Computational models of prosody in the Nguni languages

N. Govender, C. Kuun, V. Zimu, E. Barnard and M. Davel

Human Language Technologies Research Group
Meraka Institute / University of Pretoria, Pretoria, 0001
ngovender@csir.co.za

Abstract

We investigate two related issues in the computational modeling of Nguni prosody, based on annotated databases of isiZulu and isiXhosa speech. Firstly, we show that a simple template can be used to describe the tonal characteristics of vowels and adjectives spoken in isolation, and that contextual effects have only a mild impact on this template. This analysis was based on a simple mapping between pitch and tone; in the second part of the paper, we show that pitch and amplitude actually play comparable roles in producing tonal percepts.

1. Prosodic models in multilingual speech technology

Although the complexity of prosody is widely recognized [8], the lack of widely-accepted descriptive standards for prosodic phenomena have meant that prosodic systems for most of the languages of the world have, at best, been described in impressionistic rule-based terms. This situation has become particularly noticeable with the development of increasingly capable text-to-speech (TTS) systems [2]. Such systems require detailed prosodic models to sound natural, and the development of these detailed models poses a significant challenge to the descriptive systems employed for prosodic quantities. For languages such as English or Japanese, for example, the ToBI marking system [1] has gained a significant following because of its utility in producing predictions for these quantities. These models allow developers to employ the methods of pattern recognition to compute numerical targets for the fundamental frequency and amplitude of spoken utterances, based on their written representation.

For the languages of Southern Africa, the deficiencies in our modeling capabilities are acute. It has long been recognized that, for example, the languages of the Nguni family (such as isiZulu and isiXhosa) have an intricate tonal structure – in fact, the adequate description of this structure was one of the major early successes of autosegmental phonology. However, little work of a quantitative nature has been published, and as Roux [11] points out, there are significant contradictions and imprecisions in the literature on this topic, which partially stems from the lack of quantitative, measurement-driven analysis.

In the current paper we detail initial results from a program that we have initiated in order to develop detailed, reliable intonation models for the languages of Southern Africa. In particular, we discuss various measurements that have been obtained in order to model the fundamental frequency contours of isiZulu and isiXhosa, and report on initial investigations on the relationship between pitch, intensity, and lexical tone.

A wide-ranging overview over intonation in numerous languages is provided in [8]; here, we briefly review some of the facts pertinent to our investigations – partially to fix terminology, since there is not universal agreement on the semantics of this domain. We use the terms prosody and intonation interchangeably to refer to the melodic pattern of an utterance. In other words, it is the non-phonetic content of speech; at the linguistic level, this is represented by variables related to *tone*, *stress* and *rhythm*. These variables are either attached to specific words, in which case they are called lexical quantities, or to (generally) larger units, in which case they are tagged as supralexical or syntactic. Corresponding to these linguistic variables are a number of physically measurable quantities – most noticeably fundamental frequency, intensity and duration. Although fundamental frequency generally is most strongly correlated with tone, intensity with stress, and duration with rhythm, this correspondence is far from perfect. Thus, stress may be indicated with changes in fundamental frequency or duration as well.

2. Intonation in the Nguni languages

In *tone languages*, lexical tone can be used to attach different meanings to words which share the same phonemic content. Thus, the contour of fundamental frequencies that accompany a particular utterance is the result of a complicated interaction between such lexical tones and the supralexical tonal content present in any utterance. (Speakers of non-tonal languages can gain an impression of these phenomena by considering the amplitudes assigned to the word “present” in the two sentences “Present yourself!” and “He gave you a present?”. Word-level stress, pragmatic accent, and phrasal paradigm combine in subtle yet predictable ways to create several contrasts: verb vs. noun, non-emphasized vs. emphasized, command vs. question.) In the current section, we focus on the relationship between tone and fundamental frequency; later, we also investigate how intensity influences the perception of tone.

The Nguni languages (and the Southern Bantu languages in general) have interesting tonal characteristics, which have been the topic of extensive research. In early work, Doke [6] distinguished nine different lexical tone levels in isiZulu; subsequent theoretical advances have simplified this description, and three tone assignments (low, high, and falling) are currently thought sufficient to describe the words of isiZulu[10] – or possibly only the first two. However, in these modern formulations, the rules for assignment of tone levels to specific syllables are quite complex [9], and we appear to be a long way from the mathematically precise formulations that have been so useful for TTS in languages such as English, Japanese and German.

