups, *mmohi* and *oa*. In this case, the cause of the split is a high phrase component in utterance d5_0048, which is expected at the beginning of an utterance. Ideally, the phrase component must coincide with the F0 values at the segmental onset of a new phrase. In utterance d2_0019, the magnitude of the phrase component is not as high, and thus allows a single and longer tone command of higher amplitude. It should be noted that this split can also be interpreted as one tone command being represented by two tone commands. Both lead to very similar F0 contours and, since the F0 contour represents a natural utterance, it is expected that they would lead to the same perceptual effect.

The comparison of repeated phrases is summarised in Table 5.10. Note that a phrase can have more than one prosodic group associated with it, as already seen for utterances d4_0030 and d11_0148. The table shows that the number of matching prosodic groupings far outweighs the other options. This confirms that similar phrases, in terms of constituent syllables, are quite similar in terms of the F0 contours and associated tone commands and prosodic groups.

### 5.4.5 Discussion

Prosodic groups that were seen to occur in Sesotho speech after Fujisaki analysis were analysed. These groups were characterised by a single, prolonged high tone command that spans at least two syllables.

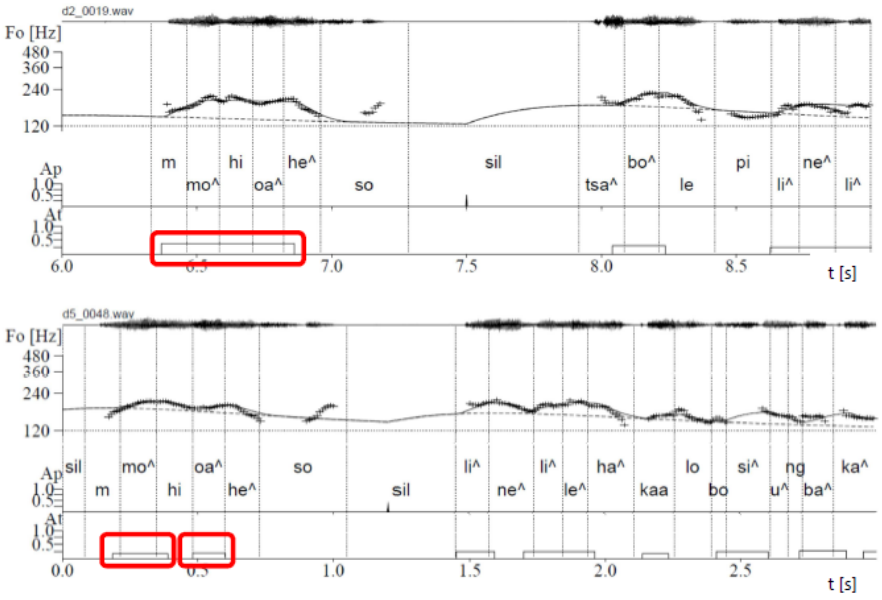



**Figure 5.9:** The prosodic group *'mmohi oa he'* is split into *'mmohi'* and *'oa he'*.

**Table 5.10:** Comparison of prosodic groupings in 35 randomly selected repeated phrases.

| Feature observed | Number of prosodic groups |
|---|---|
| Matching | 119 |
| Expansion to the left | 7 |
| Expansion to the right | 3 |
| Split | 3 |
| **Total** | **132** |

The analysis shows that these prosodic groups can be associated with certain sequential surface tone patterns. These usually take the form of a corresponding sequence of high surface tones, but may include leading, trailing or intermediate low surface tone syllables. Leading and trailing low tones are in the overwhelming majority, and support the anticipatory rising and delayed lowering (peak delay) of pitch observed for other tonal languages by other researchers. Prosodic groups may also be associated with a sequence of alternating surface tones. Also, for this case the Fujisaki analysis suggests a single prolonged tone command to be the most appropriate for F0 modelling. By considering repeated phrases in our data, it was shown that the Fujisaki analyses were highly consistent.

The patterns of behaviour seen suggest ways in which appropriate Fujisaki tone commands can be inferred from a surface tone transcription. The implementation and perceptual testing of such methods is the focus of the next section.

## 5.5 Perceptual Evaluation of Prosodic Groups

### 5.5.1 Aim

Prosodic groups of high-tone sequences and alternating tone sequences were identified and characterised in the previous experiment. In the current experiment, duration and amplitude of these groups were

**Table 5.11:** Frequency of occurrence of high-tone prosodic groups of various lengths.

| Number of syllables | Prosodic groups |
|---|---|
| 1 | 111 |
| 2 | 296 |
| 3 | 258 |
| 4 | 15 |
| **Total** | **680** |

modified by means of the Fujisaki model and perceptual evaluation performed. Tone commands of individual syllables were also considered. The purpose was to investigate how the effect of modifying the tone commands affects the naturalness of speech output. This was a way of converging surface tonal transcription and the Fujisaki model in order to achieve correct realisation of tone in an eventual Sesotho TTS system.

### 5.5.2 Data Selection and Modification

The data used for this experiment is a corpus comprising a total of 256 sentences which make 51 minutes of speech.

All prosodic groups identified were categorised into two groups: those whose surface tone labels matched with Fujisaki tone commands, and those whose tone labelling differed from the Fujisaki F0 contour. The two groups were modified in different ways. The duration and amplitude of the first group were modified in a way that would allow easy synthesis from the orthography. For the second group, instances in which the surface tone differs from the tone indicated by Fujisaki analysis were investigated, and the effect of these discrepancies on speech quality determined. The amplitude of Fujisaki tone commands was manipulated to match the surface tones. The resulting resynthesized speech from all modifications was subsequently analysed by perceptual tests, in a bid to determine the best timing and amplitude of tone commands from the Sesotho orthography. Sections 5.5.3 to 5.5.6 give detailed description of the selection and modification methods.

### 5.5.3 Selection of High-Tone Prosodic Groups

Selection of high-tone prosodic groups was based on cases for which tone commands coincided with high surface tone labels only. In other words, high tone cases in which the surface tone and Fujisaki tone commands agreed were considered.

A set of high surface tone prosodic groups was selected from the 256 utterances. An example is shown in Figure 5.10, where ellipses 1, 2, and 5 indicate groups that were included in our selection since these tone commands capture syllables with associated high surface tones. On the other hand, ellipses 3 and 4 indicate groups that were not included since they were associated with at least one low surface tone syllable. For example, the tone command in ellipse 3 has the syllable *Mo* as a low surface tone, while the other syllables have high surface tones. Table 5.11 shows the number of high-tone prosodic groups found in our data spanning 1, 2, 3, and 4 syllables respectively. The table shows that, in the 256 sentences considered, a total of 680 prosodic groups were identified. Once the high-tone prosodic groups had been selected, they were modified as described the Section 5.5.5.

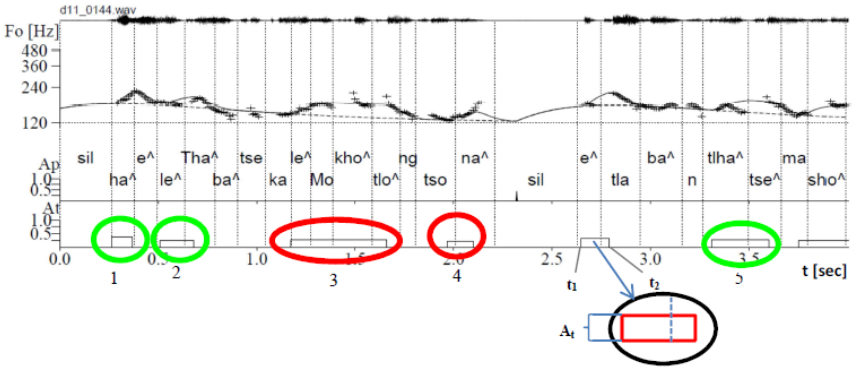



**Figure 5.10:** Selection of high-tone prosodic groups. Ellipses 1, 2 and 5 indicate groups that were selected for analysis. The symbol '^' indicates a syllable with a high surface tone. At is the amplitude of the tone command, t1 is the onset and t2 is the offset.

**Table 5.12:** Instances in which Fujisaki tone commands correspond to low surface tones (FHSL).

| Surface tone pattern | Number of syllables | Number of cases |
|---|---|---|
| Alternating, e.g. LHLHLH | $\geqslant 2$ | 40 |
| Low surface tone labels only, e.g. LLL | $\geqslant 1$ | 7 |
| Any other combination, e.g. LHHL | $\geqslant 2$ | 16 |
| **Total** | | **63** |

**Table 5.13:** Instances in which high surface tones do not correspond to Fujisaki tone commands (FLSH).

| Number of syllables | Number of cases |
|---|---|
| 1 | 26 |
| 2 | 22 |
| 3 | 13 |
| 4 | 3 |
| **TOTAL** | **64** |

### 5.5.4 Selection of Mismatched Syllables

Cases in which the Fujisaki tone command corresponded to one or more low surface tones (FHSL) were selected. Table 5.12 classifies the pattern of surface tones associated with such prosodic groups. Each low surface tone indicates a discrepancy. With this in mind, cases in which a Fujisaki tone command coincided with one or more low surface tones were selected. In total, 63 such cases were isolated from our data.

Cases in which a high surface tone was not accompanied by a Fujisaki tone command (FLSH) were also considered. Examples of 1, 2, 3 and 4-syllable sequences with high surface tones but no corresponding Fujisaki tone command were identified in the data, and are listed in Table 5.13. In total, 64 such cases were isolated from our data.

In Figure 5.11, the rounded rectangles indicate groups where the tone commands capture syllable sequences with at least one low surface tone label. Rectangle 1 is an example of an alternating sequence (LHLH), while rectangle 2 shows a different combination of high and low surface tones (line 3 of Table 5.12). Rectangle 3 is an example of a 2-syllable low surface tone sequence with associated Fujisaki tone command (line 2 of Table 5.12).

Once the selection of the syllables and groups was complete, the matching phrases were ready for

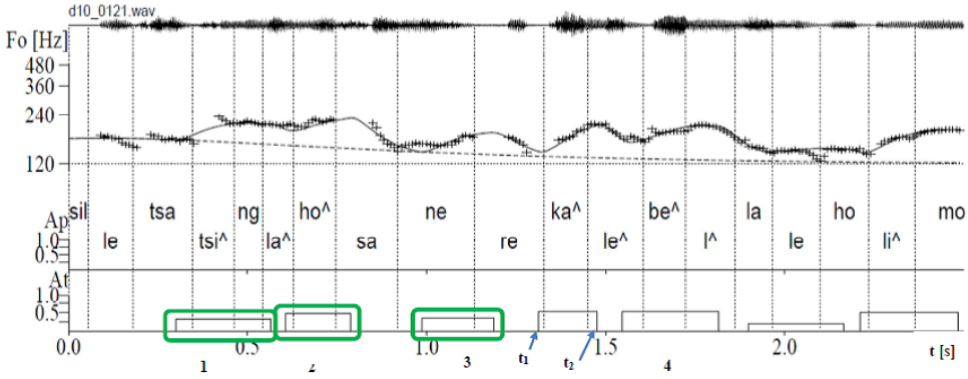



**Figure 5.11:** Selection of groups with mismatched syllables. Rectangles 1, 2, and 3 show examples of groups which were selected for our analysis.

modification.

### 5.5.5 Pitch Contour Modification of High-Tone Prosodic Groups

The tone commands produced by the Fujisaki model are in general not aligned with syllable boundaries, and each tone command has an individual amplitude. If the Fujisaki model is to be used for synthesis (as opposed to analysis), the onset T1, offset T2 and amplitudes At of each tone command must be deduced from the orthography. The perceptual effect of modifying the onset and offset of tone commands determined by Fujisaki analysis to coincide either with a syllable boundary, or a point midway between two boundaries was investigated. The amplitudes were also modified by fixing them to certain discrete levels that are uniform across the utterance. The objective was to determine whether a simple overall rule can be found with which to determine the parameters of the Fujisaki tone commands while leading to the least degradation in the perceptual quality of the speech. The subsections which follow describe the duration and amplitude modifications that were made to the selected high-tone prosodic groups.

#### 5.5.5.1 Duration modification

For those tone commands which fall within the boundaries of a single syllable, which will be referred to as single-syllable prosodic groups, and for which neither T1 nor T2 already coincide with a syllable boundary, the three duration modifications, EL, EM, and ML as described in Table 5.14, were applied. For cases in which the tone command straddled at least one syllable boundary (multiple-syllable prosodic groups), a minimum of three and a maximum of nine modifications were applied, depending on the number of syllables within the group and the length of the tone command across syllable segments. Instances where either T1 or T2 was already aligned with a syllable boundary before modification were excluded from the experiments.

#### 5.5.5.2 Amplitude modification

Amplitude modification was performed by setting all tone command amplitudes within the speech segment under analysis to the same fixed value. Six fixed values were chosen for experimentation: the minimum of the tone command amplitudes in the corpus, the mean, the mean $\pm$ one standard deviation, the mean $+$ two standard deviations, and the maximum. An amplitude of the mean $-$ two standard deviations was

**Table 5.14:** Description of duration modifications applied to high-tone prosodic groups.

| Duration modification | Description | Number of modifications done | | | | Total |
|---|---|---|---|---|---|---|
| | | 1-syl | 2-syl | 3-syl | 4-syl | |
| EE | T1 moved to preceding syllable boundary AND T2 moved to preceding syllable boundary | - | 13 | 4 | 2 | 19 |
| EL | T1 moved to preceding syllable boundary AND T2 moved to following syllable boundary | 34 | 13 | 4 | 1 | 52 |
| EM | T1 moved to preceding syllable boundary AND T2 moved to mid-syllable segment | 35 | 10 | 6 | 1 | 52 |
| LE | T1 moved to following syllable boundary AND T2 moved to preceding syllable boundary | - | - | 13 | 5 | 18 |
| LL | T1 moved to following syllable boundary AND T2 moved to following syllable boundary | - | 11 | 7 | 1 | 19 |
| LM | T1 moved to following syllable boundary AND T2 moved to mid-syllable segment | - | 7 | 10 | 2 | 19 |
| ME | T1 moved to mid-syllable segment AND T2 moved to preceding syllable boundary | - | 13 | 7 | 0 | 20 |
| ML | T1 moved to mid-syllable segment AND T2 moved to following syllable boundary | 35 | 12 | 6 | 1 | 54 |
| MM | T1 moved to mid-syllable segment AND T2 moved to mid-syllable segment of second or last syllable | - | 6 | 11 | 2 | 19 |

**Table 5.15:** Amplitude values used for modification of high-tone prosodic groups.

| Amplitude modification | Tone command amplitude |
|---|---|
| Minimum | 0.0957 |
| Mean - s.d. $(\mu - \sigma)$ | 0.1517 |
| Mean $\mu$ | 0.2657 |
| Mean + s.d. $(\mu + \sigma)$ | 0.3797 |
| Mean + 2s.d. $(\mu + 2\sigma)$ | 0.0494 |
| Maximum | 0.5676 |

not imposed because its value was lower than the minimum value. The values used for modification are shown in Table 5.15.

## 5.5.6 Pitch Contour Modification of Mismatched Syllables

The two types of mismatches identified in Section 5.5.4 were "resolved" by either removing or inserting the Fujisaki tone command in order to match the predicted surface tones. This was accomplished by setting the tone command amplitude to a zero for low surface tone syllables, and inserting a tone command for high surface tone syllables.

### 5.5.6.1 Case FHSL: High Fujisaki tone associated with low surface tone

For cases in which the Fujisaki tone command coincided with a low surface tone syllable, the amplitude of the tone command was set to zero. Figure 5.12 illustrates this modification.

