Perceptual evaluation of the effect of mismatched Fujisaki model commands and surface tone in Sesotho

Lehlohonolo Mohasi [1], *Hansjörg Mixdorff*[2], *Thomas Niesler* [1]

[1] Department of Electrical & Electronic Engineering, University of Stellenbosch, South Africa
[2] Department of Computer Science and Media, Beuth University Berlin, Germany
lmohasi@sun.ac.za, mixdorff@beuth-hochschule.de, trn@sun.ac.za

Abstract

Sesotho is a tonal Southern Bantu language which has so far received extremely little attention by the speech research community. We consider tone modelling for Sesotho using the Fujisaki model-based analysis with a view to the development of a text-to-speech (TTS) system. Fujisaki analysis can be used to indicate the tone associated with a syllable, but it often differs from the surface tone that would be available for TTS synthesis. We investigate instances in which the surface tone differs from the tone indicated by Fujisaki analysis, and determine the effect of these discrepancies on speech quality. The amplitude of Fujisaki tone commands is manipulated to match the surface tones, and the resulting resynthesized speech subsequently analysed by perceptual tests. We find that the effect of inserting tone commands at high surface tone syllables is more severe than matching the Fujisaki tone commands with low surface tone syllables, in terms of naturalness. Furthermore, some discrepancies can be attributed to errors in the surface tonal transcription. However, on average, all manipulations lead only to a mild degradation in speech quality. We conclude that the Fujisaki model is a feasible way to model tone in Sesotho even in the presence of limited and under-developed linguistic resources.

Index Terms: Fujisaki model, Sesotho, surface tonal transcription, text-to-speech synthesis

1. Introduction

Accurate prosodic modelling is crucial for natural-sounding text-to-speech (TTS) systems and can be achieved by correct modelling of pitch in tonal languages. For example, Ekpenyong et al. [1] found that the use of tone marking contributes significantly to the quality of synthetic Ibibio speech. In Sesotho, tone is not marked in the orthography, but must be deduced by the process of surface tonal transcription. This relies on morphological analysis, a tone-marked pronunciation dictionary, and a set of tonal rules. Each one of these three components can introduce errors, for instance, Schadeberg [2] and Roux [3] have pointed out the inconsistency of Sesotho tone-marked dictionaries.

The Fujisaki model is a tractable and powerful tool for prosody manipulation that has proven to be effective for modelling fundamental frequency (F0) contours. Its validity has been tested for several languages [4, 5, 6, 7] some of which are tonal languages such as Thai [8] and Mandarin [9]. It has been shown that the tone of a syllable can be represented in the Fujisaki model as a tone command, which is a pulse indicating where the rising and falling of F0 occurs.

Sesotho uses a register tone system with two tones, high and low. The Fujisaki model and surface tonal transcription have shown a strong agreement on the interpretation of tone in Sesotho sentences [10]. The purpose of this paper is to identify, investigate and perceptually evaluate those cases where the Fujisaki tone commands and surface tone labels do not agree. We want to determine if the mismatch between the actual F0 realisations and the surface tones seriously degrade the prosodic quality.

In modelling Sesotho prosody, certain phenomena must be taken into consideration. Tone sandhi is a phonological change occurring in tonal languages, in which tones occur in combination with other tones. It is present in Sesotho as pointed out by Demuth [11], and is modelled by the surface tonal rules HTD (spreading of a lexical high tone to the immediate right syllable) and IHTS (iterative spreading of a high tone to the end of a verb). Other tonal rules, however, do not model tone sandhi. Ekpenyong and Udoh [12] emphasise the importance of tone sandhi and its effect on the overall F0 contour in tone languages.

The Obligatory Contour Principle (OCP) is a phenomenon where adjacent identical tone elements are prohibited. According to Yip [13], OCP violations can be avoided in a variety of ways, such as tone deletion, blocking of spreading if it leads to adjacency, or fusion between tones. However, OCP in some cases is violable [13]. In Sesotho surface tonal transcription, OCP is observed via the RBD (dissociating 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), LBD (delinking the immediate preceding left branch of a multiply-linked high tone syllable if it is preceded by a high tone syllable) and FR (exempting syllables at the end of a phonological phrase from the application of tonal rules) rules [14]. In contrast, OCP is not fully observed by the Fujisaki analysis and this is evident in the production of prolonged tone commands instead of exclusively single-syllable tone commands. Peak delay is observed when the F0 peak corresponding to a high-toned syllable occurs in the following syllable [15, 13], while anticipation is observed when a high tone is realised on the preceding syllable [16]. The former aspect is modelled in the surface tonal transcription via the HTS1, IHTS and GTI rules. The latter feature is not modelled in surface tone transcription. Both are observed in the Fujisaki analysis.

