A Study on the Perception of Tone and Intonation in Sesotho

Hansjörg Mixdorff[1], Lehlohonolo Mohasi[2], 'Malillo Machobane[3], Thomas Niesler[2]

[1]Beuth University of Applied Sciences Berlin, Germany
[2]University of Stellenbosch, South Africa
[3]National University of Lesotho

mixdorff@beuth-hochschule.de;mohasi@sun.ac.za;emmachobane@nul.ls;trn@sun.ac.za

Abstract

This paper presents a study on the perception of Sesotho, a Southern African tonal language, employing a set of recorded minimal pairs, whose *F0* contours were analyzed in a previous study using the Fujisaki model and resynthesized. Sequences of prosodically modified stimuli were produced to examine the effect of these modifications on word identification, statement/question distinction, as well as focus identification. With few exceptions, results regarding word identification are in line with our expectations. *F0* modifications even seem to override vowel differences between words when both affect its meaning. With respect to the statement/question distinction, shortening of the penultimate syllable, higher speech rate and increased phrase command magnitude *Ap* all increase the probability of an utterance to be perceived as a question. The focus experiment, however, produced inconclusive results, possibly due to its complex setting.

Index Terms: Sesotho, Fujisaki model, tone perception

1. Introduction

Sesotho is a Southern Bantu language spoken as an official language in Lesotho and South Africa. Sesotho is a tonal language with two contrasting tonemes, high (H) and low (L). The tone of a syllable is carried by the vowel, or by the nasal, if the nasal is syllabic. The interrelationship between the tone and general intonation in Sesotho is hitherto poorly understood and technologically not addressed. This is complicated by the fact that tonal information is not indicated in the orthography of Sesotho [1] [2], as well as most other Bantu languages [3].

In a recent study [4] examining minimal pairs of words embedded in identical carrier sentences, tonal alignments and magnitudes of f0 excursions were analyzed using the Fujisaki model [5]. This model decomposes a given log *F0* contour into a base frequency *Fb*, a phrase component, which captures slower changes in the *F0* contour as associated with intonation phrases, and an accent component that reflects faster changes of *F0* associated with accents and boundary tones. Results of the study showed that high tones in Sesotho are associated with tone commands of positive polarity whereas during 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 syllables. Some only showed vowel differences, but 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 current perceptual study we aim to investigate to what extent tonal perception interacts with vowel perception to facilitate word identification. Furthermore, we will examine which prosodic features facilitate the perceptual differentiation between statements and questions. In our previous study we observed that utterances of questions were generally spoken at a higher pitch, that is, increased phrase command magnitude *Ap*, slightly increased speech rate, as well a conspicuous shortening of the penultimate syllable which is the carrier of word stress in Sesotho. This observation was also reported in [6].

Finally, we are interested in exploring whether the perceived focus and hence the meaning of an utterance can be manipulated by raising *F0* on selected words. It has been claimed that other languages in the same family do not employ *F0* for marking focus [7]. In contrast, Asian tone languages such as Mandarin are known to use F0 range for marking focus. The second and third authors of this study are native speakers of Sesotho. During informal resynthesis tests with the Sesotho material they observed that by increasing the tone command amplitude *At* on certain words we yielded natural sounding utterances with slightly changed meanings. In the current study we therefore explore whether these meanings can be identified systematically.

2. Stimulus design

All stimuli were produced using the resynthesis capability of the *FujiParaEditor* [8] which replaces the original *F0* contour of an utterance by one generated with the Fujisaki model. The actual acoustic resynthesis employs the *Praat ManipulationEditor* [9] which is based on PSOLA. Original utterances had been produced by three male native speakers of Sesotho from Lesotho.

2.1. Lexical Identification

Table 1 lists all minimal pairs of words that were used in the lexical identification task. As can be seen, '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. We were interested in the following research questions:

(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 partner - 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 partner – even without modifying the vowel quality?

