FROM TONE TO PITCH IN SEPEDI

Etienne Barnard[1], *Sabine Zerbian*[2]

[1] Human Language Technologies Research Group, Meraka Institute, CSIR, Pretoria, South Africa
[2] Department of Linguistics, University of Potsdam, Germany
ebarnard@csir.co.za, szerbian@uni-potsdam.de

ABSTRACT

We investigate the acoustic realization of tone in continuous utterances in Sepedi (a language in the Southern Bantu family). Human labelers marked each of the 271 syllables in a 15-sentence corpus produced by a single speaker as "high" or "low". Automatic pitch extraction was then used to estimate the fundamental frequencies of the voiced segments of each of these syllables. Statistical analysis of the resulting pitch contours confirms that the mean pitch frequencies of the syllabic nuclei serve as the primary indicator of tone, with the relative frequencies of successive syllables being the most relevant measure. Our analysis also suggests that additional factors may play a role in the production and perception of tone.

Index Terms— Tone languages, pitch contours, Sepedi, Southern Bantu

1. INTRODUCTION

Southern Bantu languages are tone languages in which word-level pitch variations generally convey both lexical and grammatical meaning. In contrast to tone languages like Chinese, they are agglutinative languages, i.e. several morphemes are joined together in a word. Although most Southern Bantu languages only have two level tones, namely high tone (H) and low tone (L), modeling of their prosody is complicated by the agglutinative morphology, the significant influence of grammar and the occurrence of tone sandhi within and across words. Given the role of word-level prosody in processes such as semantic interpretation and the production of natural speech, it is important that a detailed and systematic account of the prosody be given. Such an account is complicated by the fact that tonal information is not indicated in the orthography of many Bantu languages (including Sepedi, which is the focus of the current study).

We have recently presented an overview of intonation in the Southern Bantu languages [6], from which we concluded that a detailed understanding of the tone system of these languages is especially important for the creation of natural-sounding text-to-speech (TTS) systems. Such an understanding will require progress in two areas, namely (a) deriving tone assignments from text and (b) understanding the relationship between physical parameters (such as pitch frequency) and the tone levels. It is the second of these tasks that is the focus of the current investigation (initial work on the first task was presented in [8]).

Below, we briefly review a number of pertinent facts on tone in the Sotho-Tswana languages (of which Sepedi is a representative). We also summarize the goals of the current study in more detail, and present the experimental methodology that was followed in pursuit of these goals (Section 3). Our results are contained in Section 4, and Section 5 contains a discussion of our main conclusions and future work that is required to complete the current investigation.

2. TONE IN THE SOTHO-TSWANA LANGUAGES

Most Southern Bantu languages are tone languages whose surface tones can be captured by two level tones, namely high (H) and low (L) [3]. The high tone is the active tone in Sotho-Tswana languages such as Sepedi, as it participates in tone spread and is subject to positional restrictions. As is the case for most Bantu languages, the Sotho-Tswana languages show an asymmetry in the tonal characteristics of its noun and verb system with nouns being more tonal than verbs: whereas nouns can contrast tone on every syllable, verbs only contrast tone on their stem-initial syllable.

By definition, the primary distinctive feature of a level tone is the value of the pitch frequency within the nucleus of a given syllable, with H generally having a higher pitch frequency than L. This general observation was confirmed in our earlier investigations [7], which focused on the temporal alignment of a single high tone within the verbal domain. (As is common practice, we measure the fundamental frequency (F0) as a physical indicator of the pitch frequency.)

However, the more general question of how these pitch values are related to one another in a complete utterance, as well as the details of the temporal trajectories of F0 within and between syllables, have not been investigated systematically. The main aim of the current paper is to present initial findings on how these physical quantities are

related to surface tone values in Sepedi. That is, we seek to understand how a speaker of Sepedi chooses to express the difference between H and L tone levels, given the considerable latitude inherent in the specification that H should carry "higher pitch" than L.

