ANALYSIS OF SESOTHO TONE USING THE FUJISAKI MODEL

Lehlohonolo Mohasi[1], Hansjörg Mixdorff[2], Thomas Niesler[1], Sabine Zerbian[3]

[1]University of Stellenbosch, South Africa; [2]Beuth University of Applied Sciences Berlin, Germany;
[3]University of Potsdam, Germany
lmohasi@sun.ac.za; mixdorff@beuth-hochschule.de; trn@sun.ac.za; szerbian@uni-potsdam.de

Abstract

In this paper, two approaches that can be used to determine the tonal pattern of sentences in Sesotho are compared: surface tone transcription and the Fujisaki model. The tone commands of the latter technique, which represent high tones, are compared with the high surface tones predicted by the tone rules. The mismatched syllables are investigated in order to account for the discrepancies, and particular attention is given to the influence of the adjacent syllables on the tone of the target syllable. Results reveal that the discrepancies are in many cases due to minor errors in the Fujisaki model such as the effects of microprosody, or are due to inconsistent surface tone prediction. An investigation into the prosodic groups formed by the tone commands found that these sequences are mainly due to two or more adjacent syllables with high tone labels, and sometimes due to alternating tone labels between the adjacent syllables.
Index Terms: Fujisaki model, Sesotho tone, Sesotho TTS

1. Introduction

In order for text-to-speech (TTS) 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, on the other hand, is a linguistic property marked by prosodic features such as F0 and intensity. Due to the absence of prosodic marking in the written format, prosodic modeling is a challenge for tonal Bantu languages such as Sesotho [1].

In this paper, we investigate and compare two methods by means of which the tonal pattern of sentences in Sesotho can be determined. The first method is referred to as surface tonal transcription. This method is a tone labelling algorithm based on underlying (lexical) tones, as well as a set of tonal rules, a pronunciation dictionary and morphological analysis. The tonal rules applied are those described in the literature by Khoali [2].

The second method employs the Fujisaki model [3] and is reliant on the acoustics of the uttered speech. The Fujisaki model is a manageable and powerful model for prosody manipulation. It has shown a remarkable effectiveness in modelling the fundamental frequency (F0) contours and its validity has been tested for several languages [4, 5, 6, 7], including tonal languages such as Mandarin [8] and Thai [9]. The Fujisaki model decomposes the F0 contour extracted from the audio samples into three components: a base frequency, a phrase component, which captures slower changes in the F0 contour as associated with intonation phrases, and an accent component that reflects faster changes in F0 associated with high tones.

The accent commands of the Fujisaki analysis, which for tone languages are usually referred to as tone commands, are an indicator of high tones in the utterance. Sesotho has 2 tones – high (H) and low (L) and in previous work [10, 11], it was found that the Fujisaki captures tone commands of positive amplitude for the high tones. For other tonal languages that have been investigated using this technique, such as Mandarin [8], Thai [9], and Vietnamese [12], low tones are captured by tone commands of negative polarity. In contrast, low tones in Sesotho were found to be associated with the absence of tone commands.

The objective of this paper is to investigate the relationship between the surface tone, which is computed using the lexical tones, the morphology and a set of tone rules known from the literature, and the tone commands as determined by the Fujisaki model. We are particularly interested in how closely related the two predictions are, and how they compare, with the perceived tone. The ultimate goal is to develop a technique that is able to predict the tone commands based on the surface tones. This will be an important step in the development of a computational model for tone, which will be essential in a Sesotho text-to-speech system.

Section 2 gives a brief background on tonal transcription and the Fujisaki model. Section 3 discusses the compilation of the corpus, the transcription for the surface and perceived tones, and the decomposition of the Fujisaki model into its parameters. Section 4 gives the results and analysis of the two cases being investigated, while Section 5 draws a conclusion from these results.

2. Background

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, tone rules may modify the underlying tones of the word and thus lead to differing surface tones.

The underlying tone is the tonal pattern of the word in isolation and may be obtained from a tone-marked dictionary. The surface tone is derived from the underlying tone using tone rules, and is the tone given to a word when spoken as part of the sentence. We will indicate underlying high tones by underlining, and surface high tones by acute accents. For example, in the phrase

> Matsatsí á mabéli á látélańg.
> "The next two days."