To this effect we introduced the manipulations listed in Table 1. Manipulated utterances are indicated by white background and the nature of the manipulation is described. The intended target meanings are marked by grey background and it is listed what way the source has been manipulated to obtain the target. Note that 'bolla' as well as 'ts'ela' exhibit vowel differences. Vowel quality was not modified in this study.

Table 1. *List of manipulations used in the word identification task. The alternative meanings we aimed to produce are indicated by a grey background. 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
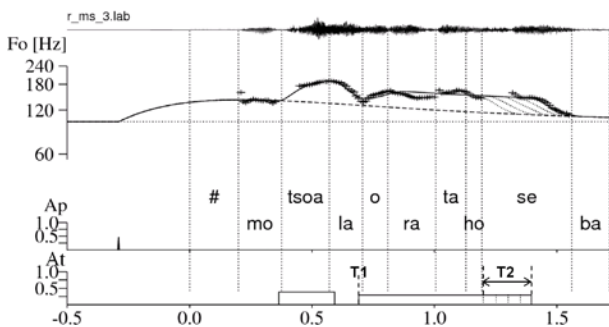
Figure 1: *Illustration of stimulus variation. The high tone on "seba" in the original is slowly transformed into a low tone by reducing the tone command offset time T2. The tone command onset time T1 is left constant in this example.*

Figure 1 shows an example of how the stimuli were created. The figure displays, from the top to the bottom: the speech waveform, the extracted (+) and modeled (-) *F0* contour, and the underlying phrase (magnitude *Ap*) and tone (amplitude *At*) commands. The syllable boundaries are indicated by the dotted vertical lines. The utterance "Motsoala o rata ho seba." means *"My cousin likes to gossip.",* with "seba" exhibiting an HL pattern. The parameter modified in this example is the offset time T2 of the second tone command. As indicated in the figure, *T2* is 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 we expect to observe as we reduced T2.

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

2.2. Question vs. Statement

For this task we selected two utterances of statements, and modified these 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 we had 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 we increased its value 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 we observed that questions were generally spoken faster than statements. We therefore used the *Praat ManipulationEditor* to increase the speech rate by 20%.

Shortening of Penultimate: Measurements had shown that the penultimate syllable in questions was considerably shorter than in statements. Therefore we created stimuli in which the penultimate was shortened to 70% of its original durations. 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. (4 levels of *Ap*, 2 speech rates, normal and shortened penultimate syllable).

2.3. Focus

All utterances in the production task had been uttered with zero contexts, yielding a default or broad focus. By post-hoc manipulation of the *F0* contour on certain words in an utterance we intended to explore whether a consistent change in meaning could be achieved, that is, whether increased prominence on certain words would modify the perception of focus or simply lead to unnaturally sounding stimuli. To this effect the amplitude *At* of tone commands associated with the items underlined in the following three sentences were increased in four steps of 0.15:

(1) Motsoala o rata ho seba. *My cousin likes to gossip.*
(2) Bona ba teng. *They are here.*
(3) Oa tena. *She is getting dressed.*

This yielded a total of 15 stimuli for the focus experiment. By providing subjects with three choices of matching questions, either (a) asking for the whole sentence, (b) the subject or (c) the predicate, we intended to detect changes in the meaning of the utterances:

choice	sentence (1)	sentence (2)	sentence (3)
(a)	What did you say?	What's happening?	What's happening?
(b)	Who likes to gossip?	Who is here?	Who is getting dressed?
(c)	What does my cousin like to do?	Where are they?	What is she doing?

3. Perceptual Experiment

The experiment was performed at the University of Stellenbosch (US) as well the National University of Lesotho (NUL). Stimuli were grouped in randomized sequences for all three sub-experiments and played back on a laptop computer over loudspeakers. The subjects consisted of 15 students of engineering at the US, 9 3rd-year students of linguistics at NUL, and 4 staff members at NUL. In total there were 17 male and 11 female subjects. 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 subject heard natural stimuli, and contrasts with respect to word difference, statement/question distinction and focus were presented. The correct answers to the warm-up trials were already listed in the questionnaire. Of the re-synthesized stimuli, 35 were presented twice in the experiment in order to test the consistency of the judgments.