When the Fujisaki tone command corresponded to a sequence of syllables with both high and low surface tones, the amplitude of the tone command was set to zero only for low surface tone syllables. The onset T1 and/or offset T2 times were unchanged for high surface tone syllables. In all other cases, T1

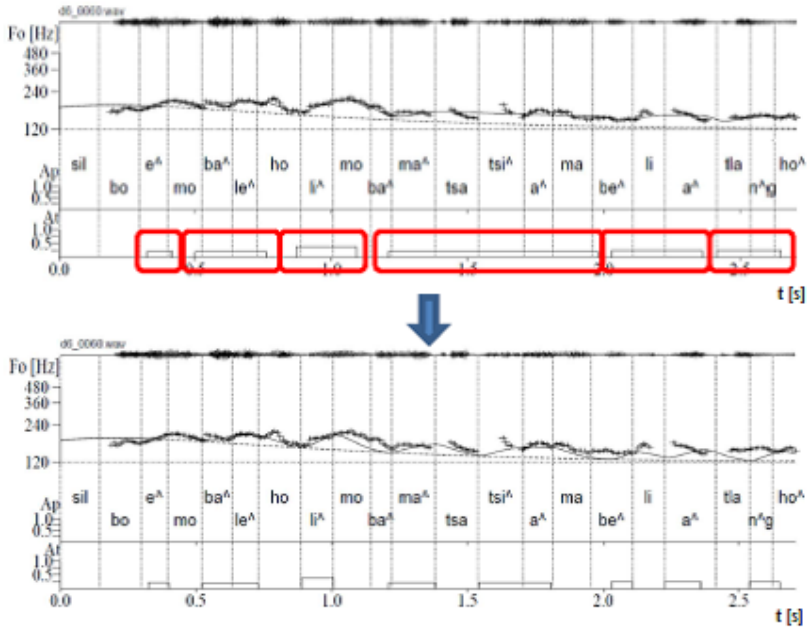



**Figure 5.12:** Modifying Fujisaki tone commands to coincide with low surface tone syllables. The top panel shows an original utterance with surface tone patterns in the following order: HL, LHHL, LHL, HHLHHLH, HLH, and LHH. The bottom panel illustrates a change in pattern after modification. The phrase reads *'Boemo ba leholimo ba matsatsi a mabeli a tlang ho ...'* – "The weather for the next two days ..."

and T2 coincided with syllable boundaries. The optimal alignment of tone command onsets and offsets appears to be language-dependent [80, 78] and its determination for Sesotho is the focus of this study.

#### 5.5.6.2 Case FLSH: High surface tone associated with low Fujisaki tone

In the case of high surface tone syllables with no Fujisaki tone command, tone commands were inserted, with the amplitude set to the average amplitude value over all tone commands. Onset and offset times were set at syllable boundaries. Sixty-four modified phrases were generated, in line with Table 5.13. Figure 5.13 illustrates an example of the modification, where the top panel shows the original utterance, and the bottom one displays the prosodic group generated (indicated by a rounded rectangle) to coincide with high surface tone markings.

### 5.5.7 Perceptual Evaluation

Phrases with modified prosodic groups were subjected to perceptual evaluation by Sesotho speakers. The classification of the data and the evaluation process are detailed below.

#### 5.5.7.1 Data

All the modified stimuli were resynthesised using the FujiParaEditor [77] and the resulting phrases were collected for perceptual testing. Approximately 10% of the utterances used in the perceptual experiments were unmodified to serve as a baseline, while some utterances were repeated (for high-tone prosodic groups) to allow the consistency of feedback from the subjects to be verified. Tables 5.16 and 5.17 give a summary of the data for different modifications and groups. Table 5.16 presents data for duration and

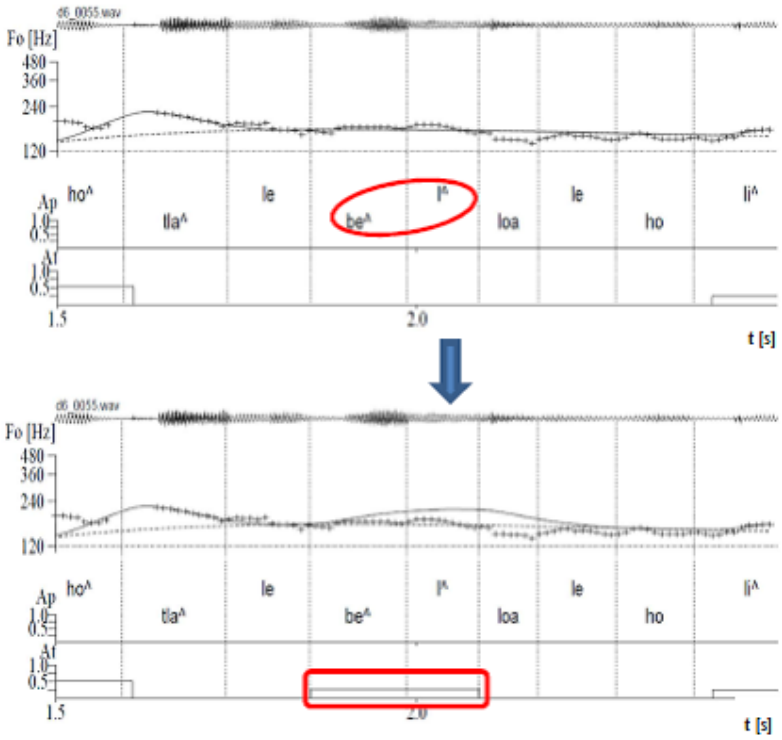



**Figure 5.13:** Inserting a tone command. The top panel shows a sequence of two high surface tone syllables, be^ and l^, with no Fujisaki tone command. In the bottom panel, a tone command (rounded rectangle) is created for these syllables. The partial phrase reads *'... ho tla lebelloa ...'* - "... the expectation will be ..."

**Table 5.16:** Data used for perceptual evaluation of high-tone prosodic groups.

| Modification | Modified phrases | Unmodified phrases | Repeated phrases | Total |
|---|---|---|---|---|
| Duration | 280 | 24 | 41 | 345 |
| Amplitude | 187 | 18 | 36 | 241 |

**Table 5.17:** Data used for perceptual evaluation of mismatched syllables.

| Modification | Modified phrases | Unmodified phrases | Total |
|---|---|---|---|
| FHSL: Fujisaki tone command removed | 63 | 6 | 69 |
| FLSH: Fujisaki tone command inserted | 64 | 6 | 70 |

amplitude modification of high-tone prosodic groups, while Table 5.17 shows modifications performed for mismatched syllables.

The data for high-tone prosodic groups was divided into eight sets, each of which constituted a perceptual evaluation. On average, there were 69 utterances per set and it took each subject approximately 15 minutes to complete an evaluation.

#### 5.5.7.2 The evaluation process

DMDX [28] was used to perform all perceptual evaluations. DMDX is a software tool designed primarily for language-processing experiments. Subjects listened to the utterances in a quiet room using a headset.

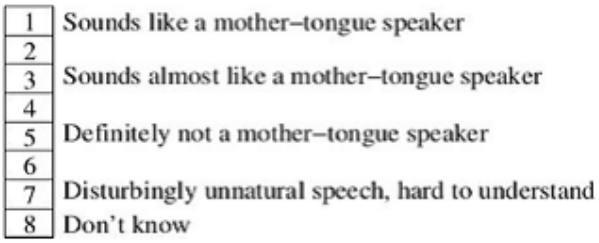



**Figure 5.14:** The rating scale used for perceptual evaluation.

It was ensured that each perceptual evaluation contained a combination of modified (80%), original (10%), and repeated (10%) phrases (for high-tone prosodic groups). Evaluation of the data was based on a rating scale intended to reflect naturalness, as shown in Figure 5.14. The ratings range from 1 to 8, where 1 represents resynthesised speech which sounds like a Sesotho mother-tongue speaker, and 7 represents an unnatural speech. The last rating, 8, is for indeterminant decisions. Subjects were asked to rate each individual audio file according to this scale.

Since the experiments were performed on different days, they were not evaluated by the same test subjects. For high-tone prosodic groups, a total of forty-five people participated in the perceptual evaluation (29 males and 16 females). All subjects are native Sesotho speakers and are from Lesotho, with the exception of two speakers from South Africa. Subjects were students at the University of Stellenbosch (22) and at the University of Cape Town (23). Perceptual evaluation for mismatched syllables was done by twenty-one Sesotho speakers (7 females, 14 males). All subjects are native Sesotho speakers from Lesotho, and were students at the University of Stellenbosch.

### 5.5.8 Results and Analysis of High-Tone Prosodic Groups

In the analysis of all our results, utterances rated with an 8 ("Don't know") were excluded from the analysis. Average scores according to the scale in Figure 5.14 were considered.

#### 5.5.8.1 Duration modification of single-syllable prosodic groups

Figure 5.15 illustrates the average perceptual score given for each type of duration modification for single-syllable prosodic groups, as well as the overall average ("All") and the average rating given to unmodified utterances. The figure also shows a 95% confidence interval (C.I.). The unmodified utterances were perceived as more natural and given an average rating closest to 1. Of the three modification cases, the EL leads to the most natural output, although it is very closely followed by the other two alternatives.

#### 5.5.8.2 Duration modification of multiple-syllable high-tone prosodic groups

Figure 5.16 shows the results of the perceptual evaluation of duration modification applied to multiple-syllable prosodic groups, each of which can be modified in up to nine different ways (Table 5.14). As was the case for single-syllable prosodic groups in Figure 5.15, the EL modification leads to the most natural output. Unmodified utterances continue to achieve the best scores by a clear margin. The EE combination on the other hand gave the least natural speech output in comparison to other alternatives.

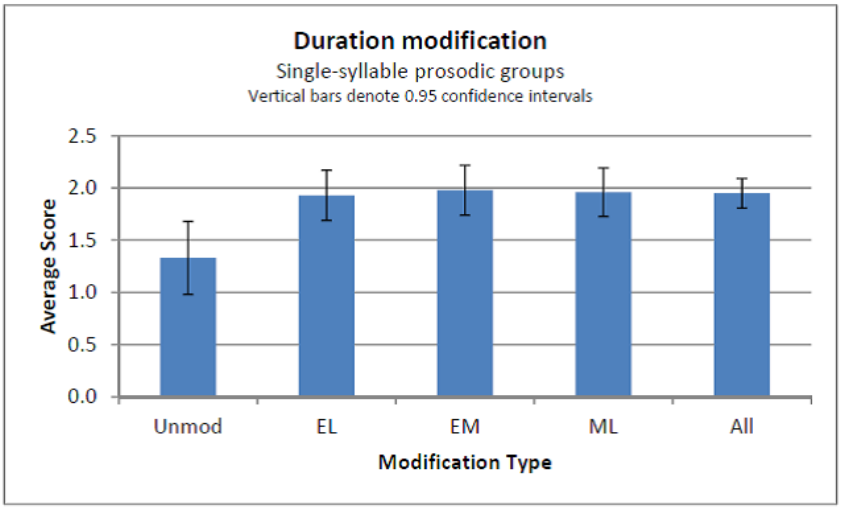



**Figure 5.15:** Comparison of average perceptual scores ascribed to each modification type for single-syllable prosodic groups, also showing the 95% C.I. The average score for unmodified utterances is also shown and that for "All" excludes unmodified utterances.

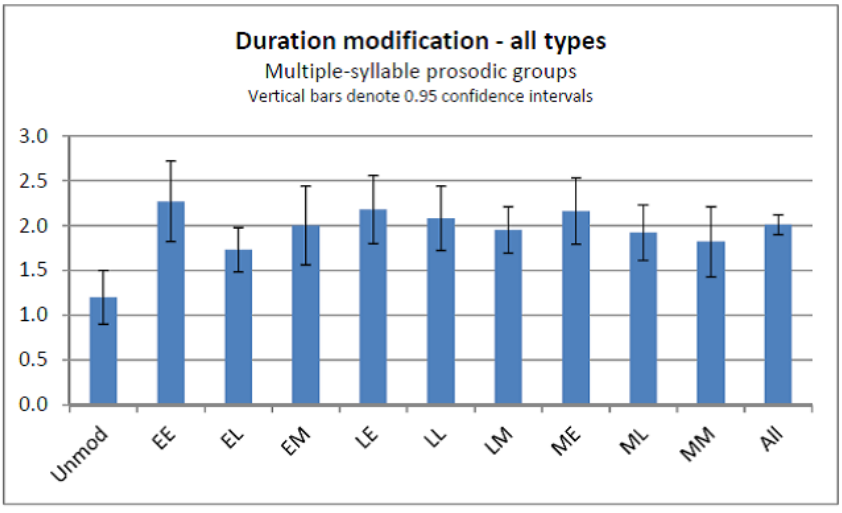



**Figure 5.16:** Comparison of average perceptual scores ascribed to each modification type for multiple-syllable prosodic groups. The average score for unmodified utterances is also shown.

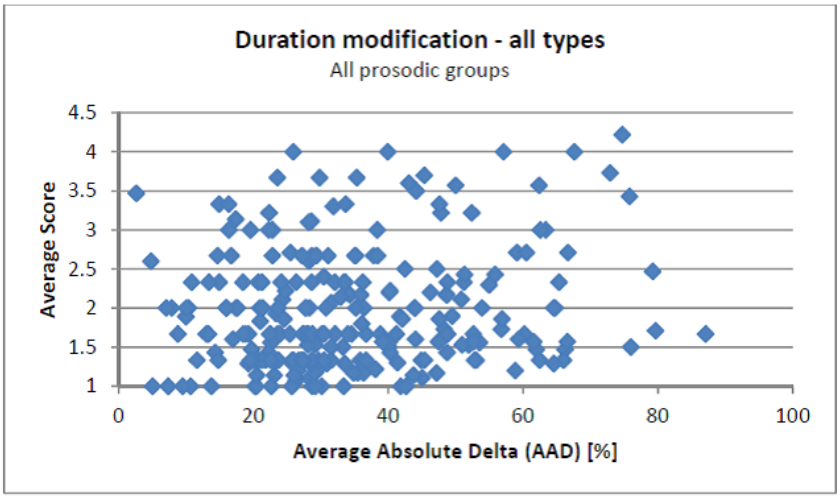



**Figure 5.17:** Average perceptual score for all duration modification types applied to both single-syllable and multiple-syllable prosodic groups as a function of the average absolute change (AAD) as defined by Equation 5.1. The average score for all utterances depicted in this figure is 1.96.

#### 5.5.8.3 Further analysis of duration modification results

Depending on how closely the onset and offset of the Fujisaki tone commands were aligned with the syllable boundaries, the duration modifications were more severe for some cases than for others. In an attempt to determine whether larger changes led to a greater perceptual deterioration, the perceptual scores were determined as a function of the degree of modification made to the timing of the tone command. Figure 5.17 illustrates the average perceptual rating given to each modified utterance (both single and multiple-syllable groups) as a function of the average absolute percentage change in the onset and offset times, which is defined by:

$$AAD = \frac{1}{2}abs\left[100 * \frac{T1 - T1(mod)}{syllableLength}\right] + \frac{1}{2}abs\left[100 * \frac{T2 - T2(mod)}{syllableLength}\right] \quad (5.1)$$

where T1 and T2 are the original onset and offset values, and T1(mod) and T2(mod) are the corresponding values after modification. This measure approaches a value of 100 when the adjustments in T1 and T2 both approach the entire length of the respective syllables. On the other extreme, a value close to zero indicates that both adjustments were very small in relation to the syllable lengths. The figure also shows that the total average score obtained ranges from 1 to about 4.5 (5 meaning "Definitely not a mother-tongue speaker").

Figure 5.18 shows a similar plot, but this time restricts itself to the EL modification, which in Figures 5.15 and 5.16 led to the best quality resynthesised speech. In the figure, most data points have an average absolute delta value of below 40%, and an average score below 3. Figures 5.17 and 5.18 indicate that the score attributed to a modified utterance is not clearly related to the degree to which the onset and offset times have been modified.

In Figure 5.19, the average perceptual scores for utterances with modifications affecting one, two, three, and four syllables respectively are shown. The figure shows that, overall, the perceptual effect of the modifications becomes less apparent as the number of syllables increases.

Figure 5.20 is similar to Figure 5.19 but restricts itself to utterances obtained with the EL modification. In this case, there were too few 4-syllable prosodic groups to calculate a meaningful average. Figure 5.20

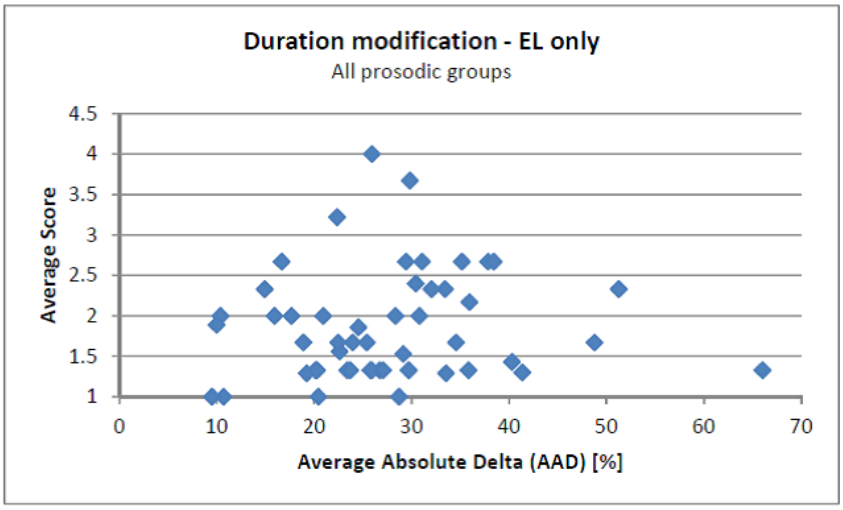



**Figure 5.18:** Average perceptual score for EL duration modification applied to both single-syllable and multiple-syllable prosodic groups as a function of average absolute delta (AAD) as defined in Equation 5.1. The average score for all utterances depicted in this figure is 1.87.