2. Data preparation

The following sections describe the compilation, preparation, annotation, and selection of the data on which our experiments are based.

2.1. Experimental data

Our data is drawn from a corpus that is based on a set of weather forecast bulletins obtained from the weather bureau in Lesotho. The corpus comprises a total of 256 sentences with an average of 23 words per sentence, and a total of 51 minutes of speech. The original audio data had a poor signal-to-noise ratio, making it unsuitable for 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.

10.21437/SpeechProsody.2014-104

For tone modelling 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 [17]. As a starting point, we compiled a suitable domain dictionary for weather forecasts using two published tone-marked dictionaries – a Sesotho dictionary by Du Plessis et al. [18], and a Northern Sotho dictionary by Kriel and van Wyk [19]. The dictionaries contain lexical tones for each word.

Next, 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 [14].

On completion of the surface tonal transcription, the sentences were annotated at word and syllable levels using Praat [20]. F0 values were extracted at a step of 10ms and inspected for errors. The F0 tracks were subsequently decomposed into their Fujisaki model components applying an automatic method originally developed for German [21]. Initial experiments in [22] had shown that the low tones in the critical words of the minimal pairs could be modelled with good accuracy without employing negative tone commands. Consequently, high tones were associated with Fujisaki tone commands with positive amplitude, while low tones had no associated tone command (i.e. zero amplitude).

2.2. Selection of mismatched syllables

First we selected from our data cases in which a Fujisaki tone command corresponds to one or more low surface tones. In general, the Fujisaki tone command spanned more than one syllable. Table 1 classifies the pattern of surface tones associated with each such tone command. Each low surface tone indicates a discrepancy. With this in mind, we selected cases in which a Fujisaki tone command coincided with one or more low surface tones. In total, 63 such cases were isolated from our data.

We also considered cases in which a high surface tone was not accompanied by a Fujisaki tone command. 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 2. In total, 64 such cases were isolated from our data.

Table 1: *Instances in which Fujisaki tone commands correspond to low surface tones (FHSL).*

Surface tone pattern	Number of syllables	Number of cases
Alternating, e.g. LHLHL	≥ 2	40
Low surface tone labels only, e.g. LLL	≥ 1	7
Any other combination, e.g. LHHL	≥ 2	16
TOTAL		63

Table 2: *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

3. Pitch contour modifications

In the previous section, two types of mismatches between the Fujisaki model and the surface tone labels were identified. In this section, these mismatches are "resolved" by either inserting or removing the Fujisaki tone command in order to match the predicted surface tones.

3.1. Case FHSL: High Fujisaki tone associated with low surface tone

For cases in which the Fujisaki tone command coincided with a syllable with a low surface tone, the amplitude of the tone command was set to zero. Figure 1 illustrates this modification.

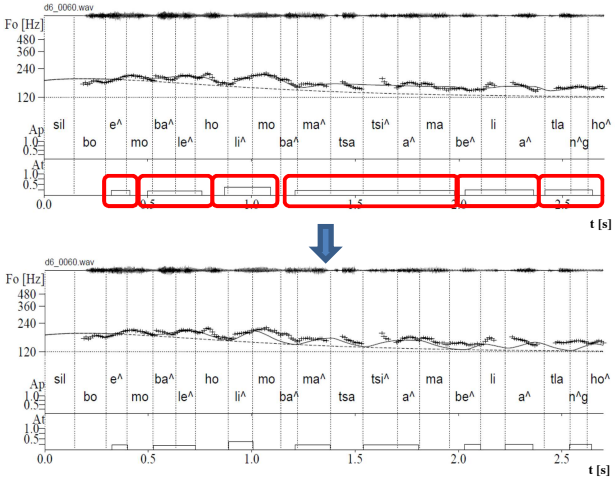

Figure 1: *Modifying Fujisaki tone commands to coincide with low surface tone syllables. The top panel shows an original utterance with surface tone patterns in the following order: HL, LHHL, LHL, HHLHHLH, HLH, and LHH. The bottom panel illustrates a change in pattern after modification. The phrase reads "Boemo ba leholimo ba matsatsi a mabeli a tlang ho …" – "The weather in the next two days …"*

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 t_1 and/or offset t_2 times were unchanged for high surface tone syllables. In all other cases, t_1 and t_2 coincided with syllable boundaries. The optimal alignment of tone command onsets and offsets appears to be language-dependent [8, 23], and its determination for Sesotho remains the subject of ongoing work.