4. Results

General Observations. At first the results from the two groups at US and NUL were evaluated separately. However, since the correlation of group result means was found to be .96, they were pooled for subsequent analysis. The correlation between results of the first and second presentation was .90, hence their average was taken for the repeated stimuli.

Lexical Identification. Figure 2 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. As can be seen, 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 we also expected to be associated with the low tone meaning of 'do mischief' are in contrast 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 second and third authors had classified them as 'do mischief' as expected.
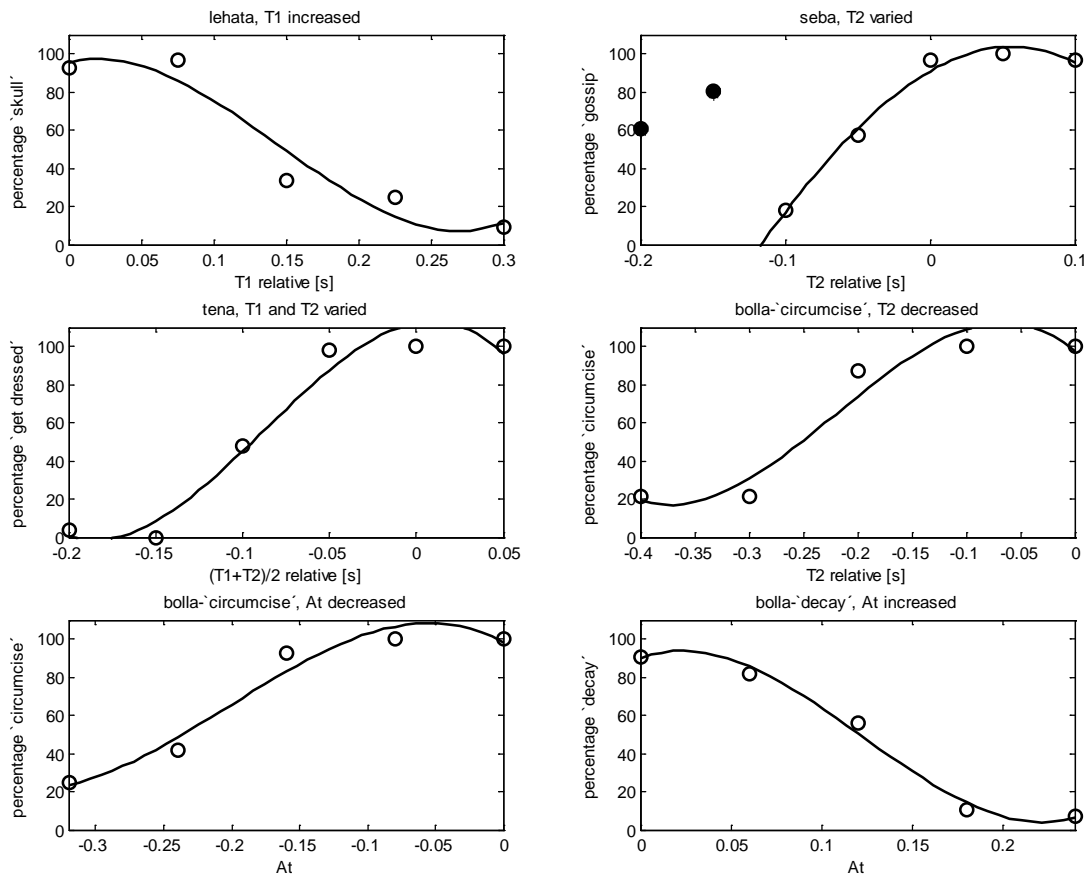

Figure 2: Results of the lexical identification perception experiment. Each panel displays the percentage of stimuli (circles) in which words are identified as having unmodified meaning as a function of the paramerter being modified. . The word meaning is that of the stimulus with unmodified *F0*. The 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.