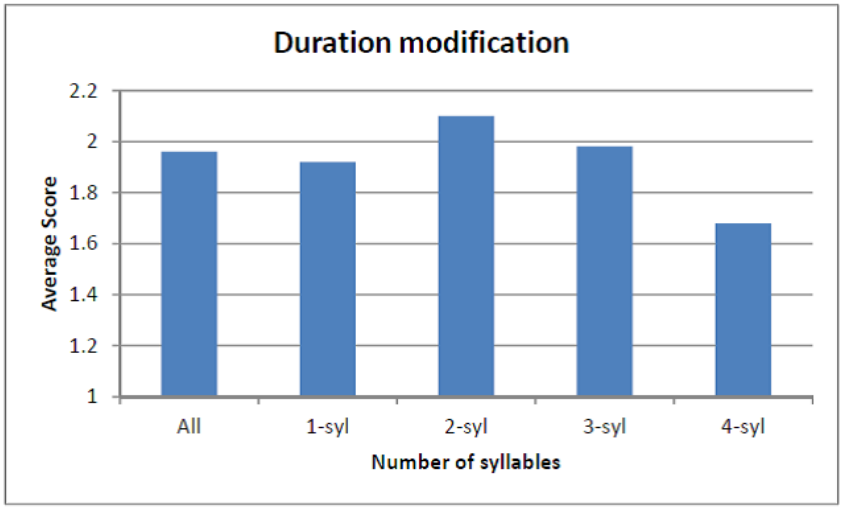



**Figure 5.19:** Comparison of average perceptual score based on number of syllables per prosodic group for all types of duration modification.

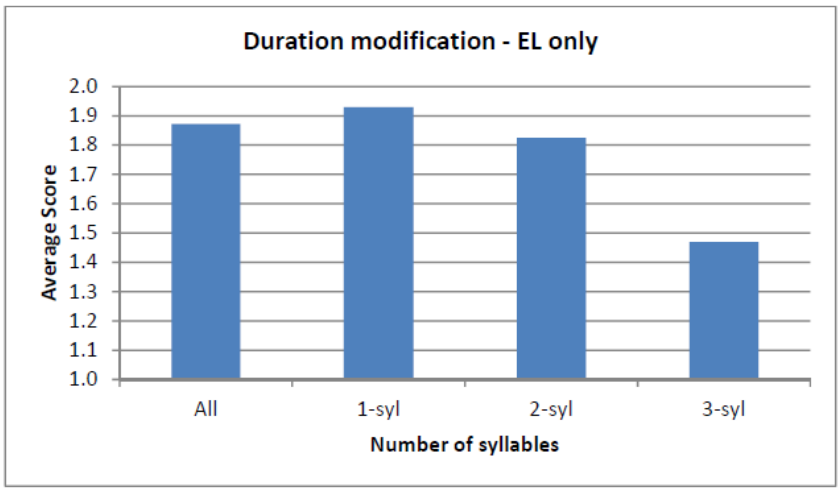



**Figure 5.20:** Comparison of average perceptual score based on number of syllables per prosodic group for the EL modification only.

again shows that the perceived quality of the modified utterance improves as the number of syllables increases. In this case, modifications to 1-syllable groups are perceptually more severe than modifications to 2-syllable groups.

#### 5.5.8.4 Analysis of amplitude modification results

Figure 5.21 shows the average scores obtained for utterances subjected to each of the amplitude modifications described in Section 5.5.5. In addition, results are shown for the same utterances when resynthesised omitting all tone commands. This represents a case in which tone is not modelled. Modifications were performed for prosodic groups between one and four syllables in length. The figure shows the average overall lengths, since the trend was found to be the same for prosodic groups of any particular length.

The figure also shows that it is best to be conservative when choosing the amplitude of tone commands, since raising these to above the mean led to a noticeable degradation in the perceptual quality of the resynthesised speech. Resynthesis at the mean or less was judged during the perceptual experiment of being closer to "mother-tongue" than to "almost mother-tongue", although unaltered amplitudes still led to the best scores. When tone is not modelled at all, the perceptual scores deteriorate strongly. This indicates that both omission and over-emphasis of tones lead to a deterioration of the perceptual quality of speech.

#### 5.5.8.5 Observation of change in meaning

One of the perceptual observations made while performing the amplitude modification experiments was change in meaning to some utterances. The meaning was changed to one that is different to the genre of the corpus, weather forecasting. This observation was made by several of the test subjects. For instance, one of the phrases for evaluation was *'Ha e-be khotso'* which could mean either "Let there be peace" or "There is no peace". For the domain in question, the appropriate meaning is "Let there be peace", where this phrase is a farewell bidding in Sesotho, usually said by a presenter at the end of a (weather forecast) presentation. Phrases which exhibited a change in meaning are given in Table 5.18, with focus words in boldface.

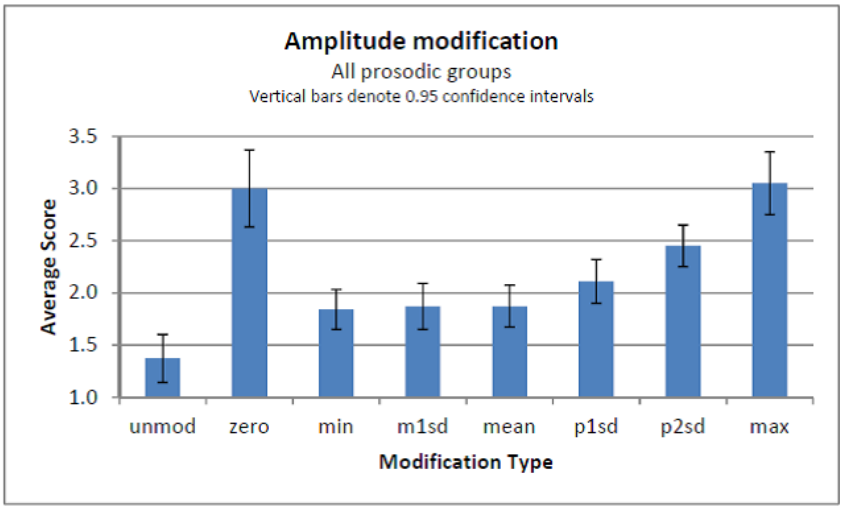



**Figure 5.21:** Average perceptual scores ascribed to utterances subjected to amplitude modification. The acronyms are min = **min**imum; m1sd = mean **m**inus **1** **s**tandard **d**eviation; p1sd = mean **p**lus **1** **s**tandard **d**eviation; p2sd = mean **p**lus **2** **s**tandard **d**eviations; max = **max**imum. The average scores given to **unmod**ified utterances and that to **zero-**amplitude modification are also shown.

**Table 5.18:** Phrases affected by amplitude modification in our corpus leading to change in meaning. Types of amplitude modification are as defined in Figure 5.21.

| Phrase | Appropriate meaning | Portrayed by ... | Alternative meaning | Portrayed by ... |
|---|---|---|---|---|
| ***Bo*** *tla rena ...* | It will dominate ... | min, m1sd, and mean | That will dominate ... | p1sd, p2sd, and max |
| *Ha e* ***be*** *khotso.* | Let there be peace. | min and m1sd | There is no peace./It doesn't bring peace. | mean, p1sd, p2sd, and max |
| ***Le*** *tla ...* | It will ... | All modifications. * | You (plural) will ... | - |
| ***E*** *tla ...* | That will ... | All modifications. * | It will ... | - |

** All modifications except for zero amplitude.*

The phrases show a grammatical tone relationship where tone is used to differentiate between a qualificative relative pronoun and a subject concord (in the first and fourth phrases); to differentiate between an imperative form and negation (in the second phrase); and to distinguish subject prefix of the third person (impersonal) and the second person (plural) (in the third phrase). The change in meaning due to amplitude modification is in line with our previous work where changes in amplitude and duration were shown to lead to a change in meaning for several minimal pairs [81]. The switch in meaning occurs at different threshold modification levels, for instance, the threshold for the first phrase is at the mean value whereas that for the second phrase is at a value of mean minus one standard deviation. However, the meanings of the last two phrases in the table were not affected by the tonal modifications deployed in this case. These are examples of minimal pairs where critical words only differ with respect to the vowel, *e* in this case, where there is an open vowel [E] and a closed vowel [e]. This was also corroborated by Mixdorff *et al.* [81] for similar minimal pairs.

#### 5.5.8.6 Checking consistency and reliability

Consistency of subject scores on repeated phrases was also analysed to verify reliability of the scores by the subjects. Inter-rater consistency measures how well the scores of a rater agree with later scores given to the same utterance. The average measure intra-class correlation coefficient (ICC) agreement and ICC consistency obtained are 0.3442 and 0.3916 respectively. These values indicate normal acceptance levels of agreement. Some of the subjects rated the "inappropriate" or "not-fitting" minimal pairs low; their decision in this instance was based on context rather than pure naturalness of the stimuli.

#### 5.5.8.7 Simultaneous modification of both duration and amplitude

The results in Section 5.5.8 indicated that the EL duration modification and a tone command amplitude that is at or slightly below the mean led to the smallest deterioration in perceived quality of the resynthesised utterances. For duration modification, EL was favoured over MM as it is easier to align onset and offset timings to syllable boundaries. In this subsection, both duration and amplitude modifications were applied simultaneously, and the resulting degradation measured by perceptual testing.

***Data selection and modification***

A set of 80 utterances with high-tone prosodic groups was selected from our corpus. These utterances consisted of a combination of single-syllable and multiple-syllable prosodic groups. For all prosodic groups, the EL duration modification was applied, and tone command amplitudes were set to the mean value, 0.2657. These modified tone command parameters were used to resynthesise each utterance using the Fujisaki model. The data prepared for perceptual evaluation consisted of 20 original utterances, and 60 with combined duration and amplitude modification. Perceptual evaluation was performed in the same manner applied in Section 5.5.7, by twenty native speakers of Sesotho at the University of Stellenbosch.

***Combined modification results***

Figure 5.22 compares the effect of the duration, amplitude, and simultaneous amplitude and duration modifications. As expected, unmodified utterances had the best score which indicates the best naturalness. Utterances subjected to combined amplitude and duration modification were awarded scores that were not significantly different from individual amplitude and duration modification. Utterances for which both duration and amplitude were modified achieved an average score of 1.77, while unmodified utterances achieved 1.38 and utterances for which tone was not modelled at all a score of 3.0.

### 5.5.9 Results and Analysis of Mismatched Syllables

Figure 5.23 illustrates the average perceptual scores resulting from the two types of modification of the mismatched syllable cases. Unmodified phrases are rated to have the most naturalness, although interestingly they were usually not awarded a score of "1". The figure also shows that, on average, removal of a Fujisaki tone command is less detrimental to perceived quality than the insertion of a tone command. However, even in the latter case, the perceived score is just above 2, corresponding to a quality between "mother tongue" and "almost mother tongue".

From the FHSL data described in Table 5.12, a subset consisting of cases where a Fujisaki tone command was associated with a sequence of between 1 and 3 low surface tone syllables was isolated. This data is summarised in Table 5.19, while Figure 5.24 shows the corresponding results of the perceptual evaluation of the modified phrases. For each case, the Fujisaki tone command amplitude was set to zero

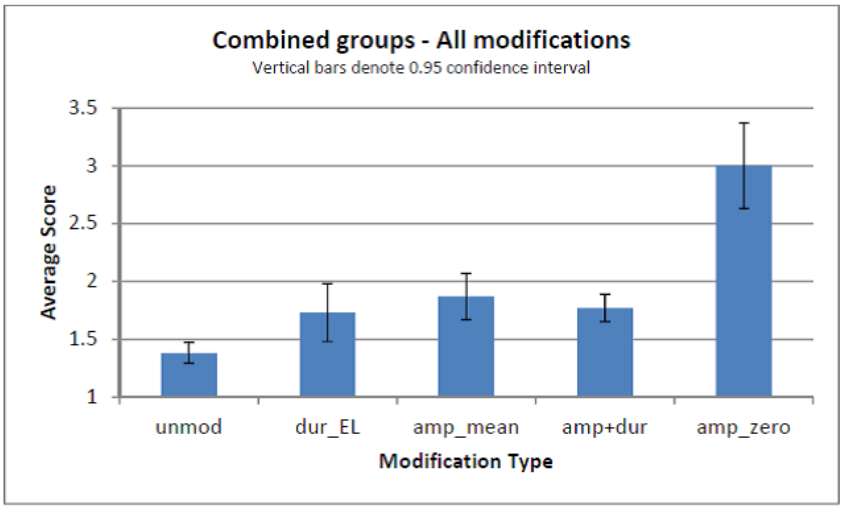



**Figure 5.22:** Comparison of average scores for individual duration and amplitude modifications (dur_EL and amp_mean respectively) based on their ideal values, and a simultaneous amplitude and duration modification (amp+dur). Average perceptual scores for unmodified utterances (unmod) and tone commands at zero amplitude (amp_zero) are also shown.

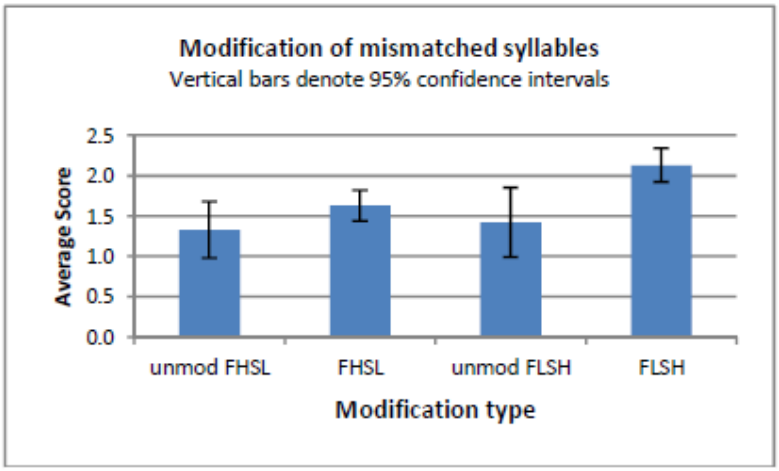



**Figure 5.23:** Comparing each type of modification with its respective unmodified counterpart, unmod FHSL for the Fujisaki high tone case and unmod FLSH for Fujisaki low tone case.

**Table 5.19:** Instances in which Fujisaki tone commands correspond to consecutive low surface tone syllables.

| Number of consecutive low surface tone syllables | Number of cases |
|---|---|
| 1 | 20 |
| 2 | 8 |
| 3 | 4 |
| **Total** | **32** |

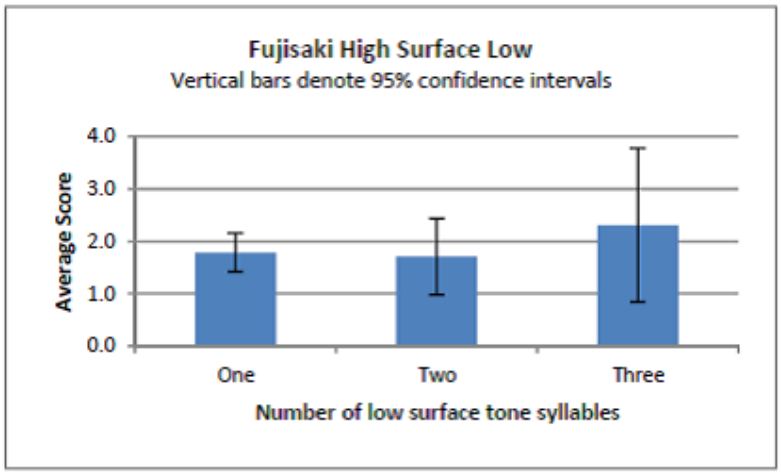



**Figure 5.24:** Average perceptual scores when removing the Fujisaki tone commands associated with 1, 2, and 3 consecutive low surface tones.

for one, two, or three consecutive syllables. The results show that 1- and 2-syllable modifications are rated similarly. Although 3-syllable modifications show larger degradation, the reliability of this average is lower due to the small number of samples (4).

Figure 5.25 shows the results of the perceptual assessment for the FLSH modifications described in Section 5.5.6. A tone command created for three syllables gave the best naturalness, closely followed by that for a single syllable. However, overall results for all numbers of syllables are similar.