3. METHODS AND CORPUS

Our analysis is based on the speech of a 30-year old male speaker, who was selected for the development of a Sepedi corpus for a concatenative text-to-speech system [5]. As part of that development, it was ensured that the speaker employs the standard Sepedi dialect, and he then recorded 299 sentences that give a balanced coverage of the most common diphones in Sepedi. In accordance with the requirements for TTS development with a limited corpus, the speaker was requested to speak naturally, but with relatively flat intonation.

Of these sentences, 15 were selected for analysis (based on factors such as the absence of loan words and proper nouns, and limitations on the mood of the verb to limit the influence of dialectal variations). These sentences were automatically aligned using a Hidden Markov Model recognizer. All syllables were subsequently labeled for tone by three labelers independently of each other. The labelers are sensitive to issues of tone but differ in their background and experience regarding Bantu tone. The individual labels were based on perception of the recorded sentences, acoustic analysis using the *Praat* software package [1] or both. The labeled sentences were compared across all three labelers, which revealed unanimous agreement on the tone labels in 72.3% of the cases (196 out of 271 syllables). A final transcription was generated based on the majority vote, i.e. the tone label selected by at least two labelers. (The flat intonation of the corpus might have been one of the reasons for cases of disagreement between labelers' decisions.)

The autocorrelation-based pitch tracker in *Praat* was employed to estimate the pitch contours (that is, the value of F0 as a function of time) for all utterances. As can be seen in Fig. 1, the computed contours are generally quite smooth (and the F0 values are found to be quite accurate). The exceptions generally occur at the edges of the voiced segments, where the voicing is generally less robust and the F0 estimates less accurate. Because of the smoothness of the pitch contours, we describe the F0 values of each syllable in terms of the smoothed initial and final F0 values of the vowel segment. These are calculated using the boundaries found by automatic alignment, as follows:

- The F0 values corresponding to the initial two pitch periods as well as the final two pitch periods are discarded.
- A least-squares linear fit of F0 as a function of time is computed from the remaining values.
- The initial and final values of F0 are estimated as the value of the linear fit at the respective boundaries of the vowel.

Mean pitch values in each segment, as well as the change in pitch across each segment, are estimated based on these linear parameters. Although this processing does markedly improve the robustness of the estimated values, it does not compensate completely for the pitch tracking errors that occur unavoidably. In addition, the automatic segmentation is not completely accurate, and the ambiguity in tone labeling also introduces some uncertainty. For all these reasons, there will be a fair amount of measurement noise in the results reported below; we return to this matter in the Conclusion.