To develop such a model, one must find a way to relate the measured values of fundamental frequency (F0) to the tone values assigned to words. In principle, this is rather straightforward: abstract tone assignments can be produced for a num-

ber of written utterances, and these can be correlated with the F0 values measured when a first-language speaker speaks those same utterances. In practice, though, several issues need to be addressed. Firstly, the measured F0 values depend on a number of factors besides the tones assigned to the word. These include

- *the nominal F0 range of the speaker* – females, for example, tend to have higher mean F0 and larger variance than males;
- *the lexical context of the word* – the tone values of surrounding words often influence the way F0 is realized in a given word;
- *the phonetic content of the word* – certain phones tend to be realized with lower F0 than others;
- *the position of the word in the sentence* – F0 tends to decline continuously throughout a phrase, and
- *pragmatic effects* – the speaker's decision to emphasize certain words, to pose a question or direct a command, may all affect the F0 values produced.

In addition, the tone values are themselves not straightforwardly assigned. Doing so from a well-formulated theory (e.g. autosegmental theory) would require knowledge of tone values for all morphemes in a languages, as well as a solid grasp of a complex set of rules. In addition, competing theories may well produce conflicting assignments[11]. Subjective assignment by first-language speakers, on the other hand, depends on the dialect of the speaker (which in turn depends on factors such as the region where the speakers grew up and currently reside, possibly their ages and socio-economic environment, etc.)

To address these issues, we have chosen to start with a small set of speakers, speaking words in isolation, carefully controlled contexts, or natural utterances. These utterances have been tone marked by speakers with similar backgrounds to those producing the recordings, but with no knowledge of formal theories of lexical tone assignment. We assume a three-step approach to prosodic modeling: the first step assigns lexical tones to words in isolation, the second applies contextual and grammatical rules to compute abstract tones for words in sentence context, and the final step is to model the observed fundamental frequency based on the sequence of abstract tones. In Section 3 below, we focus on the first two steps, and Section 4 contains results relevant to the third step.

3. Predicting abstract tone from text

3.1. Subjective evaluation of nouns and adjectives

We work with recorded utterances that fall into 4 different categories:

1. Randomly selected isiZulu words.
2. Such words embedded in "carrier sentences".
3. Randomly selected isiZulu sentences, as well as the individual words that make up those sentences.
4. isiZulu sentences that have emphasis placed on selected words.

The categories were selected in order to study the phenomena described in Section 2. For example, a carrier sentence may be "I am now going to say the word 'apple' "; the word "apple" can be substituted with any chosen phrase, to study both the intrinsic tone of the word and the effects of the carrier context on these words.

Recordings were obtained from one female and two male isiZulu speakers, and all utterances were analyzed with the Praat pitch tracker [3] in order to compute the contours of fundamental frequency (F0). Separately, a first-language isiZulu speaker marked each syllable in the text as "High" or "Low" according to her subjective expectation for the surface realization of each utterance.

Let us initially focus on the isolated words, and tentatively mark the initial syllable of a word as "(h)igh" if its mean F0 is above the mean F0 of the word, and "(l)ow" otherwise. Subsequent syllables are marked with the same label as long as they are within approximately 20 Hz of the preceding syllable; otherwise, the label becomes "h" or "l" depending on the direction of the change, or "(r)ising" or "(f)alling" if F0 changes by more than that amount *within* the syllable. (In Section 4 we describe more sophisticated algorithms for tone assignment, but this simple approach was sufficient for the current purposes.) If this convention is applied to randomly selected nouns, results such as those in table 1 are obtained. Similarly, representative results for adjectives are shown in table 2. For nouns we consistently observe a 'high(*)-falling-low' pattern, with the 'falling' label invariably assigned to the second last syllable and the last syllable being 'low'. There were very few exceptions, the most common exception being that the word displayed a high-low-low pattern. The evidence in this regard conflicts significantly with the range of noun patterns reported, for example by Goldsmith [5] as well as Poulos [10] in previous intonation studies. We therefore present further analysis of these observations below.