Finally, each of the 127 mismatches described in Tables 5.12 and 5.13 was considered individually in order to determine the source of the discrepancy between the surface tone transcription and the Fujisaki model tone commands. Tables 5.20 and 5.21 describe the results of this investigation. The totals exceed the values in Tables 5.12 and 5.13 because each mismatch can be due to more than one factor. The tables show that tone sandhi, OCP and peak delay are major contributors of mismatches, while incorrect dictionary entries are less so. One phenomenon not yet taken into consideration in the deployment of both methods is downstep, which is to be explored for future work. Figure 5.26 illustrates the perceptual score due to the mismatch effect by each phenomenon on naturalness. Overall, the discrepancies where the Fujisaki tone command was inserted significantly affect naturalness more than when the tone command was removed.

### 5.5.10 Discussion

The stimuli of high-tone prosodic groups and of cases where the surface tone and the Fujisaki tone commands mismatch were explored. For high-tone prosodic groups, the duration and the amplitude of the Fujisaki tone commands were modified in various ways, and the speech subsequently resynthesised.

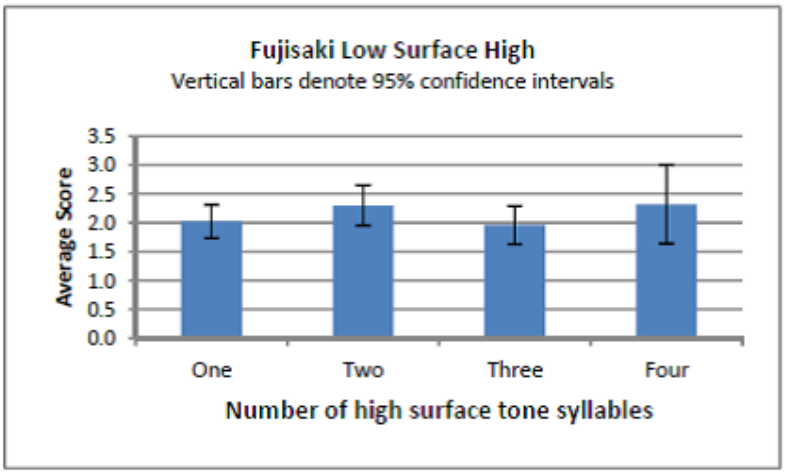



**Figure 5.25:** Average perceptual scores when inserting Fujisaki tone commands for 1, 2, 3, and 4 consecutive high surface syllables.

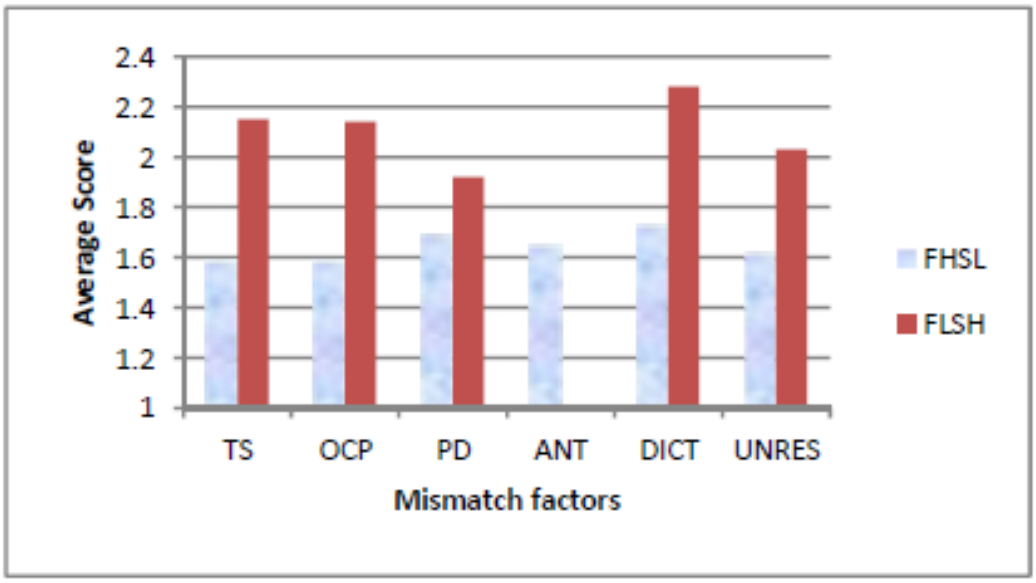



**Figure 5.26:** Sources of mismatches between surface tone and Fujisaki tone commands, and their effect on naturalness. TS = tone sandhi, OCP = Obligatory Contour Principle, PD = peak delay, ANT = anticipation, DICT = incorrect tone in dictionary, and UNRES = unresolved cases.

**Table 5.20:** Mismatches in the FHSL case.

| Description | Number of instances |
|---|---|
| Tone sandhi is observed by the Fujisaki analysis but not by the relevant surface tone rules. | 33 |
| Surface tone transcription observes OCP but Fujisaki analysis does not. | 38 |
| Peak delay is observed by the Fujisaki analysis but not by the relevant surface tone transcription rules. (The FR rules observes peak delay only for relative verbs and for ultimate syllables whose lexical tone is high.) | 18 |
| Anticipation is observed by the Fujisaki analysis but it is not modelled by the surface tone transcription rules. | 19 |
| Incorrect dictionary tone | 10 |
| Unresolved | 5 |
| **Total** | **123** |

**Table 5.21:** Mismatches in the FLSH case.

| Description | Number of instances |
|---|---|
| Tone sandhi is observed by the surface tonal transcription but not by the Fujisaki analysis. | 40 |
| OCP is violated in the surface tonal transcription by adjacent high-tone syllables, but not violated by the Fujisaki analysis. | 38 |
| The Fujisaki analysis does not observe peak delay while the surface tonal transcription does. | 23 |
| Incorrect dictionary tone | 18 |
| Unresolved | 5 |
| **Total** | **124** |

The resynthesised speech was then subjected to a perceptual evaluation to gauge its naturalness. This evaluation indicates that best results are obtained when aligning the tone commands with the syllable boundaries, and when using conservative tone command amplitudes. When this is done, the resultant resynthesised speech is judged to be midway between "mother tongue" and "almost mother tongue" quality by native speakers of Sesotho. This is a very positive result as it suggests that the Fujisaki model can be used for effective prosodic modelling when employing a simple means of synthesising the tone commands.

Instances where the tones obtained by Fujisaki analysis did not agree with the tones provided by surface tonal transcription were also investigated. These discrepancies had various causes. There may have been an error in the dictionary providing the lexical tones, or even an error or omission in the current surface tonal transcription algorithm. However, it could also be that the pronunciation of the utterance in question was incorrect or ambiguous. When considering TTS for a very poorly resourced language, such as Sesotho, the reliance of imperfect resources, including dictionaries, morphological analyses and tonal rules, is inescapable. Experiments were performed to determine the effect of these mismatches on the perceived quality of resynthesised speech. Overall, it was found that the modifications applied to "rectify" the mismatches led only to mild degradation in the perceived quality of the speech. The conclusion, therefore, is that Sesotho TTS based on the Fujisaki model for tonal and prosodic modelling is feasible, even when based on imprecise resources. An analysis of the sources of the discrepancies indicates scores of errors and that much can be gained by the improvement of the surface tone transcription process.

## 5.6 Conclusions

Sesotho is a tonal language and in order to model tonal prosody, the corpus was annotated at the tonal level using surface tone transcription. A comparison of three transcription methods was performed, these being the lexical tone, the surface tone, and the perceived tone transcriptions. The perceived tones were more in agreement with the surface tones, while the lexical tones had more mismatches with the perceived tone. The discrepancies in the three methods were observed to be as follows:

- For lexical tones, the uncertainty concerning the lexical tones of proper names, the inconsistency of the tone-marked available references, uncertainties in morphological analysis, and dialectal differences between Tswana and Sesotho in morphology and tone.
- In the case of surface tones, there is the possibility of incorrect prediction due to wrong or unknown underlying high tone labels, but also of incorrect or unknown tonal rules.
- Finally, for the perceived tones, tonal transcription is hampered by the fact that there is no standard methodology for tone transcriptions based on auditory impressions. Guidelines are missing that illustrate tonal contrasts in different contexts and that offer exercises in tone marking in order to provide transcribers with relevant training. Also, the perceptual tests should be repeated by different subjects and the tone labels achieving the greatest consensus noted. This would allow greater confidence to be placed in the annotation of perceived tone.

The Fujisaki tone commands were also compared with the perceived tones and surface tones, where they showed high agreement with both transcriptions, with a higher match for perceived tones. Surface tone transcription is more reliable than perceived tone transcription due to established rules of this procedure in literature. Inspection of the relationship between surface tones and the Fujisaki tone commands revealed two distinct patterns of prosodic groups: a sequence of high surface tone syllables, and a sequence of syllables with alternating surface tone labels. These groups, referred to as prosodic groups, were characterised by a single, prolonged tone command spanning at least two syllables.

The stimuli of these groups and of cases where the surface tone and the Fujisaki tone commands mismatched were explored and modified. For high-tone prosodic groups, the duration and amplitude of the Fujisaki tone commands were modified in various ways: the onset and offset were either aligned with syllable boundaries and/or set at mid-segment of the syllable, and the amplitude was set in the range from the minimum amplitude to the average value plus two standard deviations. The evaluation of resynthesised speech from these modifications indicated that best results were obtained when aligning the tone commands with syllable boundaries, and when using conservative tone command amplitudes. For mismatched syllables, a Fujisaki tone command was inserted for high surface tone syllables and the tone command removed for low surface tone syllables. It was found that the modifications applied to "rectify" these mismatched syllables led only to mild degradation in the perceived quality of speech. The conclusion therefore is that Sesotho TTS based on the Fujisaki model for tonal and prosodic modelling is feasible, even when based on imprecise resources.

# Chapter 6

# A Prosodic Model for Sesotho

## 6.1 Introduction

The Fujisaki tone commands are not in general aligned with syllable boundaries, since they are extracted exclusively from the audio signal. For speech synthesis, however, the start and end times of the tone commands relating to the syllable boundaries must be specified. The convergence of the surface tonal transcription and the Fujisaki model is a means of achieving the correct realisation of tone in an eventual Sesotho text-to-speech system. Thus, in Section 5.5, a heuristic model was derived where the onset, T1, offset, T2, and the tone command amplitude, At, of the Fujisaki model were modified as a means of determining appropriate values for a natural speech output from the Sesotho orthography. This work is extended in the current chapter whereby a statistical prosodic model for the determination of these quantities is developed. The purpose of this model is to predict the onset, offset and amplitude of the tone commands from prosodically-marked Sesotho text.

Before delving into the development of the model, factors which influence T1 and T2 were investigated in Section 6.2. According to Mixdorff [75], T1 can be predicted accurately either with respect to the syllable onset or the onset of the nuclear vowel, provided the internal timing of the syllable is known. Otherwise, the syllable onset is the more appropriate reference. T1 is also influenced by the type of the consonant in the onset of the syllable. T2 coincides with the segmental offset of the syllable.

In Section 6.3, T1, T2 and At were statistically predicted from a given set of Sesotho sentences with the use of regression trees. This led to the development of a prosodic model that can determine the parameters of a Fujisaki tone command from Sesotho text which is prosodically marked, i.e., that includes surface tone labelling. In Section 5.5, we showed that the simultaneous modification of duration and amplitude of the tone commands on the basis of some simple heuristics led to a mild deterioration in the naturalness of the speech output. The resulting utterances were rated as having almost "mother tongue" quality. In the current chapter, the naturalness of speech output from the model described in Section 6.3 is compared with that of Section 5.5, and also with the baseline Fujisaki-modelled utterances. All utterances and stimuli were resynthesised using the FujiParaEditor [77]. The comparison was achieved by performing perceptual evaluation tests.

As pointed out in Section 2.4, the syllable is the prosodic unit in Sesotho and thus the basic feature on which this study is based. It plays a crucial role in the production as well as perception of speech. Furthermore, the syllable duration has an influence on the F0 contour and the accurate prediction of syllable durations must be regarded as the centrepiece of a prosodic model [75]. It was thus imperative to explore the syllable duration in relation to the predicted values of T1, T2 and At in Section 6.4.

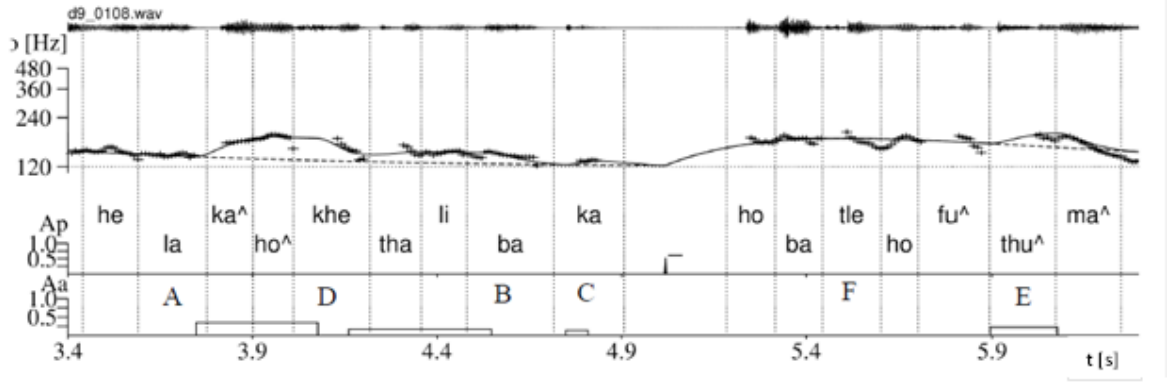



**Figure 6.1:** The onset and offset of tone commands at strategic positions within syllable segments depicted by symbols A - F. The vertical dotted lines represent syllable boundaries.

## 6.2 The Influence of the Syllable Structure on the F0 Contour

### 6.2.1 Aim

This experiment investigated the influence of the syllable structure on the onset and offset times of the tone commands of the Fujisaki model. The syllable structure aspects which were explored are the type of the syllable and the type of the constituent consonant. Factors which influence T1 and T2 were examined. In this experiment, the onset of the nuclear vowel was excluded.

### 6.2.2 Speech Data and Method of Analysis

A corpus of 150 utterances, which comprise a sum of 2 953 syllables, was used for this experiment. This dataset is a sub-corpus of that compiled in Chapter 5.

Analysis of the Fujisaki tone commands in previous experiments has revealed syllables at six strategic points, as illustrated in Figure 6.1. Point A shows the onset, T1, of a tone command spanning multiple syllables while B shows the offset, T2, of another multiple-syllable tone command. Point C shows T1 and T2 within one syllable segment. This is a monosyllabic tone command. Point D shows T1 and T2 within a single syllable segment, but T1 and T2 belong to two separate tone commands. Point E shows T1 and T2 coinciding with syllable boundaries. Point F shows a syllable with no tone command. In this experiment, syllables at points A to E were considered. Those at point F were excluded due to the absence of a tone command.

In our corpus, there were 560 instances of case A, 564 of case B, 53 of case C, 182 of case D, and 976 of case E.

In order to examine the temporal alignment of tone commands, the following parameters were investigated, as illustrated in Figure 6.2.

- T1 with respect to syllable onset (a, c, and e in the figure). The points a, c, and e depict the onset of a tone command after segmental onset.
- T2 with respect to segmental offset of the syllable (b, d, and f in the figure). The points b, d, and f denote the offset of a tone command before segmental offset.
- T1_dist = T1 distance expressed as the fraction of the toned syllable duration.

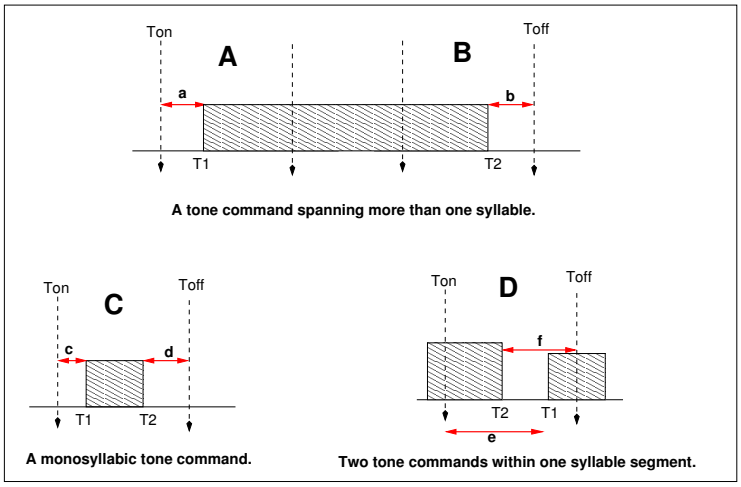



**Figure 6.2:** Parameters being investigated for temporal alignment of tone commands. These are a, b, c, d, e, and f. The parameters a, c, and e are the distances between the syllable onset and tone command onset, T1, while b, d, and f are the distances between tone command offset, T2 and syllable offset. A, B, C, and D are the cases (strategic points) of the tone commands described in Section 6.2.2. Only case E is not illustrated in the figure due to minimal temporal parameters. Vertical dotted lines represent syllable boundaries. Ton = syllable segment onset, Toff = syllable segment offset, T1 = tone command onset, T2 = tone command offset, Toff - Ton = syllable duration.