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

Whereas the surface tone is determined using a set of tone rules, the Fujisaki model 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. This method was first proposed by Fujisaki and

his co-workers in the 70s and 80s [13] 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. A chief attraction of the Fujisaki model lies in its ability to offer a 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' [14].

The Fujisaki model has been integrated into a German TTS system and proved to produce high naturalness compared with other approaches [15]. The inverse model, automated by Mixdorff [3], 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 [16]. Therefore, there is always a trade-off between minimizing the approximation error and obtaining a set of linguistically meaningful commands.

3. Data preparation

Our experiments required significant data collection and preparation, and this is outlined in the following.

3.1 Corpus compilation and recording

The data used for our corpus is based on a set of weather forecast bulletins obtained from the weather bureau in Lesotho, Lesotho Meteorological Services (LMS). The original data was compiled and broadcast for Lesotho TV. The corpus we use consists of the weather forecasts for three consecutive days. The orthographic transcriptions were compiled using the Lesotho orthographic conventions. The corpus consists of 53 sentences, with an average of 23 words per sentence. The future tense was found to be dominant, appearing in 51 of the 53 sentences.

The original audio data 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 by the first author, who is a female native speaker of Sesotho. Recording was performed in a quiet studio environment using a large membrane SHURE KSM32SL microphone. All recordings were made at a sampling rate of 48kHz.

3.2 Surface tone transcription

One of the requirements for surface tonal transcription is knowledge of the tonal rules. For Sesotho, such rules have been described in the literature, although scholars differ on their specific nature. Work has been carried out by Kunene [17], Doke and Mofokeng [18], and by Khoali [2]. Khoali, whose work on Sesotho tone is the most recent, points out possible deficiencies in previous studies. We chose the rules due to Khoali mainly because these rules are the most detailed and some have already been implemented as algorithms in the analysis of other Sotho-Tswana corpora [19].

For tone modeling in (African) tonal languages, in which tone is not indicated by the orthography, an algorithm that predicts the tonal labels of syllables in a word is a prerequisite [20, 1]. We compiled a pronunciation dictionary for use as our reference for underlying tone labels. A Sesotho pronunciation dictionary, in which tones are not marked [21], was adopted as a starting point. To determine the underlying tone of a specific phone for each work in our pronunciation dictionary, we referred to two tone-marked dictionaries – a Sesotho dictionary by Du Plessis et al. [22], and a Northern Sotho dictionary by Kriel et al. [23].

Once the pronunciation dictionary was complete, the sentences in our corpus were annotated with underlying tones from the dictionary. From this underlying tone transcription, a surface tone transcription was deduced by means of a morphological analysis as well as the tonal rules described in [2]. These include for example, the high tone spreading rule (HTS) which describes the spreading of an underlying high tone to the succeeding syllable, and the right branch delinking rule (RBD), which dissociates the immediate right branch of a multiply linked high tone when there is a high tone immediately after the target. For detailed information on these and other tone rules, the reader is referred to Khoali [2].

3.3 Fujisaki model-based analysis

Once the surface tone transcription was complete, the sentences were annotated at word and syllable levels using Praat TextGrid editor [24]. F0 values were extracted using Praat [24] at a step of 10ms and inspected for errors. The F0 tracks were subsequently decomposed into their Fujisaki components applying an automatic method originally developed for German [25]. The F0 components in question are the base frequency, the 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 high tones. Initial experiments in [10] had shown that the low tones in the critical words of the minimal pairs could be modeled with sufficient accuracy without employing negative tone commands. This means that the tone commands of the Fujiaki model showed positive polarity for high tones and zero polarity for low tones. As a consequence, only high tones were associated with tone commands. Adopting this rationale, automatically calculated parameters were viewed in the FujiParaEditor [26] 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 Perceived tone annotation

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 [26], 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 resynthesised for perceptual verification. This was carried out by two of the authors. The aim was to identify all cases where the modeled F0 contour