Word	Segmentation	Tone
amanzi	a-ma-n-zi	hhfl
ilanga	i-la-n-ga	hhfl
abantu	a-ba-n-tu	hhfl
inja	i-n-ja	hfl
ubuhle	u-bu-hle	hfl

Table 1: *Examples of observed F0 values for the* ***Nouns***

Word	Segmentation	Tone
abathathu	a-ba-tha-thu	hhfl
omusha	o-mu-sha	hfl
amahle	a-ma-hle	hfl
esikhulu	e-si-khu-lu	hhfl
enkulu	e-n-ku-lu	hhfl

Table 2: *Examples of observed F0 values for the* ***Adjectives***

3.2. Measurements

In order to accurately categorize the observed tone values in the utterances, it is necessary to take two types of measurements: the initial and final value of the pitch in every syllable of a word, as well as the average value of each syllable in a word. These parameters were measured using Praat [3]. Three separate cases were evaluated, the nouns up to a maximum of six syllables (46 words were used), the adjectives with a maximum of 5 syllables (21 words were used), and nouns occurring in a sentence with a maximum of 4 syllables (9 instances). A rough representation of the pitch as computed by Praat in shown in figure 1; the figure

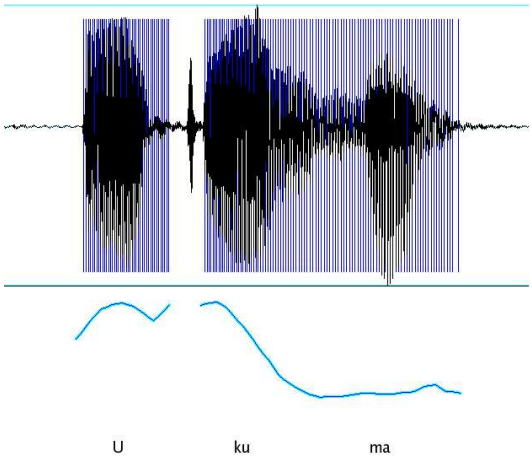

Figure 1: *F0 contour computed by Praat; the word used in this case is "ukuma".*

indicates a typical high-falling-low pattern.

3.3. Results

Measurements were obtained by listening to each utterance, finding the syllable boundaries, and then using F0 measurements at the appropriate boundaries. The initial and final values of the syllables were used to compute the change in frequency across each syllable, referred to here as Δf. The mean value of the Δf values, for each of the cases examined were then computed and can be seen in figures 2, 3 and 4.

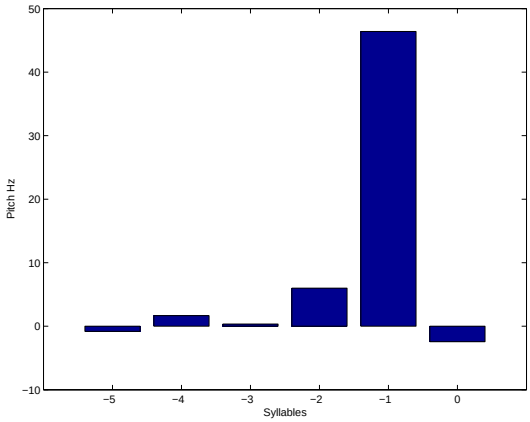

Figure 2: *Mean change in F0 across each syllable for 46 nouns spoken in isolation. In order to group words with different numbers of syllables together, syllables are counted backwards from the end of the word.*

In figures 5, 6 and 7 the mean values in the center of each syllable, and the variances around these means, are represented by a solid line and a set of error bars, respectively (each error bar represents a mean plus or minus one standard deviation). Note that these values are all computed as changes with respect to the F0 value at the beginning of the word.