- T2_dist = T2 distance expressed as the fraction of the toned syllable duration.
- Syllable duration associated with each case.
- Types of consonants and types of syllables associated with each case.

### 6.2.3 Results of Analysis

Tables 6.1 to 6.3 depict values of the above parameters determined from our corpus. Table 6.1 shows the mean and standard deviation values for the five cases, with metrics for syllable duration only for case E as other parameters are not applicable. Table 6.2 shows the types of syllables prevalent in each case, while Table 6.3 shows the consonant types associated with the same groups.

In Table 6.1, case C, which depicts monosyllabic tone commands, has the smallest distance between the syllable onset and tone command onset, c, and syllable offset and tone command offset, d, while case D, where there are two tone commands within a syllable segment, shows the largest distance. This means that, on average, for monosyllabic tone commands, T1 is triggered almost immediately after the syllable onset, within approximately 20% (T1_dist for c) of the syllable segment. The F0 contour starts to fall at a point approximately 80% into the same syllable segment (T2_dist for d). For a tone command spanning more than one syllable, both T1 and T2 are triggered approximately mid-syllable segment; T1 in the first syllable segment and T2 in the last syllable segment of the tone command. In cases where T1 and T2 of two tone commands are within one syllable segment (case D), T1 occurs approximately 68% into the syllable and this is subsequent to the F0 peak drop at 16% of the previous tone command. In terms of syllable duration, the monosyllabic tone commands have the longest duration while case E, where T1 and T2 are aligned with syllable boundaries, has the shortest syllable duration. Interestingly, syllable duration for a multiple-syllable tone command is almost the same for syllables where T1 and T2 occur. Syllable duration for case D is just greater than that of cases A and B.

**Table 6.1:** Parameters investigated associated with cases A - E as shown in Figure 6.2. Other measurements are not applicable to case E except for syllable duration.

| Case | Parameter | | | T1_dist (in bold) & T2_dist | | Syllable duration | |
|---|---|---|---|---|---|---|---|
| | | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ |
| A | a | 0.0797 | 0.0596 | **0.4707** | **0.3097** | 0.1704 | 0.0513 |
| B | b | 0.0925 | 0.0616 | 0.5162 | 0.2898 | 0.1787 | 0.0568 |
| C | c | 0.0459 | 0.0410 | **0.2063** | **0.1653** | 0.2208 | 0.0518 |
| | d | 0.0445 | 0.0369 | 0.1913 | 0.1380 | | |
| D | e | 0.1247 | 0.0579 | **0.6841** | **0.2374** | 0.1813 | 0.0532 |
| | f | 0.2985 | 0.1534 | 1.8420 | 1.1520 | | |
| E | - | - | - | - | - | 0.1398 | 0.0431 |

**Table 6.2:** Types of syllables associated with cases A - E. V denotes vowels, C consonants, and CV a vowel-consonant combination type.

| Syllable type | A | B | C | D | E | Overall | Overall (incl. syls with no tone commands) |
|---|---|---|---|---|---|---|---|
| V | 36 | 44 | 5 | 12 | 159 | 256 | 307 |
| C | 55 | 34 | 6 | 6 | 104 | 205 | 273 |
| CV | 468 | 481 | 41 | 164 | 711 | 1865 | 2364 |

The overall observation in Table 6.2 is that CV-type syllables are dominant in all cases, covering at least 66% of the corpus. These are followed by vowels (V-type) and then consonants (C-type). In terms of the consonant-only syllables, case A is the only one from which the number of consonants significantly exceeds the number of vowels. On the other hand, for case C, the number of vowels and consonants match closely. The values for cases A and B show that more consonants are likely to be found at T1 while T2 is dominated by vowels. CV-type syllables are more prevalent for T2 than for T1.

Different types of consonants (as per Guma [44]) associated with each case are given in Table 6.3. Overall, more unvoiced than voiced consonants are observed. However, cases A and C indicate the opposite. Plosives, bilabials, laterals and ejectives are less frequent. Clicks (e.g. *q* in *Moqebelo*) and pre-palatals (e.g. *ny* in *monyetla*) are the least frequent. Case E has the most frequent number of each type except for the alveolar, for which case A has the most frequent number, though the two values are close. Cases A and C include the largest number of plosives, while cases B and D include the largest number of fricatives. T1 of a multiple-syllable tone command is ruled by voiced consonants whereas T2 is ruled by unvoiced consonants.

### 6.2.4 Discussion

With case E having the smallest syllable duration, it is expected that T1 and T2 will align with syllable boundaries, because the F0 contour in this case does not have enough room to rise and eventually fall. In addition, the high number of instances of case E is in line with results from Section 5.5 where the alignment of T1 and T2 with syllable boundaries produced more natural-sounding speech compared with other modifications. Long syllable durations are also not surprising for monosyllabic tone commands, as this provides more time for the rising and falling of the F0 contour. This observation also applies to case D, though here F0 rises almost immediately after the fall, and occurs much faster, thus syllable

**Table 6.3:** Types of consonants associated with cases A - E. These types are as per Guma [2].

| Type of consonant | A | B | C | D | E | Overall | Overall (incl. syls with no tone commands) |
|---|---|---|---|---|---|---|---|
| voiced | 304 | 254 | 40 | 87 | 445 | 1130 | 1471 |
| unvoiced | 256 | 310 | 13 | 95 | 531 | 1205 | 1482 |
| plosive | 127 | 88 | 8 | 17 | 170 | 410 | 549 |
| click | 3 | 0 | 0 | 0 | 5 | 8 | 9 |
| fricative | 73 | 109 | 7 | 36 | 156 | 381 | 489 |
| lateral | 67 | 38 | 5 | 28 | 99 | 237 | 312 |
| rolled | 7 | 19 | 2 | 6 | 27 | 61 | 89 |
| bilabial | 66 | 83 | 6 | 22 | 84 | 261 | 327 |
| alveolar | 59 | 27 | 7 | 13 | 57 | 163 | 206 |
| pre-palatal | 0 | 0 | 0 | 0 | 1 | 1 | 6 |
| velar | 31 | 25 | 6 | 7 | 72 | 141 | 181 |
| ejective | 48 | 80 | 0 | 24 | 52 | 204 | 234 |
| aspirated | 17 | 18 | 0 | 13 | 43 | 91 | 99 |
| flapped | 19 | 26 | 6 | 2 | 49 | 102 | 121 |

duration is shorter than for case C. Results also show, for a multiple-syllable tone command, T1 and T2 are triggered at mid-syllable and the syllable durations where these occur are almost equal. So T1 and T2 occur within syllable segments of approximately equal durations. For monosyllabic tone commands, T1 is observed 20% into the syllable segment and T2 is realised 20% before the end of the segment.

The observation of syllable types indicates that, for a multiple-syllable tone command, T1 is mostly realised in a C-type syllable and T2 in a V-type. The CV-type is dominant and T1 and T2 are almost equally likely in the two. The possibility of occurrence of vowels or consonants at T1 and T2 within a monosyllabic tone command is equally likely. Alignment with syllable boundaries has a high probability for CV-type syllables, and is twice as likely for vowels than for consonants in case D. Case E is more likely to be associated with a CV-type, followed by vowels, then consonants.

In Table 6.3, for multiple-syllable tone commands, there is a higher likelihood for realisation of voiced consonants at T1 and higher likelihood for unvoiced consonants at T2. Plosives are more likely to be observed at T1 and fricatives have a more frequent occurrence at T2. Two different tone commands are likely to occur in a syllable with a fricative consonant followed by a lateral but will not be realised in a syllable with a pre-palatal or a click. A monosyllabic tone command will not be realised in a click, a pre-palatal, an ejective, or an aspirated consonant. Most realisations are seen for plosives, fricatives and alveolar consonants. The alignment with syllable boundaries is most likely to occur for plosives, followed by fricatives. The alignment is least likely to occur for clicks and pre-palatals.

It must be noted that the results above are based on a corpus in the weather genre, a limited domain entity. In this case, there is a high likelihood of repeated words and phrases, which do not provide a phonetically or syllabically balanced corpus. An open domain corpus would be necessary to verify the above results.

## 6.3 A Statistical Prosodic Model for Sesotho

### 6.3.1 Aim

The aim of this experiment was to derive a statistical model which will automatically determine the F0 pattern of Sesotho utterances in terms of the Fujisaki model. The parameters to be predicted were T1, T2 and At of the Fujisaki tone commands.

### 6.3.2 Data and Evaluation Method

A selection of 150 utterances for the training set and 50 utterances for the test set from the main corpus was made for the purpose of this experiment. The selection avoided repeated utterances, as are common in collected data from a weather report.

Table 6.4 lists the predictor variables which were investigated in order to predict T1, T2 and At. These are:

- The syllable, *syl*, and the syllable type, *syl_type*.
- The manner and place of articulation, *vibration* and *articulation* respectively.
- The tone level of the current syllable, *curr_syl_tone*, the previous syllable, *prev_syl_tone*, and the following syllable, *next_syl_tone*.
- The length of the word in terms of the number of syllables, *wrd_dur*.
- The position of a syllable in a word, *syl_in_wrd*, in a phrase, *syl_in_phr*, and in a sentence, *syl_in_sent*.
- The position of the current syllable with respect to the previous high-tone syllable, *syl_psn_prev_H*, and with respect to the next high-tone syllable, *syl_psn_next_H*.
- The word, *word*, and the part of speech each word belongs to, *pos*.

Three regression trees for T1, T2 and At were trained and implemented using WEKA, a freely-available data mining software package developed at the University of Waikato [3]. The training was performed with an M5P pruned model tree which uses smoothed linear models.

The approximation methods used in our experiment are different from those used by Teixeira *et al.* [102], Mixdorff [75] and Navas *et al.* [85]. Teixeira *et al.* consider each tone command to be associated with each single syllable while Mixdorff and Navas *et al.* consider one tone command per tone group. In our study, prediction of tone commands (also referred to as prosodic groups) which span more than one syllable as well as of monosyllabic tone commands was considered. This is in line with our observations of tone commands in Chapter 5.

The statistical modelling was done in an interdependent and accumulative manner, which means that the approximation of one target variable was based on one or the other predicted parameter(s). Preliminary experiments where independent predictions were performed gave reasonable results (correlation coefficient (CC) values of 0.8774, 0.8768, 0.7488 and root mean square errors (RMSE) of 1.0153, 1.0642, and 0.0893 for T1, T2 and At respectively). However, these positive indications were not matched by the perceptual tests. The interdependent approach, on the other hand, showed better performance.

Our approach was to predict T1 first, followed by T2, and then At. Predicted T1 values were added as an additional input variable to the training data set for the prediction of T2. For the approximation

**Table 6.4:** Attributes for prosody model prediction of T1, T2, and At. The values are: C = consonant, V = vowel, CV = a syllable type that consists of a consonant and a vowel, U = unvoiced, V = voiced, PL= plosive, CL = click, FR = fricative, LT = lateral, R = rolled, BL = bilabial, AL = alveolar, PR = pre-palatal, VL = velar, EJ = ejective, AS = aspirated, FL = flapped, H = high tone, L = low tone. The articulation and vibration variables are as described by Guma [44].

| Attribute | Value |
|---|---|
| syl | All syllables in the dataset, e.g. mo, nye, tla. |
| syl_type | C, V, CV |
| vibration | U, V |
| articulation | PL, CL, FR, LT, R, BL, AL, PR, VL, EJ, AS, FL |
| curr_syl_tone | H, L |
| prev_syl_tone | H, L |
| next_syl_tone | H, L |
| wrd_dur | Numeric value starting at 1. |
| syl_in_wrd | Numeric value starting at 1. |
| syl_in_phr | Numeric value starting at 1. |
| syl_in_sent | Numeric value starting at 1. |
| syl_psn_prev_H | Numeric value starting at 1. |
| syl_psn_next_H | Numeric value starting at 1. |
| word | All words in the dataset. |
| pos | Part of speech for each word. |

of At, both the predicted T1 and T2 values were included as input variables in the training data. During the modelling process, T1 and T2 for tone commands spanning more than one syllable were checked to verify that they did not overlap. In this way, it was ensured that T2 of a preceding tone command coincided with T1 of the following tone command.

The section that follows presents the results obtained from the three regression trees.

### 6.3.3 Prediction Results

Prediction metrics for different test options are given in Table 6.5. As expected, prediction is best for the training set (in terms of the highest CC between actual and predicted values, and the lowest RMSE). There were 3 359 instances in the training set and 994 for the test set.

Table 6.6 shows the results in terms of CC and RMSE results between the baseline, the prediction model (from the regression trees) and the heuristic model. The CC is a value between -1 and 1 that indicates the strength and direction of the linear relationship between two variables. A value close to 1 indicates a strong positive linear relationship between the variables, a zero means no linear relationship, and a value close to -1 denotes a strong negative linear relationship. Negative values should not occur for reasonable prediction methods. RMSE, on the other hand, is based on residuals. Residuals are the difference between the actual values and the predicted values. RMSE is calculated by using the formula

$$\sqrt{\frac{\sum_{i=1}^{n}(y_i - x_i)}{n}},$$

where $x_i$ denotes the actual value for the $i^{th}$ observation, $y_i$ is the predicted value, $n$ is the total number of values. Residuals can be positive or negative as the predicted value under- or over-estimates the actual value. RMSE is used as a measure of the spread of the $y$ values around the actual $x$ values. The smaller the RMSE value, the more preferred (or the better the prediction) as there is less deviation from the actual values. In general, good performance is indicated by a large positive value of the CC and a small value of the RMSE.

**Table 6.5:** Performance metrics for prediction results on three different test options.

| Performance measure | Training set | | | Test set | | | Cross-validation of the training set (10-fold) | | |
|---|---|---|---|---|---|---|---|---|---|
| | T1 | T2 | At | T1 | T2 | At | T1 | T2 | At |
| CC | 0.929 | 0.9473 | 0.8107 | 0.5779 | 0.551 | 0.3642 | 0.7395 | 0.9162 | 0.6695 |
| RMSE | 0.7846 | 0.708 | 0.0789 | 3.9294 | 4.287 | 0.1497 | 1.4584 | 0.8868 | 0.1016 |

**Table 6.6:** Correlation coefficient (CC) and RMSE performance values for the baseline, the regression tree (prediction) model and the heuristic model.

| | CC | | | RMSE | | |
|---|---|---|---|---|---|---|
| | T1 | T2 | At | T1 | T2 | At |
| Baseline vs regression tree | 0.9290 | 0.9089 | 0.7751 | 0.7846 | 0.9231 | 0.0851 |
| Baseline vs heuristic model | 0.8588 | 0.8532 | 0.3411 | 0.0649 | 0.0638 | 0.0279 |

In Table 6.6, as expected, the correlation between the baseline and the regression tree values show a stronger linear relationship than that between the baseline and the heuristic model. This is observed across the three parameters, T1, T2 and At, though the correlation coefficient values for At are smaller, with a weak relationship of 0.3411 between the baseline and the heuristic model. Conversely, the RMSE indicates less deviation of values between the baseline and heuristic model than that between the baseline and the prediction model.

#### Variable Ranking

The importance of each variable for the prediction of T1, T2 and At was evaluated and the summary of the ranking is shown in Table 6.7. Ranking is an indication of the use of the variable in the regression tree. A ranking value of 1 means the most used variable in the regression tree, thus, the larger the ranking value, the less its use in the tree. These rankings are listed for each target variable. In the table, the most important variable for prediction of T1 is the position of the syllable in a sentence, followed by the knowledge of the tone level of the next syllable. Prediction of T2 is highly dependent on the predicted value of T1, followed by knowing what the tone level of the next syllable is. In order to predict At, six factors are highly and equally ranked. These are *syl_type*, *vibration*, *curr_syl_tone*, *syl_in_word*, *word*, and *predicted_T2*. The variable *next_syl_tone* is important in the prediction of both T1 and T2. The high ranking of T1 for prediction of T2, and that of predicted T2 for determination of At show the interdependence of these features.

### 6.3.4 Perceptual Evaluation and Results

In order to evaluate the effectiveness of the prediction model, a perceptual experiment was conducted with resynthesised stimuli produced with the Fujisaki model. The effectiveness of the model was compared with the baseline and the heuristic approach described in Chapter 5.

Twenty-one utterances were randomly selected for perceptual evaluation tests. The stimuli were from unmodified utterances from the baseline, those produced from a heuristic model, and those from the regression tree. Each stimulus was paired with each of the other two approaches and this resulted in 63 pairs of stimuli.