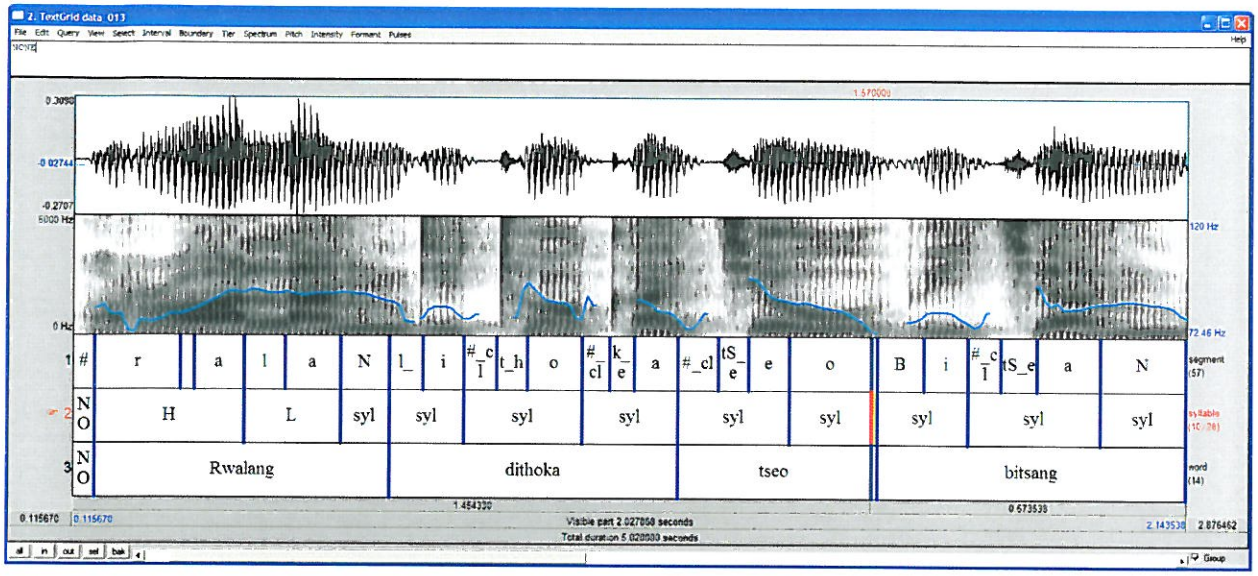

Figure 1: *Spectrogram and segmentation of typical Sepedi utterance used in the current study. The pitch contour is shown in blue, superimposed on the spectrogram.*

4. RESULTS

Figure 2 shows the overall distribution of mean F0 values for all segments in the corpus. As expected, the H segments generally have higher mean F0 values than the L segments, but there is considerable overlap between the two classes. This overlap is a predictable consequence of the fact that pitch generally declines throughout an utterance, so that both H and L pitch values are systematically reduced towards the end of each utterance. (The same tendency was, for example, observed for pitch levels in Mandarin [2].)

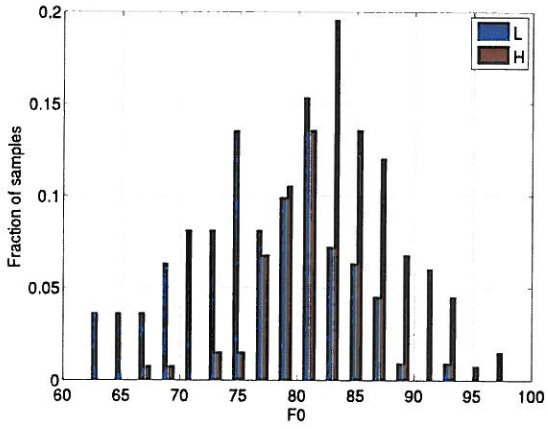

Figure 2: *Distribution of mean F0 values for H and L syllables, respectively*

If the declination in pitch were a complete explanation for the overlap in H and L pitch values, one would expect the relative values between consecutive segments to be a better indicator of the intended tone – such relative F0 values were indeed found to be indicators of F0 perception in Vietnamese [4]. In Figures 3 and 4 we therefore show histograms of the difference between the mean F0 values of successive pairs of vowels, where the first vowel is labeled as H and L, respectively. It can be seen that L-to-H transitions tend to produce an increase or slight decrease in the mean pitch, whereas H-to-L transitions tend to result in a large decrease in mean pitch; L-to-L and H-to-H transitions fall somewhere between these extremes. Statistics confirming these tendencies are presented in Table 1. Note, however, that significant overlaps occur between all four cases, suggesting that the relative mean pitch values do not offer a complete expression of the speaker's intended tone level.

Condition	Mean change in F0 (Hz)	Standard deviation of change in F0
L-H transitions	4.881	5.650
H-H transitions	-0.183	4.959
L-L transitions	-3.452	5.888
H-L transitions	-6.409	4.576

Table 1: *Statistics of changes in mean F0 values between successive syllables.*

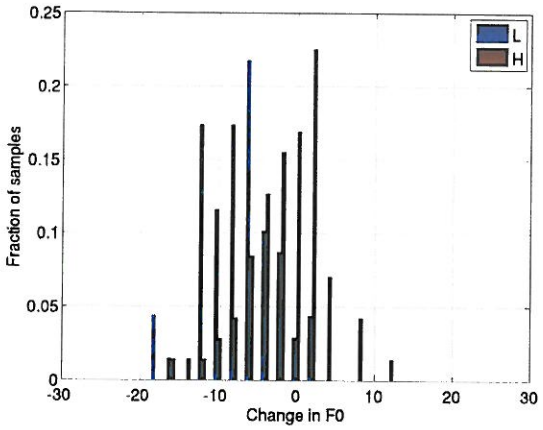

Figure 3: *Distribution of change in mean F0 values* ***between successive*** *syllables, when the first syllable was H.*