3.2. 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 with average amplitude were inserted and t_1 and t_2 were set to syllable boundaries. Sixty-four modified phrases were generated, in line with Table 2. Figure 2 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.

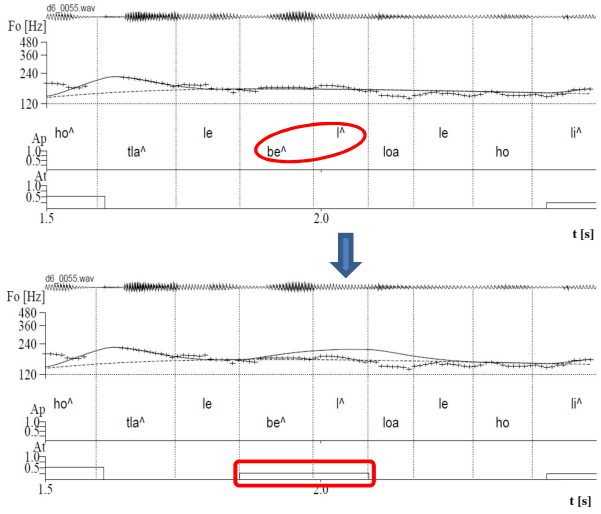

Figure 2: *Inserting a tone command. The top panel indicates a sequence of two high surface tone syllables, be^ and l^, with no Fujisaki tone command. In the bottom panel, a tone command is created for these syllables. The partial phrase reads "ho tla lebelloa leholi…" – "the sky will be expected …"*

4. Perceptual Evaluation

Perceptual evaluation of the modified phrases was carried out by twenty-one Sesotho speakers (7 females, 14 males). All subjects are native Sesotho speakers from Lesotho, and are students at the University of Stellenbosch. The classification of the data and the evaluation process are detailed below.

4.1. Data

For each group in Section 3.1, modifications were applied to the Fujisaki tone commands, after which the utterance was resynthesized using the PSOLA-based Praat ManipulationEditor (20) [19]. Tone commands which were inserted in Section 3.2 were also resynthesized in a similar manner. The resulting phrases were collected for perceptual testing.

Table 3: *Data used for perceptual evaluation.*

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

The data for evaluation included a number of unmodified phrases to serve as a baseline. Table 3 gives a summary of the data for the two modifications and groups.

4.2. The evaluation process

DMDX [24] was used to perform all perceptual evaluations. Subjects listened to the utterances in a quiet room using a headset. Evaluation of the data was based on a rating scale intended to reflect naturalness, as shown in Figure 3. Subjects were asked to rate each individual audio file according to this scale.

An informal SUS intelligibility test [25] of the modified phrases was also performed. Words from the modified and unmodified phrases were randomly selected to form semantically unpredictable sentences (Table 4). The sentences generated were each five words long. These were then resynthesised and the listener requested to indicate what they heard. Due to time constraints, this was performed by only one native Sesotho speaker.

1	Sounds like a mother–tongue speaker
2	
3	Sounds almost like a mother–tongue speaker
4	
5	Definitely not a mother–tongue speaker
6	
7	Disturbingly unnatural speech, hard to understand
8	Don't know

Figure 3: *The rating scale used for perceptual evaluation.*

Table 4: *Data used for the SUS intelligibility test.*

	FHSL	FLSH	Unmodified
Number of words	106	115	108
Number of sentences	21	23	22

5. Results

Phrases rated with an 8 ("Don't know") were excluded from the following analysis. In other cases, average scores according to the scale in Figure 3 have been considered.

Figure 4 illustrates the overall average perceptual scores resulting from the two types of modification described in Section 3. Unmodified phrases are perceived to be most natural, 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 slightly 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, neighbouring "almost mother tongue".