The perceptual evaluation tests were undertaken by 14 native speakers of Sesotho at the University

**Table 6.7:** Ranking of variables in the regression tree for prediction of T1, T2 and At.

| Target variable | Predictor variable | Rank in the regression tree |
|---|---|---|
| T1 | syl_in_sent | 1 |
| | next_syl_tone | 2 |
| | pos | 3 |
| T2 | predicted_T1 | 1 |
| | next_syl_tone | 2 |
| | curr_syl_tone | 3 |
| | syl_type | 4 |
| At | syt_type | 1 |
| | vibration | 1 |
| | curr_syl_tone | 1 |
| | syl_in_word | 1 |
| | word | 1 |
| | predicted_T2 | 1 |
| | syl_psn_next_H | 2 |
| | syl | 3 |

**Table 6.8:** Preferences observed from the perceptual evaluations.

| | Best audio | Both natural | Both unnatural |
|---|---|---|---|
| Baseline vs regression tree | Baseline (100%) | 58% | - |
| Baseline vs heuristic model | Baseline (100%) | 25% | - |
| Regression tree vs heuristic model | Regression tree (83%) | 17% | 4% |

of Stellenbosch, six females and eight males. The subjects were offered pairs of stimuli from the 63 pairs and had to decide which version they found more natural or if both were equally (un)natural. DMDX [28], a tool designed specifically for language-processing experiments, was used to perform perceptual evaluations. Table 6.8 shows the preferences observed.

The unmodified stimuli were always preferred over the modified utterances. However, resynthesised speech from the prediction model was found to sound more natural than that from the heuristic model by 83% of the participants. There were instances where the paired models at a specific instance were both judged to produce natural-sounding speech. Although it was rare (4% of cases), the regression tree model did sometimes produce 'unnatural' speech.

### 6.3.5 Discussion

The Fujisaki tone command parameters, T1, T2 and At were predicted from tone-marked orthography using three regression tree models, one for each parameter. The prediction of T2 was dependent on the predicted value of T1, and the predicted value of At was dependent on the predicted values of both T1 and T2. The performance of the regression tree model was then compared with the heuristic model developed in Chapter 5. Performance was also tested by undertaking perceptual evaluations of resynthesised stimuli from the two models. While unmodified utterances were still unanimously preferred by more than 80% of respondents, the utterances resynthesised using the regression trees were preferred to the utterances resynthesised using heuristics. This indicates that a statistically-based generation of prosody in Sesotho for TTS is promising.

**Table 6.9:** Ranking of variables in the prediction of syllable duration from a tone-marked Sesotho orthography.

| Predictor variable | Ranking |
|---|---|
| syl_type | 1 |
| prev_syl_tone | 1 |
| syl_psn_next_H | 1 |
| word | 2 |
| pos | 2 |
| syl | 3 |
| articulation | 4 |

## 6.4 A Statistical Prediction of Syllable Duration

### 6.4.1 Aim

The Fujisaki parameters T1, T2, and At determine the alignment of tone commands with syllables. However, the length of the syllable itself is also a variable for TTS systems. In this section, the aim is to determine whether estimated T1, T2, and At values are useful for the determination of the syllable length.

### 6.4.2 Data

The same training dataset used in Section 6.3, with additional predictor variables as T1, T2 and At, was used in this experiment. The same regression tree was trained for prediction of syllable durations. The additional variables were predicted values obtained in Section 6.3.

### 6.4.3 Results and Discussion

The correlation coefficient between the baseline and predicted syllable durations is 0.755 and the RMSE is 0.0413. These metrics are based on the training data set with predicted parameters, T1, T2 and At from the previous experiment in Section 6.3. In order to check the influence of other factors in the prediction of syllable duration, their relevance was determined from the prediction model. Table 6.9 shows the ranking of the predictor variables, with the most important first.

From the table, it is evident that the type of syllable, the tone level of the previous syllable, and the position of the current syllable with respect to the following high-tone syllable are the most influential, and equally so, in the prediction of syllable duration. Conversely, place of articulation in pronouncing syllables (or words) has the least importance on the list.

To determine the effect of these variables in the prediction of syllable duration, they were each removed from the dataset and the regression tree run each time, whereby the performance metrics were calculated. Table 6.10 summarises the performance measures obtained. The table includes values for prediction with and without T1, T2 and At. Only variables whose ranking value was less than or equal to three were tested.

The presence or knowledge of T1, T2 and At has the largest correlation coefficient, an indication that these parameters are reasonably correlated with syllable duration. However, their absence does not have much impact as shown by the correlation coefficient value of 0.7405 in the table, which differs slightly from that of their inclusion in the dataset. Due to this insignificant difference, these parameters were removed completely from the dataset and then each of the other attributes were removed one at a time, in turn, and calculations performed. The variable *syl* was not excluded from the dataset though, as it is

**Table 6.10:** Performance measures of variables influential in the prediction of syllable duration.

| | Correlation coefficient | RMSE |
|---|---|---|
| With predicted T1, T2 and At | 0.755 | 0.0413 |
| Without predicted T1, T2 and At | 0.7405 | 0.0423 |
| *syl_type* removed | 0.7418 | 0.0422 |
| *prev_syl_tone* removed | 0.66 | 0.0473 |
| *syl_psn_next_H* removed | 0.6889 | 0.0456 |
| *word* removed | 0.7323 | 0.0429 |
| *pos* removed | 0.7504 | 0.0416 |

the core factor in the study.

Lack of knowledge about the tone of the previous syllable has the most impact in the determination of a syllable duration. This is followed by the attribute *syl_psn_next_H*, which means that the position of the syllable with respect to the next high tone syllable is of importance. This is in line with the ranking depicted in Table 6.9. However, in Table 6.10, the absence of the word has more impact than the type of syllable, though their values are comparable. This is to be expected in that the syllable, or its type, can only be derived from the word in which it is constituted.

In as much as the syllable structure influences T1 and T2, as was shown in Section 6.2, these parameters are not highly influential in the determination of syllable duration. The important factors are *prev_syl_tone*, s*yl_psn_next_H*, *word*, *syl_type*, and *pos*, in that order. The place of articulation in the pronunciation of syllables (or their words) is also of importance, though to a lesser extent.

## 6.5 Conclusions

The characteristics and structure of a syllable, which include the type of the syllable, the type of constituent consonant, the position of the syllable in the word or utterance, and the tone of the neighbouring syllables, can be determining factors on where the onset and offset of the tone command occur. This means these features are instrumental in predicting the F0 contour.

This knowledge, combined with other attributes of the syllable in a prosodically-marked text, can be used by a machine learning algorithm to approximate the parameters T1, T2 and At used by the Fujisaki model. Three regression trees were trained with sixteen such attributes to develop a data-driven prosodic model for the Sesotho language. Speech resynthesised using this model sounded more natural than that resynthesised using simpler heuristics. This demonstrates the potential for synthesising Sesotho speech with natural prosody from tone-marked text using data-driven techniques. The speech output could be further improved by including additional linguistic factors such as tone sandhi and downstep, and also by predicting the phrase command parameters, both in terms of magnitude Ap and position, T0. However, this is left for future exploration and investigation.

In the case of predicting syllable duration, which is also a crucial factor for the F0 contour, the most important factors are knowledge of the previous syllable tone, the position of the syllable with respect to the following high-toned syllable, the word, the syllable type, and the part of speech of the word. Unlike in the instance where syllable duration is one of the variables which are determining factors for T1 and T2, the procedure is not repeated in reverse. Knowledge of T1, T2 and At is not highly ranked for the prediction of syllable duration.

# Chapter 7

# The Fujisaki Model and HTS

## 7.1 Introduction

As already highlighted in Chapter 1, the purpose of this study was to derive a prosody modelling algorithm for Sesotho from the orthography using the Fujisaki model. This algorithm is to be integrated into a Sesotho TTS system. In the previous chapters, the Fujisaki model has been shown to be a promising candidate for this purpose. However, many current TTS systems incorporate prosody generation into hidden Markov model-based speech synthesis (HTS) [106, 117]. In this chapter, we compare these two approaches to prosody modelling for text-to-speech systems. Furthermore, this comparison is carried out for two tonal languages: Sesotho and Serbian[1].

Sesotho and Serbian belong to different families: the former is a tonal Bantu language while the latter is a pitch-accent language [96]. For both Serbian and Sesotho, the orthography does not include prosodic marking. In this study, the focus is on the modelling and prediction of F0, which is perceptually the most important element of sentence prosody.

The Fujisaki model [30] approach to prosody analysis relies only on the acoustics of uttered speech. A second approach, widely used in conjunction with hidden Markov model-based speech synthesis (HTS), employs a set of trained statistical models (context-dependent HMMs), which are used to predict prosodic parameters such as durations of phonetic segments and values of log F0, as well as spectral coefficients for synthesis. Both models rely on a previously recorded speech corpus, used to train the model, i.e. to set the values of its relevant parameters.

## 7.2 HMM-based Speech Synthesis

Statistical parametric speech synthesis uses hidden Markov models (HMMs) for speech waveform generation, and is typically called HMM-based speech synthesis. It has been demonstrated to be very effective in synthesising speech. The main advantage of this approach over other successful approaches, like unit selection, is its flexibility in changing speaker identities, emotions, and speaking styles. HMMs are used to represent not only the phoneme sequences but also prosodic information such as F0. HMMs are trained on a speech corpus in order to be able to predict values of prosodic and spectral parameters of speech for a given text. However, since it is impossible to prepare training data for all possible linguistic contexts, a

[1]The Serbian system was developed by researchers at the University of Novi Sad, Serbia, as part of a research collaboration.

number of tree-based clustering techniques have been proposed in order to allow HMMs to share model parameters among states in each cluster.

The synthesis part of the system converts a given text to be synthesised into a sequence of context-dependent labels. According to the label sequence, a sentence-level HMM is constructed by concatenating context-dependent HMMs. The duration of each state is determined to maximise its probability based on a set of state duration probability distributions. A sequence of speech parameters including spectral and excitation parameters is subsequently determined so as to maximise its output probability using a speech parameter generation algorithm [104]. Finally, a speech waveform is resynthesised directly from the generated spectral and excitation parameters by using a speech synthesis filter, such as the mel-log spectral approximation filter for mel-cepstral coefficients and the all-pole filter for linear-prediction-based spectral parameter coefficients [105].

## 7.3 The Serbian Language

Serbian is the standarised variety of the Serbo-Croatian language, spoken in Serbia as the official language, and in some other countries of the Balkan peninsula as well. Serbo-Croatian is the only Slavic language which uses a pitch accent, assigning it at the level of the lexicon and using it to differentiate between word meanings or values of morphological categories. Traditional grammars define the pitch accent of Serbo-Croatian through four distinct accent types, which involve a rise or fall in pitch associated to either long or short vowels, and with optional post-accent lengths. However, more recent analyses [43, 94] have shown that these accent types can be interpreted as tonal sequences, i.e. reduced to sequences of high and low tones, without loss of representativeness, provided that phonemic length contrast is preserved. Thus, words can be thought of as strings of syllables following tonal patterns, and the surface tone of the utterance can be derived from its underlying tone using appropriate tonal rules.

## 7.4 Speech Data

The Sesotho corpus contained 40 minutes of speech and the utterances were 12 seconds long on average. These utterances were taken from the main corpus of the weather forecast described in Section 5.2.

The Serbian speech corpus contained approximately four hours of speech, recorded in a sound-proof studio and sampled at 44 kHz. All sentences were uttered by a single female voice talent, a professional radio announcer using the ekavian standard pronunciation. General intonation in the database ranged from neutral to moderately expressive, and the effort was made to keep the speech rate approximately constant. However, in order to avoid considerable difference in the experimental setup for Serbian and Sesotho, only a 40-minute portion of the corpus corresponding in size to the entire Sesotho corpus was used for the principal experiment involving the comparison between the Fujisaki model and HMM prosody generation for both languages.

## 7.5 Surface Tone Transcription and Fujisaki Analysis

The method followed for surface tone transcription of Sesotho data was as stipulated in Section 2.5.3.

The sequence of tones for the Serbian corpus was determined based on the pitch accent assigned by the system for automatic part-of-speech (POS) tagging [95], with tagging errors manually corrected. Appropriate tonal rules were used to convert the underlying tone to surface tone.

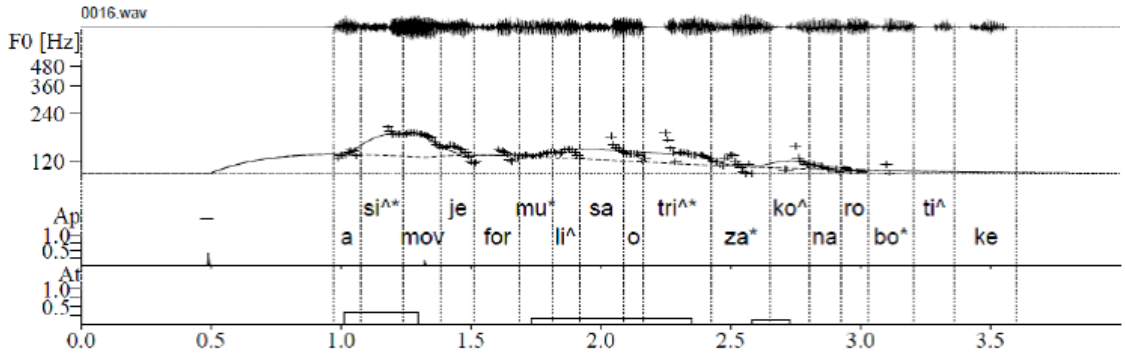



**Figure 7.1:** Serbian sentence illustrating the Fujisaki-modelled F0 contour and its surface tone labels. The high surface tones are indicated by the symbol '^' and stress is indicated by the symbol '*'. Vertical dotted lines mark syllable boundaries. The sentence reads '*Asimov je formulisao tri zakona robotike.*' - "Asimov formulated the three laws of robotics."

**Table 7.1:** Comparison between original sentences and those resynthesized by a Fujisaki model.

| | Duration (ms) | | Pitch (Hz) | |
|---|---|---|---|---|
| | Sesotho | Serbian | Sesotho | Serbian |
| RMSE | 0.00 | 0.00 | 9.18 | 14.41 |
| MAE | 0.00 | 0.00 | 6.28 | 8.38 |
| CC | 1.00 | 1.00 | 0.89 | 0.63 |

**Table 7.2:** Comparison between original utterances and those synthesized by a HTS system.

| | Duration (ms) | | Pitch (Hz) | |
|---|---|---|---|---|
| | Sesotho | Serbian | Sesotho | Serbian |
| RMSE | 35.76 | 22.39 | 23.41 | 16.10 |
| MAE | 28.14 | 17.27 | 18.65 | 12.28 |
| CC | 0.27 | 0.74 | 0.03 | 0.59 |

Once the surface tone transcription was complete for both languages, the sentences were annotated at syllable levels using Praat TextGrid editor [12]. F0 values were extracted using Praat at a step of 10 ms and inspected for errors. The F0 tracks were subsequently decomposed into their Fujisaki components applying an automatic method originally developed for German [73]. Initial experiments in [82] for Sesotho have shown that the low tones in the critical words of the minimal pairs could be modelled with sufficient accuracy without employing negative tone commands. Serbian, as illustrated in Figure 7.1, also shows positive tone commands only. As a consequence, only high tones were associated with tone commands for both languages. Adopting this rationale, automatically calculated parameters were viewed in the FujiParaEditor [77] and corrected when necessary, as stipulated in Section 3.4.

## 7.6 Experimental Results

Ten utterances from each language were selected for testing by both models. The resulting synthesised data were then compared with original utterances, where root mean squared error (RMSE), mean absolute error (MAE), and correlation coefficient (CC) were calculated for duration and pitch. Tables 7.1, 7.2, and 7.3 show the results obtained from these calculations.

**Table 7.3:** Comparison between utterances resynthesized by a Fujisaki model and those synthesized by a HTS system.

| | Duration (ms) | | Pitch (Hz) | |
|---|---|---|---|---|
| | Sesotho | Serbian | Sesotho | Serbian |
| RMSE | 35.76 | 22.39 | 23.75 | 17.54 |
| MAE | 28.14 | 17.27 | 19.15 | 14.57 |
| CC | 0.27 | 0.74 | -0.04 | 0.37 |

In Table 7.1, both languages show a positive high correlation for both duration and pitch, with Sesotho presenting a closer relationship than Serbian. Duration displays perfect correlation of the value of 1 for the two languages, which suggests that the duration of the utterances was not affected during resynthesis. This is also an indication that the F0 extracted by the Fujisaki model is very similar to that of the original utterances. The RMSE and MAE values do not show any variation for duration, and the variation in pitch is quite small.