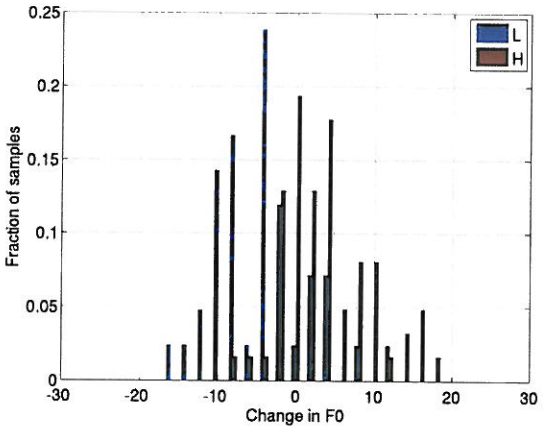

Figure 4: *Distribution of change in mean F0 values* ***between successive syllables****, when the first syllable was L.*

Inspection of pitch tracks such as that shown in Fig. 1 suggests another possible source of distinction between H and L, namely the overall slope of the pitch contour within a syllable (or syllable nucleus). As can be seen in Fig. 5, which represents a histogram of the overall changes in (smoothed) F0 values within each syllable nucleus, this feature does indeed take on somewhat different values for the two tones (though it is not strongly distinctive). The histograms of this feature for the various transitions (Figures 6 and 7) demonstrate that this feature is virtually irrelevant for syllables following an H syllable, but that it is somewhat distinctive for syllables preceded by an L syllable.

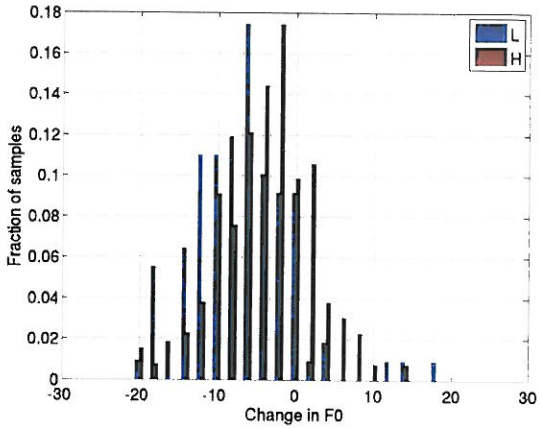

Figure 5: *Distribution of change in F0 within the syllable nucleus for H and L syllables, respectively*

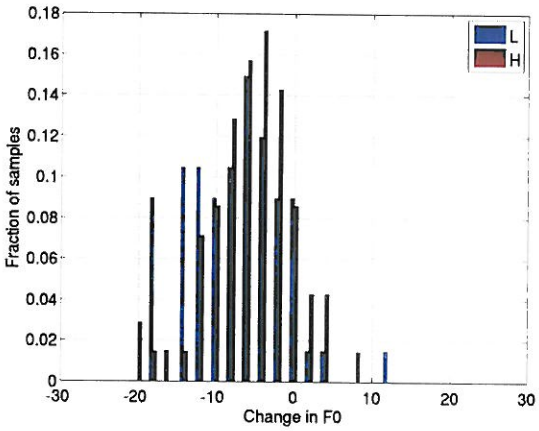

Figure 6: *Distribution of change in mean F0 values* ***within the syllable nucleus****, when the first syllable was H.*

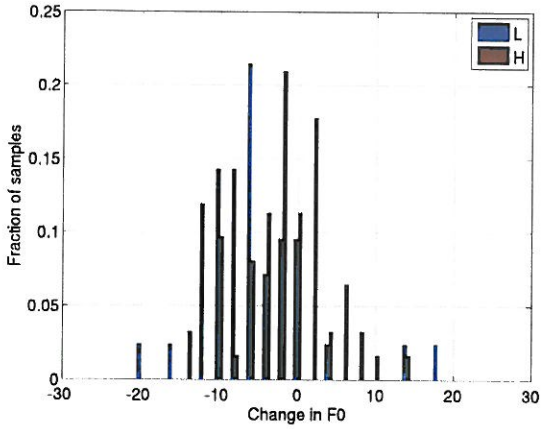

Figure 7: *Distribution of change in mean F0 values* ***within the syllable nucleus****, when the first syllable was L.*