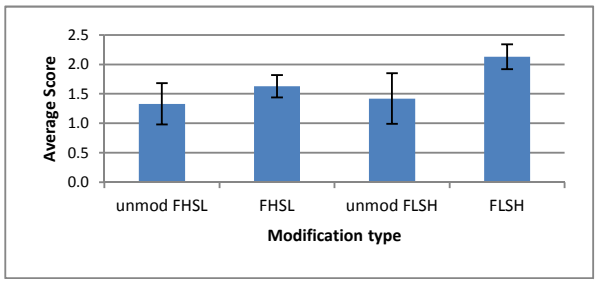

Figure 4: *Overall average perceptual scores for FHSL and FLSH modifications. Vertical bars denote 95% confidence intervals.*

From the FHSL data described in Table 1, we isolated a subset consisting of cases where a Fujisaki tone command was associated with a sequence of between 1 and 3 low surface tone syllables. This data is summarised in Table 5, while Figure 5 shows the corresponding results of the perceptual evaluation. For each case, the Fujisaki tone command amplitude was set to zero 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 low due to the small number of samples (4).

Table 5: *Instances in which Fujisaki tone commands correspond to consecutive low surface tone syllables.*

Consecutive low tone syllables	Number of cases
1	20
2	8
3	4
TOTAL	32

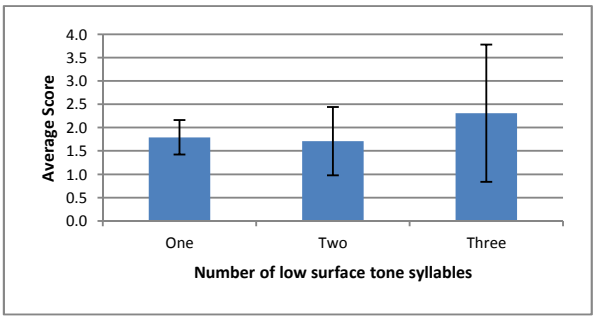

Figure 5: *Average perceptual scores when removing the Fujisaki tone command associated with 1, 2, and 3 consecutive low surface tones. Vertical bars denote 95% confidence intervals.*

Figure 6 shows a similar analysis for the FLSH data described in Table 2. The results are similar across number of syllables.

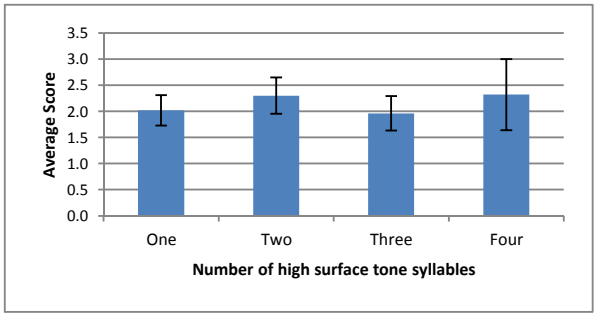

Figure 6: *Average perceptual scores when inserting Fujisaki tone commands for 1, 2, 3, and 4 consecutive high surface tone syllables. Vertical bars denote 95% confidence intervals.*

Table 6 shows the results from the informal intelligibility test described in Section 4.2. The intelligibility score is based on the number of words identified correctly in each sentence. Sentences composed of unmodified words have the highest intelligibility score, with the lowest score obtained for FHSL modifications. Removal of the Fujisaki tone command therefore appears to have a much higher detrimental effect on intelligibility than its insertion.

Table 6: *Intelligibility scores for FHSL, FLSH, and unmodified utterances.*

Modification	Intelligibility Score [%]
FHSL	38.1
FLSH	60.9
Unmodified	77.3

Finally, each of the 127 mismatches described in Tables 1 and 2 was considered individually in order to determine the source of the discrepancy. Tables 7 and 8 describe the results of this investigation. The totals exceed the values in Tables 1 and 2 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 will be explored in future work.

Figure 7 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 than when the tone command was removed.

Table 7: *Mismatches in the FHSL case.*

Description	Number of instances
Tone sandhi is observed by the Fujisaki analysis but not by the relevant surface tonal 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 rule 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 8: *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

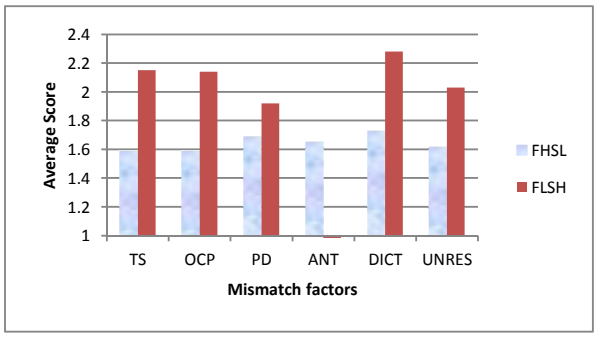

Figure 7: *Sources of mismatch between surface tone and Fujisaki tone commands, and their effect on naturalness. TS = tone sandhi, PD = peak delay, ANT = anticipation, DICT = incorrect tone in dictionary, and UNRES = unresolved cases.*

6. Conclusions

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. In this paper we have performed experiments to determine the effect of these mismatches on the perceived quality or resynthesized speech. We find that, overall, the modifications applied to "rectify" the mismatch lead only to mild degradation in the perceived quality of the speech. From this we conclude 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 error and that much can be gained by the improvement of the surface tone transcription process.