In the comparison between the original utterances and those synthesised by a standard HTS system [86], given in Table 7.2, the values here demonstrate a substantial difference from those found in Table 7.1. Correlation for Serbian is significant for both duration and pitch, while that for Sesotho is quite low in both instances, with almost no correlation for pitch. The RMSE and MAE variations have increased for both languages, considerably so for duration. For pitch, the increase in variation is by a small margin for Serbian, while that for Sesotho almost triples.

Table 7.3 compares the two tools and the scores attained are the same as in Table 7.2 for duration. For pitch, the increase in variation values is small in comparison to values obtained in Table 7.2. However, correlation has decreased for both languages, with Sesotho showing a negative correlation.

In general, the results above show that the Fujisaki model allows more accurate F0 modelling than a standard HTS. Although the Fujisaki model performed better for Sesotho, HTS showed a significantly higher performance for Serbian. This is due to the fact that Serbian had more data available for training in HTS, approximately four hours of speech though only 40 minutes of data was used in this experiment, whereas Sesotho data of approximately 40 minutes did not meet the minimum requirements.

## 7.7 Conclusions

In this experiment, the F0 modelling capabilities of the Fujisaki model and HTS for two languages, Sesotho and Serbian[2], were explored and compared. The results obtained show the Fujisaki model to be more accurate than HTS for Sesotho. On the other hand, HTS produced a more natural-sounding speech for Serbian than for Sesotho. The prosody generation accuracy of the HTS system for Sesotho can be improved by training more data, preferably more than 4 hours of speech.

[2]The study for Serbian was sponsored by the Ministry of Education and Science of the Republic of Serbia under the Research grant TR32035.

# Chapter 8

# Summary, Conclusions and Recommendations

## 8.1 Summary and Conclusions

The objective of this thesis was to develop a prosodic model for the Sesotho language based on the Fujisaki analysis. This was approached by first introducing the Sesotho language, a tonal Southern Bantu language, in Chapter 2. The language is spoken nationally in Lesotho and is one of the official languages in South Africa. The syllable, which plays an important role as a carrier of tone, and its properties were described. The surface tonal transcription process, together with the tonal rules involved, were elaborated on.

In Chapter 3, the Fujisaki model, which is the selected prosody modelling tool for this study, was described. A detailed explanation of both the modelling and the analysis processes was given. Other intonation modelling tools, those used in the past and/or in current use, were also highlighted.

Although the Fujisaki model has been shown by other researchers to be effective for tonal languages such as Mandarin, its effectiveness for an African tonal language such as Sesotho has never been considered. The differences and similarities between Sesotho and other tonal languages to which the Fujisaki model has been successfully applied were explored in Chapter 4. Experiments were performed where the polarity of the tone commands in Sesotho was examined. Subsequently, the intonation of sentences and questions and the effect of their prosodic manipulation using the Fujisaki model was also investigated. Results revealed the following:

- High tones are associated with Fujisaki tone commands of positive polarity. Low tones are associated with the absence of a tone command, and not negative polarity as in many other languages.
- Onset and offset of the tone commands can vary considerably in natural Sesotho speech.
- Tonal and vowel differences interact in the perceptual assessment of tones. In several instances, tonal minimal pairs (according to dictionary) actually appear to differ in vowel quality and not tone, but this difference is perceived as a tonal difference.
- Reducing tone command duration and/or amplitude leads to perceived change in meaning in the minimal pairs.

Once the effectiveness of the Fujisaki analysis on the F0 contour for Sesotho was established, the subsequent step was to carry out experiments which would allow a deterministic strategy for prosody modelling for a Sesotho text-to-speech system using the Fujisaki model. A first approach was presented in Chapter 5. First, a speech corpus based on the weather forecast was compiled. The text was prosodically-marked by the process of surface tonal transcription. The relationship between the surface tones and Fujisaki tone commands were investigated. It was established that two or more adjacent syllables of high surface tone formed a prosodic group, a long tone command spanning more than one syllable, in the Fujisaki model. Adjacent syllables of alternating surface tone levels also formed prosodic groups. Cases where the Fujisaki model and the surface tone transcription differed were analysed. Reasons for discrepancies included linguistic phenomena which were not captured in the tonal rules, incorrect dictionary tone, and other factors not observed by the Fujisaki analysis.

Subsequently, the duration and amplitude of Fujisaki tone commands were modified in ways which would allow easy synthesis from the orthography. The modified utterances, whose speech was resynthesised using the FujiParaEditor, were subjected to perceptual testing to establish their naturalness. The evaluation indicated that the best results are obtained when simply aligning the tone commands with the syllable boundaries, and when using conservative tone command amplitudes. The resultant resynthesised speech was judged to be midway between "mother tongue" and "almost mother tongue" quality by native speakers of Sesotho. Further experiments were performed to address some of the mismatches that were observed between the Fujisaki model and the surface tone transcription. Perceptual testing revealed that these modifications led to additional mild degradation in the perceived quality of the resynthesised speech.

In Chapter 6, a statistical prosodic model for Sesotho was developed. This was first approached by investigating the properties of a syllable and their effect on the tone command parameters, the onset T1, the offset T2, and the amplitude At. Three regression tree models were trained to predict each of these parameters from tone-marked Sesotho text. The three models were interdependent because the prediction of T2 was dependent on the predicted T1, and the prediction of At was dependent on the predicted values of T1 and T2. The resynthesised stimuli from the regression trees were compared with those from the baseline (Fujisaki-extracted utterances) and those from the heuristic approach described in Chapter 5. Perceptual evaluation indicated that, while the baseline was the most preferred, more than 80% of the respondents preferred resynthesised speech from the regression trees over that from the heuristic model. An overall conclusion is that data-driven prediction of the Fujisaki model parameters from a tone-marked Sesotho text is viable.

Finally, the prosody generation capability of the Fujisaki model and that commonly integrated into HMM-based speech synthesis (HTS) were investigated in Chapter 7 for two tonal languages of different families: Sesotho and Serbian. Comparative tests showed that the Fujisaki model has a better F0 modelling capability for Sesotho whereas HTS showed a significantly higher performance for Serbian.

## 8.2 Recommendations for Further Work

In order to extend this study, the following points are recommended:

- The accurate tone transcription of words not found in Sesotho tone-marked dictionary by a Sesotho linguist remains part of an ongoing work, and this needs to be explored further.
- Measures in terms of algorithms to take into consideration linguistic phenomena such as tone sandhi, downstep, peak delay, and anticipation must be put into place.

- In addition to the prediction of the Fujisaki tone commands, the phrase command parameters, the position T0 and magnitude Ap, must also be predicted from a tone-marked orthography as a means to augment the current prosodic model.
- The size of the Sesotho annotated speech corpus must be increased. Experience with the Serbian TTS indicates that approximately 4 hours are needed.
- Full end-to-end TTS experiments still need to be performed. Our research has addressed resynthesis from modelled Fujisaki parameters, and TTS given known Fujisaki parameters. An evaluation still needs to be performed in which these are modelled jointly.
- Other prosodic factors, such as emphasis and rhythm, must still be considered.
- Extend the current study to other Bantu languages.

# Bibliography

[1] Language Resource Management Agency. Lwazi Sesotho TTS Corpus. http://rma.nwu.ac.za/index.php/lwazi-english-tts-corpus-989.html [Last accessed September 12, 2014].

[2] Aguero, P.D., Wimmer, K. and Bonafonte, A. "Automatic analysis and synthesis of Fujisaki's intonation model for TTS." In: Proceedings of Speech Prosody. Nara, Japan, 2004.

[3] Machine Learning Group at the University of Waikato. Weka 3: Data Mining Software in Java. http://http://www.cs.waikato.ac.nz/ml/weka/ [Last accessed November 3, 2014].

[4] Auberge, V. "Prosody modeling with a dynamic lexicon of intonative forms: Application for text-to-speech synthesis." In: Proceedings of ESCA Workshop on Prosody. pp. 62-65. 1993.

[5] Bailly, G. and Bartroli, A. "Generating Spanish intonation with a trainable prosodic model." In: Proceedings of Speech Prosody. 2008.

[6] Bailly, G. and Gorisch, I. "Generating German intonation with a trainable prosodic model." In: Proceedings of Interspeech. 2006.

[7] Bailly, G. and Holm, B. "SFC: A trainable prosodic model." Speech Communication 46 (2005): 364-384.

[8] Battiste, M. "Language, and Culture in Modern Society." Reclaiming Indigenous Voice and Vision 192 (2000).

[9] Benesty, J., Sondhi, M.M. and Huang, Y. "Prosodic Processing: A Survey of Current Approaches." In: Springer Handbook of Speech Processing, pp. 479-483. Springer, 2008.

[10] Berment, V. "Méthodes pour informatiser des langues et des groupes de langues peu dotées." PhD Thesis. J. Fourier University - Grenoble, 2004.

[11] Black, A. and Hunt, A. "Generating F0 contours from the ToBI lables using linear regression." In: Proceedings of the 4th International Conference on Spoken Language

Processing (ICSLP). pp. 1385-1388. 1996.

[12] Boersma, P. "Praat, a system for doing phonetics by computer." Glot International 5, no. 9/10 (2001): 341-345.

[13] Breiman, L., Friedman, J.H, Olshen, R.A. and Stone, C.J. Classification and Regression Trees. Wadsworths Brooks, Monterey, 1984.

[14] Brown, M.K. et al. "Web-based platform for interactive voice response (IVR)." *U.S. Patent No. 6,587,822*. Washington, DC: U.S. Patent and Trademark Office. 2003.

[15] Calteaux, K. et al. Lwazi II Final Report: Increasing the impact of speech technologies in South Africa. Pretoria: CSIR Meraka Institute, 2013.

[16] Campbell, N. and Venditti, J. "J-ToBI: An intonation labelling system for Japanese." In: Proceedings of Autumn Meeting Acoustics Society of Japan. pp. 317-318. 1995.

[17] Chen, G-P. et al. "A superposed prosodic model for Chinese text-to-speech synthesis." In: IEEE International Symposium on Chinese Spoken Language Processing. 2004.

[18] Cohen, A. and 't Hart, J. "On the anatomy of intonation." Lingua 19 (1967): 177-192.

[19] Demuth, K. "Issues in the acquisition of the Sesotho tonal system." Journal of Child Language 20 (1993): 275-301.

[20] Demuth, K. and African Studies Centre. "Unifying organizational principles in the development of orthographic conventions." Workshop on Orthography, Maputo, Mozambique, 1988.

[21] Doke, C.M. and Mofokeng, M. Textbook of Southern Sotho Grammar. Cape Town: Longmans, Green and Co., 1957.

[22] Du Plessis, J.A., Gildenhuys, J.G. and Moiloa, J.J. Tweetalige Woordeboek Afrikaans-Suid-Sotho /Bukantswe ya maleme-pedi Sesotho-Seafrikanse. Kaapstad: Via Afrika Bpk, 1974.

[23] Dung, T.N. et al. "Fujisaki model based F0 contours in Vietnamese TTS." In: Proceedings of International Conference on Speech and Language Processing (ICSLP). Jeju, South Korea, 2004.

[24] Dusterhoff, K. and Black, A.W. "Generating F0 contours for speech synthesis using the Tilt intonation theory." In: Intonation: Theory, Models and Applications. 1997.

[25] Ekpenyong, M.E. "Statistical Parametric Speech Synthesis for Ibibio." Speech Communication 56 (2014): 243-251.

[26] Ekpenyong, M.E. and Udoh, E-O. "Intelligent prosody modelling: A framework for tone language synthesis." In: Proceedings of the 6th Langauge and Technology Conference (LTC'13). Poznan, Poland, 2013.

[27] Max Planck Institute for Evolutionary Anthropology - Department of Linguistics. Leipzig Glossing Rules. http://www.eva.mpg.de/lingua/resources/glossing-rules.php [Last accessed September 2, 2014].

[28] Foster, K.I. and Foster, J.C. "DMDX: A Windows display program with millisecond accuracy." Behaviour Research Methods, Instruments, and Computers: A Journal of the Psychonomic Society, Inc. 35, no. 1 (2003): 116-124.

[29] Fujisaki, H. "Analysis and modeling of fundamental frequency contours of Korean utterances – A preliminary study." In: Proceedings of Phonetics and Linguistics. pp. 640-657. 1996.

[30] Fujisaki, H. and Hirose, K. "Analysis of voice fundamental frequency contours for declarative sentences of Japanese." Journal of the Acoustic Society of Japan (E) 5, no. 4 (1984): 233-241.

[31] Fujisaki, H. and Nagashima, S. "A model for the synthesis of pitch contours of connected speech." Annual Report of Engineering Research Institute 28 (1969): 53-60.

[32] Fujisaki, H. and Ohno, S. "Analysis and modeling of fundamental frequency contours of English utterances." In: Proceedings of Eurospeech. pp. 985-988. 1995.

[33] Fujisaki, H., Halle, P. and Lei, H. "Application of F0 contour command-response model to Chinese tones." Reports of Autumn Meeting, Acoustics Society of Japan 1 (1987): 197-198.

[34] Fujisaki, H., Hirose, K., Halle, P. and Lei, H. "Analysis and modeling of tonal features in polysyllabic words and sentences of the Standard Chinese." In: Proceedings of the International Conference on Spoken Language Processing (ICSLP). Kobe, Japan. pp 841-844. 1990.

[35] Fujisaki, H., Ljungqvist, M. and Murata, H. "Analysis and modeling of word accent and sentence intonation in Swedish." In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing. pp. 211-214. 1993.

[36] Fujisaki, H., Ohno, S. and Wang, C. "A command-response model for F0 contour generation in multilingual speech synthesis." In: Proceedings of the 3rd ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis. Australia, 1998.

[37] Fujisaki, H., Ohno, S. and Yagi, T. "Analysis and modeling of fundamental frequency contours of Greek utterances." In: Proceedings of Eurospeech. pp. 465-468. 1997.

[38] Fujisaki, H., Ohno, S., Nakamura, K., Guirao, M. and Gurlekian, J. "Analysis of accent

and intonation in Spanish based on a quantitative model." In: Proceedings of the International Conference on Spoken Language Processing (ICSLP). pp. 355-358. 1994.

[39] Fujisaki, H., Ohno, S., Yagi, T. and Ono, T. "Analysis and interpretation of fundamental frequency contours of British English in terms of a command-response model." In: Proceedings of the International Conference on Spoken Language Processing (ICSLP). 1998.

[40] Fujisaki, H., Wang, C., Ohno, S. and Gu., W. "Analysis and synthesis of F0 contours of Standard Chinese using the command-response model." Speech Communication 47 (2005): 59-70.

[41] Giroux, H.A. "Literacy and the pedagogy of voice and political empowerment." Educational theory 38, no. 1 (1988): 61-75.

[42] Goldsmith, J.A. Autosegmental and Metrical Phonology. Oxford: Blackwell, 1990.

[43] Gođevac, S. "Translating Serbo-Croatian Intonation" In: Prosodic Typology: The Phonology of Intonation and Phrasing. pp. 146-171. Edited by S.-A. Jun. Oxford Linguistics, UK, 2005.

[44] Guma, S.M. An Outline Structure of Southern Sotho. Shuter and Shooter, 1971.

[45] Hirst, D. "A Praat plugin for MOMEL and INTSINT with improved algorithms for modelling and coding intonation." In: Proceedings of International Congress of Phonetic Sciences (ICPhS XVI). pp. 1233-1236. 2007.

[46] Holm, B. and Bailly, G. "Generating prosody by superposing multi-parametric overlapping contours." In: Proceedings of the International Conference on Spoken Language Processing (ICSLP). Beijing, China. pp. 203-206. 2000.

[47] Hyman, L.M. Universals of tone rules: 30 years later. Vol. 1, in Typological Studies in Word and Sentence Prosody, edited by T. Riad and C. Gussenhoven. 2007.

[48] CSIR Meraka Institute. Human Language Technologies – Projects. http://www.csir.co.za/meraka/hlt/hlt-projects.html [Last accessed October 15, 2013].

[49] Jacottet, E. A Practical Method to Learn Sesuto. Morija: Sesuto Book Depot, 1914.

[50] Jilka, M., Mohler, G. and Dogli, G. "Rules for the generation of ToBi-based American English intonation." Speech Communication 28 (1999): 83-108.

[51] Khoali, B.T. "A Sesotho Tonal Grammar." PhD Thesis. University of Illinois at Urbana-Champaign, 1991.

[52] Kisseberth, C. and Odden, D. "Tone: The Bantu Languages." Edited by D. Nurse and G. Philippson. London: Routledge, 2003. 59-70.