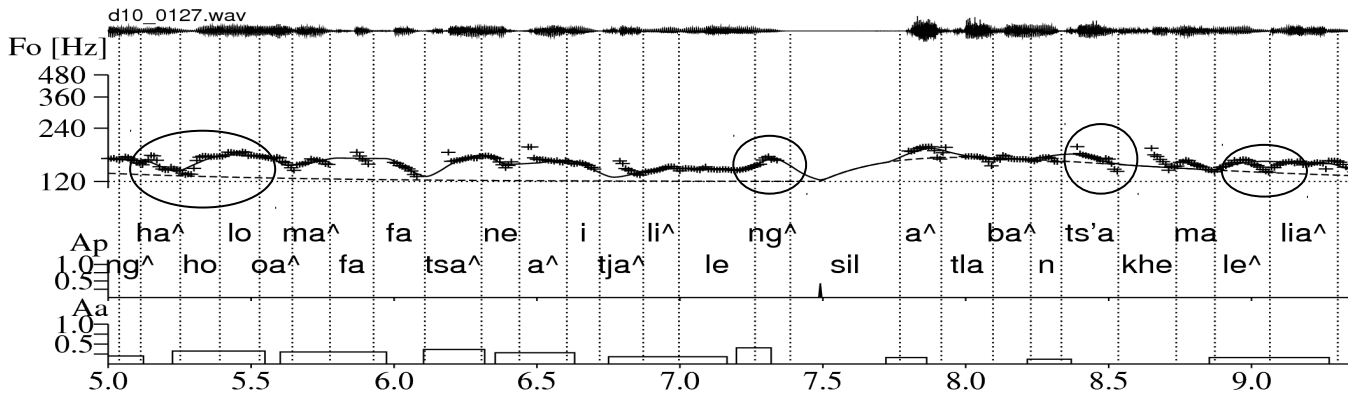

Figure 1: *Sample from sentence d10_0127 illustrating some of the discrepancies between the Fujisaki-modelled F0 contour and the surface tone labels. The predicted high surface tones are indicated by ^, the modeled F0 contour by a solid line, and the extracted F0 contour by the crosses (+). The horizontal dotted line in the top graph indicates the base frequency (Fb), while Ap and Aa show the phrase and tone commands respectively. The part-sentence reads* "... haholo oa mafafatsane a itjalileng, a tla ba nts'a khema la lia(luma)." *(... of scattered showers in particular, with a bit of (thunder).)*

contradicted the predicted surface tone, and the perceived tone. These mismatches were then subjected to more careful analysis.

4. Analysis of experimental results

An overall comparison between the surface, perceived and tone commands are depicted in Tables 1, 2 and 3. Table 1 shows a percentage match between the predicted surface tone and the perceived tone of the 1096 syllables. Table 2 illustrates the percentage match between the surface tone and tone commands derived from the Fujisaki model. Finally, Table 3 presents the percentage match between the 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 does consider the effect of neighbouring syllables is presented in Section 4.1.

Table 1: *Syllable match between the surface tone and the perceived tone.*

	% Perceived tone high	**% Perceived tone low**
Surface tone high	79.8	19.1
Surface tone low	18.2	81.3

Table 2: *Syllable match between surface tone and the tone commands.*

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

Table 1 shows that, overall, there is good agreement between the predicted surface tones and the perceived tones. Table 2 illustrates the correspondence between tone labels derived from the Fujisaki tone commands and the surface tones.

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, while 38% are predicted to have low surface tones, despite the presence of the tone commands.

Table 3: *Syllable match between the perceived tone and the tone commands.*

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

Table 3 shows a similar comparison between the perceived tones and the Fujisaki tone 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 2 and 3 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 Tables 2 and 3, 29% (of the 38%) and 19% (of the 29%) respectively fall into this category. Thus a more discerning analysis would give more optimistic percentages in Tables 2 and 3.

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

4.1 Tonal influence of neighbouring syllables

The tones of the syllables were studied with reference to the influence of their immediate left and right neighbours. We 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 the word or phrase boundary.

We again considered the relationship between the predicted surface tone, the perceived tone, and the output of the Fujisaki model. This was investigated by comparing the tonal patterns of each syllable in the 24 sentences described in Section 3.4, and focusing our attention on the discrepancies that occur among the three 'predictions'.