5. CONCLUSION AND OUTLOOK

We have found that the mean pitch within the syllable nucleus is a strong indicator of the tone perceived in Sepedi speech. Not surprisingly, we find that the absolute pitch level is less important than relative pitch, which implies that the *change* in the mean pitch is the strongest indicator of tone amongst the signatures investigated here.

The change in mean pitch is nevertheless not a perfect indicator of tone in our data, as indicated by the overlap of the histograms shown in Figures 3 and 4. It is possible that these overlaps are simply the result of ambiguities in the tone labels and errors in alignment and pitch extraction (as discussed in Section 3). Some of the outliers in our results can certainly be attributed to such factors; however, the large number of syllables with overlapping values for the change in F0 leads us to suspect that other factors may be at stake. Figure 7 suggests that, in some cases, the intra-syllabic trend of F0 may be used to indicate tone, for syllables following an L syllable. Other factors that we have investigated were less promising – for example, consideration of the tone and mean pitch values of surrounding syllables does not produce better separation of the low and high tones. We have seen some evidence that the segmental make-up of a syllable may have an effect on the way that tone is expressed [6], but in the current corpus that influence is not evident.

To resolve these issues, we plan to analyze larger sets of sentences. It will be useful to consider speech from other speakers, to learn whether different speakers employ different strategies to communicate tone. It will also be interesting to perform comparative analyses of other Southern Bantu languages: whereas closely related languages such as Sesotho and Setswana are expected to be quite similar to Sepedi with respect to the phonetics of tone, somewhat more distant languages (e.g. isiXhosa and isiZulu) are likely to display some additional phenomena (e.g. depressor consonants [3]).

The successful application of these insights in speech-technology systems will be strong confirmation of their validity. We are in the process of developing all the components necessary to build a tone-aware TTS system for the Sotho-Tswana languages – the algorithm for tone assignment from text [8] is partially worked out, and the compilation of a sufficiently complete tone-marked pronunciation dictionary remains as the biggest challenge in that regard. The completion of this TTS system will allow us to carry out comprehensive perceptual tests to evaluate our ability to model tonal processes in Sepedi.

6. ACKNOWLEDGEMENTS

The authors gratefully acknowledge the assistance of Ben Khoali, who assisted with the tone marking process and Charl van Heerden, who provided technical assistance. EB was generously supported by a Google Research Award.

7. REFERENCES

[1] Boersma, P., *Praat, a system for doing phonetics by computer.* Glott International, Amsterdam 2001

[2] Chen, S-H and C-C. Kuo, "Perceptual Relevance of Pitch Contours of Mandarin Tones and its Efficacy in Prosody Generation of Speech Synthesis" In *Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech 2007)*, Antwerp, Belgium, pp. 2669-2672, 2007

[3] Kisseberth CW and D. Odden. "Tone." In Nurse D & Philippson G (eds) *The Bantu languages.* Routledge, London, New York: , pp. 59–70. 2003

[4] Tran D.D, Castelli E., Serignat JF., Le X.H., Trinh V.L., "Influence of F0 on Vietnamese syllable perception", In *Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech 2005)*, Lisbon, Portugal, pp. 1697-1700, 2005.

[5] Van Niekerk, D.R. and E. Barnard, "Phonetic alignment for speech synthesis in under-resourced languages" In *Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech 2009)*, Brighton, UK, 2009

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

[7] Zerbian, S. and E. Barnard. "Realizations of a single high tone in Northern Sotho", *Southern African Linguistics and Applied Language Studies* 27(4): 357–379, 2009

[8] Zerbian, S. and E. Barnard, "Word-level prosody in Sotho-Tswana", *submitted for publication*, 2010