[53] Kochanski, G.P. and Shih, C.. "STEM-ML: Language independent prosody description." In: Proceedings of the International Conference on Spoken Language Processing (ICSLP). pp. 239-242. 2000.

[54] Kochanski, G.P. and Shih, C. "Prosody modeling with soft templates." Speech Communication 39, no. 3-4 (2003): 311-352.

[55] Kock, L.J. and Moeketsi, R.H. An Introduction to Sesotho Phonetics. Marius Lubbe Publishers, 1990.

[56] Kohler, K.J. "Studies in German intonation." Arbeitsberichte des institus für Phonetik und digitale Sprachverarbeitung 25 (1991): 295-360.

[57] Kohler, K.J. "The Kiel Intonation Model (KIM), its implementation in TTS Synthesis and its Application to the Study of Spontaneous Speech." 1995. http://www.ipds.uni-kiel.de/kjk/forschung/kim.en.html [Last accessed November 12, 2014].

[58] Kohler, K.J. "Parametric control of prosodic variables by symbolic input in TTS synthesis." Edited by J. van Santen , R. Sproat, J. Olive and J. Hirschberg. In: Proceedings of Progress in Speech Synthesis. pp. 459-475. 1997.

[59] Krauwer, S. "The basic language resource kit (BLARK) as the first milestone for the language resource roadmap." In: Proceedings of the 2003 International Workshop Speech and Computer (SPECOM). Moscow, Russia, 2003.

[60] Kriel, T.J. and van Wyk, E.B. Pukuntsu Woordeboek Noord Sotho-Afrikaans. 4th. Pretoria: Van Schaik, 1989.

[61] Kunene, D.P. "The Sound System of Southern Sotho." PhD Thesis. University of Cape Town, 1961.

[62] Kunene, D.P. "A preliminary study of downstepping in Southern Sotho." African Studies 24 (1972): 233-262.

[63] Ladd, D.R. Intonational Phonology. Cambridge: Cambridge University Press, 1996.

[64] Lee, T., et al. "Modeling tones in continuous Cantonese speech." In: Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP). pp. 2401-2404. 2002.

[65] Li, Y., Lee, T. and Qian, Y. "Analysis and modeling of F0 contours for Cantonese text-to-speech." ACM Transactions on Asian Language Information Processing 3, no. 3 (2004): 169-180.

[66] Lombard, D.P. "Aspekte van toon in Noord-Sotho." PhD Thesis. University of South Africa, 1976.

[67] Louw, J.A., Davel, M. and Barnard, E. "A general-purpose isiZulu speech synthesizer." South African Journal of African Languages 25 (2005): 92-100.

[68] Mabille, A. and Dieterlen, H. Southern Sotho-English Dictionary. Morija: Sesuto Book Depot, 1959.

[69] Marantz, A. "No escape from syntax: Don't try morphological analysis in the privacy of your own lexicon." University of Pennsylvania Working Papers in Linguistics 4, no. 2 (1997).

[70] Martine, G. et al. "Consistency in transcription and labelling of German intonation with GtoBI." In: Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP). Philadelphia, PA, USA. pp. 1716-1719. 1996.

[71] Marx, M.T. *et al*. "System and method for developing interactive speech applications." *U.S. Patent No. 6,173,266*. Washington, DC: U.S. Patent and Trademark Office. 2001.

[72] Mixdorff, H., Fujisaki, H., Chen, G. and Hu, Y. "Towards the automatic extraction of Fujisaki model parameters for Mandarin." In: Proceedings of Eurospeech/Interspeech. Geneva, Switzerland. pp. 873-876. 2003.

[73] Mixdorff, H. "Intonation Patterns of German - Model-based Quantitative Analysis and Synthesis of F0 Contours." PhD Thesis. TU Dresden, 1998.

[74] Mixdorff, H. "A novel approach to the fully automatic extraction of Fujisaki model parameters." In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Istanbul, Turkey. pp. 1281-1284. 2000.

[75] Mixdorff, H. An Integrated Approach to Modeling German Prosody. Eckhard Richter & Co. OHG, 2002.

[76] Mixdorff, H. "Quantitative tone and intonation modeling across languages." In: Proceedings of the International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages. Beijing, China. pp. 137-142. 2004.

[77] Mixdorff, H. FujiParaEditor. October 2009. http://public.bht-berlin.de/~mixdorff/thesis/fujisaki.html [Last accessed February 11, 2014].

[78] Mixdorff, H. and Barbosa, P.A. "Alignment of intonational events in German and Brazilian Portugese - A quantitative study." In: Proceedings of Speech Prosody. Shanghai, China, 2012.

[79] Mixdorff, H. and Mehnert, D. "Exploring the naturalness of several German high- quality text-to-speech systems." In: Proceedings of Eurospeech. Budapest, Hungary, 1999.

[80] Mixdorff, H., Luksaneeyanawin, S., Fujisaki, H. and Charnvivit, P. "Perception of tone

and vowel quality in Thai." In: Proceedings of the International Conference of Spoken Language Processing (ICSLP). Denver, Colorado, USA, 2002.

[81] Mixdorff, H., Mohasi, L., Machobane, M. and Niesler, T. "A study on the perception of tone and intonation in Sesotho." In: Proceedings of Interspeech. Florence, Italy. pp. 3181-3184. 2011.

[82] Mohasi, L., Mixdorff, H., and Niesler, T. "An acoustic analysis of tone in Sesotho." In: Proceedings of the International Congress on Phonetic Sciences (ICPhS XVII). Hong Kong, China. pp. 17-21. 2011.

[83] Myers, S. "Surface underspecification of tone in Chichewa." Phonology 15 (1998): 367-391.

[84] Myers, S. "Tone association and F0 timing in Chichewa." Studies in African Linguistics 28, no. 2 (1999): 215-239.

[85] Navas, E., Hernaez, I., and Sanchez, J. M. "Basque intonation modelling for text to Speech conversion." In: Proceedings of the International Conference of Spoken Language Processing (ICSLP). Denver, USA. pp. 2409-2412. 2002.

[86] Pakoci, E. and Mak, R. "HMM-based speech synthesis for the Serbian language." In: Proceedings of ETRAN. Zlatibor, Serbia, 2012.

[87] Paroz, R.A. Elements of Southern Sotho. Morija: Sesuto Book Depot, 1946.

[88] Petten, C.V. "Influences of semantic and syntactic context on open- and closed-class words." Memory and Cognition 19 (1991): 95-112.

[89] Pierrehumbert, J. "Synthesizing intonation." The Journal of the Acoustics Society of America 70, no. 4 (1981): 985-995.

[90] Raborife, M. "Tone Labelling Algorithm for Sesotho." Unpublished MSc Thesis. University of Witwatersrand, 2011.

[91] Raborife, M., Zerbian, S. and Sigrid, E. "Developing a corpus to verify the performance of a tone labelling algorithm." In: Proceedings of the 22nd Annual Symposium of the Pattern Recognition Association of South Africa (PRASA). pp. 126-131. 2011.

[92] Reyelt, M., Grice, M., Benzmuller, R., Mayer, J. and Batliner, A. "Prosodische Etikettierung des Deutschen mit ToBI." In: Proceedings of Natural Language and Speech Technology. pp. 144-155. 1996.

[93] Schadeberg, T. "Tone in South African Bantu Dictionaries." Journal of African Languages and Linguistics 3 (1981): 175-180.

[94] Sečujski, M., Jakovljević, N. and Pekar, D. "Automatic prosody generation for Serbo-Croatian speech synthesis based on regression trees." In: Proceedings of Interspeech.

pp. 3157-3160. 2011.

[95] Sečujski, M. and Delić, V. "A Software Tool for Semi-Automatic Part-of-Speech Tagging and Sentence Accentuation in Serbian Language." In: Proceedings of IS-LTC. 2006.

[96] Sečujski i, M., Obradović, R., Pekar, D., Jovanov, Lj., and Delić, V. "AlfaNum System for Speech Synthesis in Serbian Language." In: Proceedings of 5th Conference on Text, Speech and Dialogue. pp. 8-16. 2002.

[97] SouthAfrica.info. 2014. http://www.southafrica.info/about/people/language.htm [Last accessed November 14, 2014].

[98] 't Hart, J., Collier, R. and Cohen, A. A Perceptual Study of Intonation: An Experimental-Phonetic Approach to Speech Melody. Cambridge: Cambridge University Press, 1990.

[99] Taylor, P.A. "The Rise/Fall/Connection model of intonation." Speech Communication 15 (1995): 169-186.

[100] Taylor, P.A. Text-to-Speech Synthesis. University of Cambridge, 2007.

[101] Taylor, P.A. and Black, A.W. "Synthesizing conversational intonation from a linguistically rich input." 1994.

[102] Teixeira, J.P., Freitas, D. and Fujisaki, H. "Prediction of accent commands for the Fujisaki intonation model." In: Proceedings of the European Conference on Speech Communication and Technology. Nara, Japan. pp. 451-455. 2004.

[103] Terken, J. "Synthesizing natural-sounding intonation for Dutch: rules and perceptual evaluation." Computer Speech Language 7 (1993): 27-48.

[104] Tokuda, K. et al. "Speech parameter generation algorithms for HMM-based speech synthesis." In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). 2000.

[105] Tokuda, K. et al. "Speech synthesis based on hidden Markov models." In: Proceedings of the IEEE. 2013. 1234-1252.

[106] Tokuda, K., Zen, H. and Black, A.W. "An HMM-based speech synthesis system applied to English." In: IEEE Workshop on Speech Synthesis. 2002.

[107] Traber, C. "F0 generation with a database of natural F0 patterns and with a neural network." Edited by C. Benoit and T. Sawalli G. Bailly. In: Talking Machines: Theories, Models, and Designs. pp. 287-304. 1992.

[108] Traber, C. "Syntactic processing and prosody control in the SVOX TTS system for German." In: Proceedings of Eurospeech. pp. 2099-2102. 1993.

[109] van Hemert, J., Adriaens-Porzig, U. and Adriaens, L. "Speech synthesis in the SPICOS

project." In: Proceedings of Eurospeech. pp. 34-39. 1987.

[110] van Niekerk, D. R. and Barnard, E. "Tone realisation in a Yorùbá speech recognition corpus." In: Spoken Language Technologies for Under-Resourced Languages. 2012.

[111] van Niekerk, D. R. and Barnard, E. "Predicting utterance pitch targets in Yorubá for tone realisation in speech synthesis." Speech Communication, 56, 229-242. 2014.

[112] van Santen, J. and Mobius, B. "A quantitative model of F0 generation and alignment." In: Proceedings of Intonation: Analysis, Modeling and Technology. pp. 269-288. 1999.

[113] van Santen, J., Mobius, B., Venditti, J. and Shih, C. "Description of the Bell Labs intonation system." In: Proceedings of the 3rd ESCA Speech Synthesis Workshop. Jenolan Caves. pp. 293-298. 1998.

[114] Wang, R., Liu, Q. and Tang, D. "A new Chinese text-to-speech system with high naturalness." Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP). pp. 1441-1444. 1996.

[115] Willems, N., Collier, R. and 't Hart, J. "A synthesis scheme for British English intonation." Journal of Acoustics Society of America 84 (1988): 1250-1261.

[116] Yip, M. Tone. New York: Cambridge University Press, 2002.

[117] Zen, H. et al. "The HMM-based speech synthesis system (HTS) version 2.0." In: Proceedings of SSW. 2007.

[118] Zerbian, S. "Phonological phrasing in Northern Sotho." The Linguistic Review 24 (2007): 233-262.

[119] Zerbian, S. "The relative clause and its tones in Tswana." Papers from the Workshop on Bantu Relative Clauses. Vol. 53. Edited by L. Downing, A. Rialland, J. Beltzung, S. Manus, C. Patin, and K. Riedel. ZAS Papers in Linguistics, 2010.

[120] Zerbian, S. and Barnard, E. "Influences on tone in Sepedi, a Southern Bantu language." In: Proceedings of Interspeech. 2008.

[121] Zerbian, S. and Barnard, E. "Phonetics of intonation in South African Bantu languages." Southern African Linguistics and Applied Language Studies 26, no. 2 (2008): 235-254.

[122] Zerbian, S. and Barnard, E. "Realisation of a single high tone in Northern Sotho." Southern African Linguistics and Applied Language Studies 27, no. 4 (2009): 357-380.

[123] Zerbian, S. and Barnard, E. "Realisation of two adjacent high tones: Acoustic evidence from Northern Sotho." Southern African Linguistics and Applied Language Studies

28, no. 2 (2010): 101-121.

[124] Zerbian, S. and Barnard, E. "Word-level prosody in Sotho-Tswana." In: Proceedings of Speech Prosody. Illinois, Chicago, USA, 2010.

# Appendix A

# Minimal Pairs in Sesotho

**Table A.1:** Examples of Sesotho minimal pairs used in Chapter 4. The underlined words are the critical words which give a different meaning based on the tone of the syllables.

| Sentence | English meaning |
|---|---|
| O ile a bolla thabeng. | He was circumcised in the mountains. |
| | He (his body) decayed in the mountains. |
| Oa tena. | He/She is putting on clothes. |
| | He/She is irritating. |
| Motsoala o rata ho seba. | My cousin likes to gossip. |
| | My cousin likes to do mischief. |
| Re ile ra e ts'ela. | We crossed it (the river). |
| | We poured it. |
| Ke lehata la ngoanana. | It is a girl's skull. |
| | The girl is a liar. |
| O ile a e hlola. | He created it. |
| | He conquered it. |
| Hona ke bobatsi. | The is the nettle. |
| | This is the width. |
| O sehile monoana oa hae ke lehare. | He cut his finger with a razor. |
| | He cut his finger in the middle. |
| O ne a pota ha a cho joalo. | He was coming over when he said that. |
| | He was talking crap when he said that. |
| Ke bone lekhala ka metsing. | I saw an aloe plant in the water. |
| | I saw a crab in the water. |
| O hlopha batho ka nako eohle. | He prepares people all the time. |
| | He torments people all the time. |
| Bona, ba teng. | As for them, they are here. |
| | See, they are here. |
| Ke ile ka lapa. | I patched (it). |
| | I got hungry. |
| Ke hlaba se seholo. | It is a big plateau. |
| | I stab the big one. |

# Appendix B

# Questionnaire Example

**Figure B.1:** An example of the questionnaire which was used for the prosody modification of minimal pairs experiment in Section 4.3

**Perceptual Experiment Questionnaire**

The aim of this questionnaire is to get your feedback by listening to various Sesotho s. This information will be used for research purposes and publication only, and we guarantee that your identity will not be revealed to anyone else.

**Instructions**
You will be required to listen to audio files, with each corresponding to the ones in the questionnaire. Please listen carefully and tick an answer which you consider most appropriate. Also, please be aware of the following:

1) no comparing of answers with anyone
2) no copying is allowed
3) no answers should be left blank

Thank you for your co-operation.

The following minimal pairs will be used for Experiment 1.

| | **English meaning** |
|---|---|
| O ile a bolla thabeng. | He was circumcised in the mountains. |
| | He (his body) decayed on the mountain. |
| Oa tena. | He/she is putting on clothes. |
| | He/she is irritating. |
| Motsoala o rata ho seba. | My cousin likes to gossip. |
| | My cousin likes to do mischief. |
| Re ile ra e ts'ela. | We crossed it (the river). |
| | We poured it. |
| Ke lehata la ngoanana. | It's a girl's skull. |
| | The girl is a liar. |

Full Name: ….....

Gender: ….....

Age: ….....

Signature: ….....  Date: ….....

**Experiment 1 : What is the meaning you perceive in each of the following s?**

| **1: Oa tena** | | **2: Ke lehata la ngoanana.** | | **3: Re ile ra e ts'ela.** | |
|---|---|---|---|---|---|
| a) putting on clothes | x | a) skull | | a) cross | |
| b) irritating | | b) liar | x | b) pour | x |
| **4: Motsoala o rata ho seba.** | | **5: O ile a bolla thabeng.** | | **6: Oa tena.** | |
| a) do mischief | | a) decay | x | a) irritating | x |
| b) gossip | x | b) circumcised | | b) putting on clothes | |
| **7: Ke lehata la ngoanana.** | | **8: Motsoala o rata ho seba.** | | **9: O ile a bolla thabeng.** | |
| a) liar | | a) gossip | | a) circumcised | x |
| b) skull | x | b) do mischief | x | b) decay | |
| **10: Re ile ra e ts'ela.** | | **End of warm-up<br>Start of sequence 5** | | **73: Ke lehata la ngoanana.** | |
| a) pour | | | | a) liar | |
| b) cross | x | | | b) skull | |
| **74: Motsoala o rata ho seba.** | | **75: Oa tena.** | | **76: O ile a bolla thabeng.** | |
| a) gossip | | a) putting on clothes | | a) circumcised | |