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

Figure 1 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.

For the utterance in Figure 1, in the word *haholo* (first ellipse), *ha* is perceived as low and F0 shows a low excursion, yet the surface prediction is high. Although *haholo* is underlyingly LLL, the high tone on *ha* has been spread from the syllable *ng^* preceding it. Also, although the F0 contour for *holo* is high, these syllables are perceived as low. The question here is whether this is a pronunciation error, or a surface tone prediction error. We suspect that the mistake originates from the lexical tone pattern, since *haholo* is ascribed two different tonal patterns by the two tonal dictionaries we used as our references – LHL and LLL. From the F0 contour and from the perceptual test, 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.

The second ellipse in Figure 1 highlights the *ng* 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 agreeing answers to this question. The tone should rise due to the continuation rise. In addition, it should because we are dealing with a relative verb. According to [27], relative verbs always have a high tone on their last syllable, *ng*.

The tone command on *ts'a* (third ellipse in Figure 1), 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 1) is the influence of the d-like sound. (In Lesotho orthographic format, the lateral 'l' followed by i or u is pronounced *di* and *du* respectively.)

As mentioned in Section 3.1, the corpus was dominated by future tense 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 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 where 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-toned syllable was surrounded by high tone neighbours. In these instances, the F0 contour stays high. Our explanation for this kind of discrepancy is that since the F0 contour takes time to rise and to fall, it may in such a case simply stay high. The high-tone 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 we assumed the latter pattern. This word appears 7 times in the corpus of 24 sentences and is located at phrase-final position in 6 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 4 instances and HL in 2. In mid-sentence position, the observed Fujisaki 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, we suspect that the continuation rise overrides the tone prediction rule.

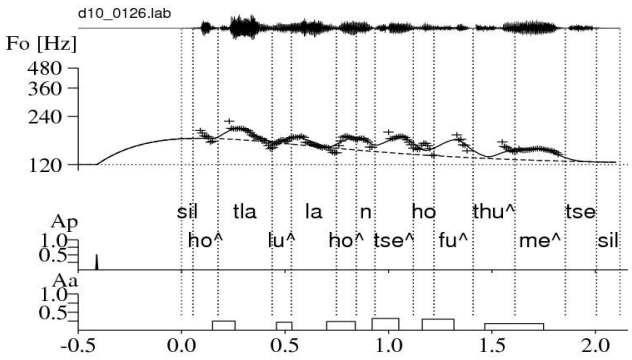

Figure 2: *Sentence d10_0126 showing delays in the maximum of the F0 waveform. The sentence is: "*Ho tla lula ho futhumetse*." (It will remain warm.).*

A further observation is that the tone command pertaining to a high tone can be delayed in some sentences. A high tone corresponds to a period of relatively high F0, and sentence 126 in Figure 2 illustrates such a case. In this figure, *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 [28], it is stated that peak delay is quite common cross-linguistically. In a study for Northern Sotho [29], 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. In his study for Chichewa, another Bantu language, Myers [30] 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 [29], the peak shift inducing object concord has been confirmed for Sesotho (through personal communication), but no detailed data and accounts are available. Since this issue is not the main aim of our investigation, we leave it for future study to confirm or disprove if the F0 peak delay in Sesotho tallies with the findings for Northern Sotho, or for Chichewa.

4.2 Prodosic groupings of high tone syllables

In the Fujisaki model, sequences of consecutive high tones were sometimes observed. In these sequences, the words and