When *At* is decreased in the high tone of the word bolla-*'circumcise'*, the hypothetical 50% boundary between the categories is located at *At*=-.24. In contrast, when increasing *At* on the low of tone bolla-*'circumcise'*, this boundary lies at *At*=+.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 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. We calculated 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*. These correlations are -.25, -.36, and -.83, respectively, but only the two latter values are significant (p < .05) and highly significant (p < .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 a statement by 97.6% of subjects. Shortening of the penultimate in the aforementioned stimulus reduces this figure to 66.1%. If in turn *Ap* is increased to the highest level, while all other features are left unchanged, the percentage of stimuli judged to be *statements* only reduced to 84.8%. If only the speech rate is increased, this figure drops 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.

Focus. Results for this experiment are inconclusive. In general, subjects complained about the difficulty of choosing the right question to match a given stimulus. Judgments are almost uncorrelated across subjects and often remained constant for one subject across a stimulus sequence or even for all stimuli. If we only consider the extremes of the stimulus continuum, that is, the unmodified versions and those with strong *F0* boosts on 'o rata ho seba', 'bona', and 'tena', respectively, the difference in the judgments is very small. 'o rata ho seba' is classified by 38.4% as the focused item in the unmodified version and by 57.5% when its *F0* is boosted by 10 semitones. For 'bona' the respective figures are 34.6% and 46.2%. In the case of 'tena' the effect is even reversed: In the unmodified version it is identified as the focused item in 76.9% of cases and only in 53.8% when its *F0* is raised. This might be due to the fact that raising 'tena' actually had the effect of turning the high tone 'oa'-*'she'* into the low tone 'ua'-*'you'*. Possibly as a consequence, broad focus judgments increased from 15.4 to 30.8%.

5. Discussion and Conclusions

This study presented a first study of the perception of tone and intonation in the Southern African language Sesotho. The limited dataset only permits tentative conclusions. All stimuli were produced by resynthesis with the Fujisaki model-based *F0* contours. Hence, 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 convert them into their low tone counterparts. Increasing *At* for a low tone word has the opposite effect. *F0* modifications even seem 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 second and third author. We can only speculate 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 thee modifications, the shortening of the penultimate syllable had the strongest impact on the judgments.

The focus experiment produced inconclusive results, possibly due to the fact that subjects were overtaxed in choosing the most appropriate question to match a given stimulus. Since *F0* manipulations were performed ad-hoc, and were not based on observations of natural utterances they may not have captured the way in which focus shift functions in Sesotho. Manipulations even affected the tones of certain words, as was observed for "Oa tena". In future experiments we will therefore investigate other strategies for eliciting focus judgments, either by having subjects underline the most prominent word, or by performing A/B comparisons.

6. Acknowledgements

7. References

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

[2] Jacottet, E. 1914. A practical method to learn Sesuto. Morija Sesuto Book Depot.

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

[4] Mohasi, L., Mixdorff, H. and Niesler, T., "An Acoustic Analysis of Tone in Sesotho", submitted to ICPhS2011, Hong Kong.

[5] Fujisaki, H., Hirose, K. 1984. Analysis of voice fundamental frequency contours for declarative sentences of Japanese, J. of the Acoust. Society of Japan (E) 5(4), 233-241.

[6] Doke, C. M., Mofokeng, M. 1957. Textbook of Southern Sotho grammar. Longmans, Green and Co., Ltd.

[7] Zerbian, S. 2007. Investigating prosodic focus marking in Northern Sotho. In: Hartmann, K., Aboh, E. & Zimmermann, M. (eds.). Focus strategies: evidence from African languages. Berlin: Mouton de Gruyter, pp. 55-79.

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

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