Acknowledgements - 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.

7. References

[1] Ekpenyong, M.E. et al. *Statisitical parametric speech synthesis for Ibibio*. Speech Communication, Vol. 56, pp. 243-251, 2014.

[2] Schadeberg, T. *Tone in South African Bantu languages*. Journal of African Languages and Linguistics, Vol. 3, pp. 175-180, 1981.

[3] Roux, J. On the perception and description of tone in the Sotho and Nguni languages. Kaji Shigeki [ed.]. *Proceedings of the symposium cross-linguistic studies of tonal phenomena. Historical Development, Phonetics of Tone and Descriptive Studies*. Tokyo : Tokyo University of Foreign Studies, ILCAA, 2003.

[4] Narusawa, et al. *A method for automatic extraction of model parameters from fundamental frequency contours of speech*. Orlando, Florida, USA : Proceedings of ICASSP, 2002.

[5] Rossi, P.S., Palmieri, F., and Cutugno, F. *Inversion of F0 model for natural-sounding speech synthesis*. Hong Kong, China : Proceedings of IEEE ICASSP, 2003.

[6] Moberg, M. & Parssnen, K. *Comparing CART and Fujisaki intonation models for synthesis of US-English names*. Nara, Japan : Proceedings of Speech Prosody, 2004.

[7] Aguero, P.D. *Automatic analysis and synthesis of Fujisaki's intonation model for TTS*. Nara, Japan : Proceedings of Speech Prosody, 2004.

[8] Mixdorff, H., Luksaneeyanawin, S., Fujisaki, H. & Charnvivit, P. *Perception of tone and vowel quantity in Thai*. Denver, Colorado, USA : Proceedings of ICSLP, 2002.

[9] Mixdorff, H., Hu, Y. & Chen, G. *Towards the automatic extraction of Fujisaki model parameters for Mandarin*. Geneva, Switzerland : Proceedings of Eurospeech, 2003.

[10] Mohasi, L., Mixdorff, H. & Niesler, T. *Characterisation of prosodic groups in Sesotho using the Fujisaki model*. Journal of Chinese Linguistics Monograph, 2013.

[11] Demuth, K. *Issues in the acquisition of the Sesotho tonal system*. Journal of Child Language, Vol. 20, pp. 275-301, 1993.

[12] Ekpenyong, M.E. and Udoh, E-O. *Intelligent prosody modelling: A framework for tone language synthesis*.

[13] Yip, M. *Tone*. Cambridge University Press, 2002.

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

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

[16] Hyman, L.M. Universals of tone rules: 30 years later. Thomas and Gussenhoven, Carlos Riad [ed.]. *Typological Studies in Word and Sentence Prosody*. Mouton de Gruyter, Vol. 1, pp. 1-34, 2007.

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

[18] Du Plessis, J.A. et al. *Tweetalige Woordeboek Afrikaans-Suid-Sotho*. Kaapstad : Via Afrika Bpk, 1974.

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

[20] Boersma, P *Praat - A system for doing phonetics by computer*. 9/10, Glot International, Vol. 5, pp. 341-345, 2001.

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

[22] Mohasi, L., Mixdorff, H. & Niesler, T. *An acoustic analysis of tone in Sesotho*. Hong Kong, China : Proceedings of ICPhS XVII, 2011.

[23] Mixdorff, H. and Barbosa, P.A *Alignment of intonational events is German and Brazilian Portuguese - a quantitative study*. Shanghai, China : Proceedings of Speech Prosody, 2012.

[24] Foster, K.I. & Foster, J.C. *DMDX: A Windows display program with millisecond accuracy*. Behavior Research Methods, Instruments, and Computers: A Journal of the Psychonomic Society, Inc, Vol. 35 (1), pp. 116-124, 2003.

[25] Benoit, C., Grice, M. and Hazan, V. *The SUS test: A method for the assessment of text-to-speech intelligibility using Semantically Unpredicatable Sentences*. Speech Communication, Vol. 18, pp. 381-392, 1996.