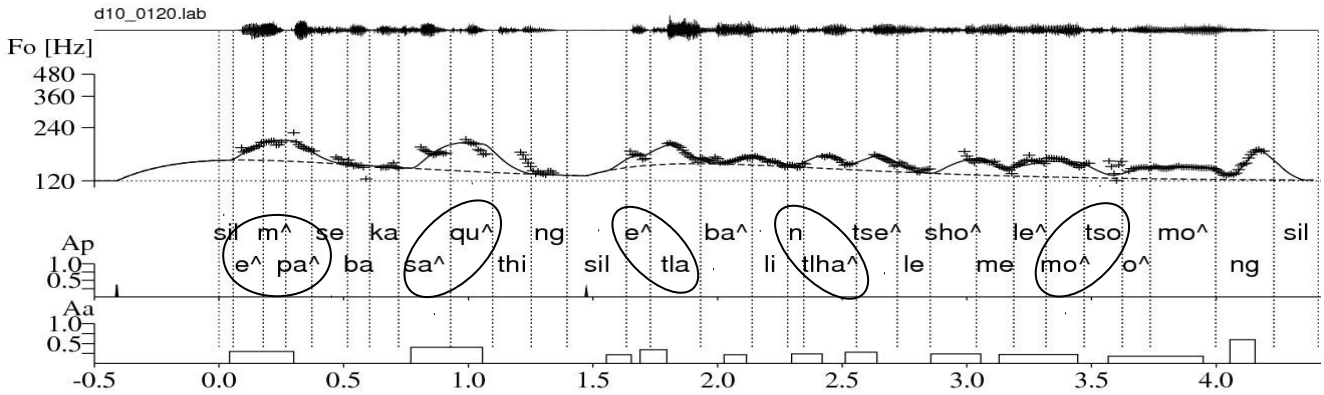

Figure 3: *A sample sentence illustrating prosodic groupings at different levels.*

syllables link up so closely that they appear to be one unit, indicated by long tone commands. We therefore decided to investigate how these prosodic groups are formed, and how the syllables or words within them are related. Our findings for the high tone groupings are given below, and Figure 3 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 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 3) 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 3 – *tso* preceded by *mo*^). Of these low-toned syllables at the start of the sequence, 9 were *ng* and 5 were *n* (fourth ellipse in Figure 3).
- 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, both across word boundaries (***sa qu***thing) and within words (***empa***) – (first and second ellipses in Figure 3). 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 5 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 single-syllable concords. The prosodic effect of two or more adjacent high-tone syllables, the alternating tone labels, and the carrying-over of the high tone holds here too.

As observed in our examples, and as confirmed by [31], the F0 does not drop rapidly within an utterance after it has reached a peak. A gradual decline is observed, rather than a steep drop to a subsequent low tone [31]. Myers [32] shows that the pitch value of a low tone is determined by the tonal environment. Therefore, in the instance 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 [28], the actual pitch value of a low-toned syllable thus depends on the presence and location of preceding high tones. This is confirmed by the examples considered in our analysis above.

5. Discussion and Conclusions

Since neither the surface tone transcriptions nor the perceived tone labels are completely reliable, care must be excercised in the interpretation of the Fujisaki model commands. F0 can be inaccurately estimated, surface tones are sometimes incorrectly specified, and there can be flaws in the Fujisaki model, like the insertion of a phrase command instead of a long tone command of small amplitude. Then there is also the influence of microprosody 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 surface high 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 tone 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 our annotation of perceived tone.

Overall, we have found that 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 microprosody 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 grouping 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.

An overall aim of our ongoing work is to develop a technique that can predict the tone commands of the Fujisaki model from the orthography and surface tones. This would allow the Fujisaki model to be integrated into a Sesotho TTS system for the purpose of prosody modelling. The good correspondance between perceived tone and Fujisaki tone commands presented in the paper, as well as the additional insights into the behaviour of tone in Sesotho obtained from our detailed analysis of discrepancies, give us optimism for the prospects of this approach.

6. Acknowledgements

This work is supported by DFG International collaboration grant Mi 625/16-1 for Mixdorff, Mohasi and Niesler. Mohasi gratefully acknowledges further financial support by Telkom South Africa.

7. References

[1] Zerbian, S. & Barnard, E., "Word-level prosody in Sotho-Tswana", Proceedings of Speech Prosody 2010, 2010.

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

[3] Mixdorff, H., "A novel approach to the fully automatic extracting of Fujisaki model parameters", IEEE Int. Conference on Acoustics, Speech, and Signal Processing Istanbul 3, pp 1281-1284, 2000.

[4] Narusawa, S. et al., "A method for automatic extraction of model parameters from fundamental frequency contours of speech", Proceedings of ICASSP, pp 509 – 512, 2002.

[5] Moberg, M. & Parssnen, K., "Comparing CART and Fujisaki intonation models for synthesis of US-English names", in Speech Prosody, 2004.

[6] Aguero P. D. & Bonafonte, A., "Consistent estimation of Fujisaki intonation model parameters", in SPECHOM, 2005.

[7] Rossi, P. S., Palmieri, F., & Cutugno, F., "Inversion of F0 model for natural-sounding speech synthesis", IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 1, pp 320-523, 2003.

[8] Mixdorff, H., Hu, Y. & Chen, G., "Towards the automatic extraction of Fujisaki model parameters for Mandarin", Proceedings of Eurospeech 2003, 2003.

[9] Mixdorff, H., Luksaneeyanawin, S., Fujisaki, H. & Charnvivit, P., "Perception of tone and vowel quantity in Thai", Proceedings of ICSLP, 2002.

[10] Mohasi, L, Mixdorff, H. & Niesler, T,. "An acoustic analysis of tone in Sesotho", Proceedings of ICPhS XVII, pp17-21, 2011.

[11] Mixdorff, H., Mohasi, L., Machobane, M., & Niesler, T., "A study on the perception of tone and intonation in Sesotho", Proceedings of Interspeech2011, pp 3181-3184, 2011.

[12] Dung, T. N., Mixdorff, H. et al., "Fujisaki model based F0 contours in Vietnamese TTS", Proceedings of ICSLP2004, 2004.

[13] Fujisaki, H. & Hirose, K., "Analysis of voice fundamental frequency contours for declarative sentences of Japanese", *Journal of the Acoustics Society of Japan* (E) 5(4), 233-241, 1984.

[14] Taylor, P. A. "The Rise/Fall/Connection Model of Intonation", *Speech Communication*, vol. 15, pp 169-186, 1995.

[15] Mixdorff, H. & Mehnert, D., "Exploring the Naturalness of Several German High-Quality-Text-to-Speech Systems", *Proceedings of Eurospeech '99*, vol. 4, pp 1859-1862, 1999.

[16] Aguero, P. D., Wimmer, K. & Bonafonte A., "Automatic Analysis and Synthesis of Fujisaki's Intonation Model for TTS", Proceedings of Speech Prosody 2004, 2004.

[17] Kunene, D. P., "The Sound System of Southern Sotho", Unpublished PhD thesis, University of Cape Town, 1961.

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

[19] Raborife, M., "Tone labelling algorithm for Sesotho," Unpublished MSc thesis, University of Witwatersrand, 2011.

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

[21] Resources for the pronunciation dictionary and phoneset by the Lwazi Project at Meraka Institute http://www.meraka.org.za/lwazi/downloads.php

[22] Du Plessis, et al., *Tweetalige Woordeboek Afrikaans-Suid-Sotho*, 1974.

[23] Kriel & Van Wyk. *Pukuntsu Woordeboek Noord Sotho-Afrikaans*, Pretoria: Van Schaik, 4th edition, Pretoria, 1989.

[24] Boersma, P. Praat, "A system for doing phonetics by computer", Glot International 5 (9.10), 341-345, 2001.

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

[26] Mixdorff, H. (1/10/2009). FujiParaEditor, http://public.bht-berlin.de/~mixdorff/thesis/fujisaki.html

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

[28] Zerbian, S. & Barnard, E., "Phonetics of intonation in South African Bantu languages". *Southern African Linguistics and Applied Language Studies* 2008: 26(2): 235-254, 2008.

[29] Zerbian, S. & Barnard, E., "Realisation of a single high tone in Northern Sotho", *Southern African Linguistics and Applied Language Studies* 27(4): 357-380, 2009.

[30] Myers, S., "Tone association and F0 timing in Chichewa". *Studies in African Linguistics* 28(2): 215-239, 1999.

[31] Zerbian, S. & Barnard, E. "Realisation of two adjacent high tones: Acoustic evidence from Northern Sotho", *Southern African Linguistics and Applied Language Studies* 28(2): 101-121, 2010.

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