Finally, figures 8, 9 and 10 represent the classifications as-

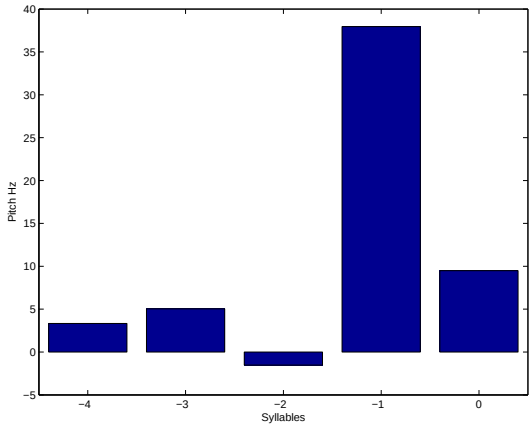

Figure 3: *Mean change in F0 across each syllable for 21 adjectives spoken in isolation.*

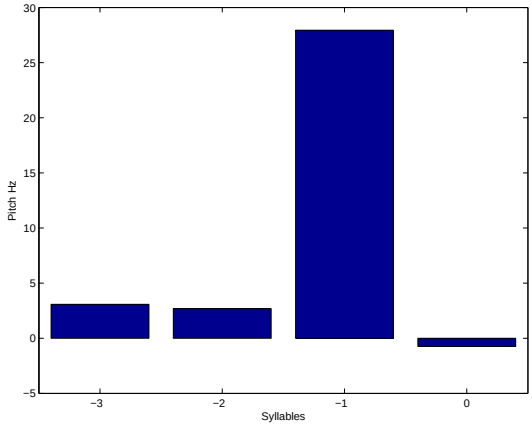

Figure 4: *Mean change in F0 across each syllable for 9 nouns spoken in sentence context.*

signed to each syllable according to the algorithm described in Section 3.1.

3.4. Discussion

For nouns and adjectives spoken in isolation, a highly consistent "h(*)fl" pattern is observed – as seen in the Δf values, the mean values and standard deviations, and the classifications assigned. For the words in sentence context, the observations are somewhat less consistent, but broadly similar.

These observations are to be contrasted with the wide variety of tonal patterns assigned to nouns and adjectives by, for example Doke [6], Rycroft [12] and Poulos [10]. Although those assignments are phonemic, one would expect to see at least some of the variability of the different tone patterns in the surface form. It therefore seems that our results are not compatible with the proposals of these earlier authors. A number of explanations for these differences may be considered:

1. *It may be that our words were by chance all selected from the same tonal class (in classical terms).* However, to the extent that the classification by Doke is still accepted, we can unambiguously state that this is not the case.
2. *The dialects of our speakers may have lost the classical tonal distinctions.* Although two of our three speakers

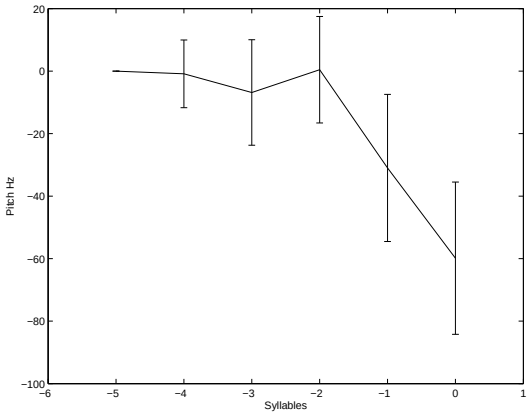

Figure 5: *Mean F0 of each syllable for 46 nouns spoken in isolation; error bars represent the mean plus or minus one standard deviation.*

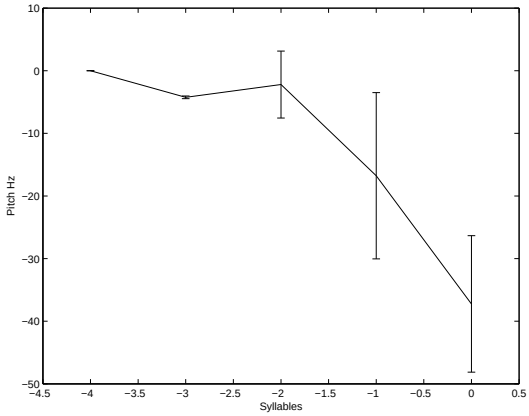

Figure 6: *Mean F0 of each syllable of 21 adjectives spoken in isolation*

are originally from the Kwazulu region, all three have spent at least five years in Gauteng, and all three are between 20 and 35 years old.

3. *Our experimental protocol may somehow have suppressed the tonal variation classically suggested.* For example, the speakers may implicitly have de-emphasized all words, and a specific request to produce emphasized or contrasted forms may elicit the expected differences. (If this is the case, the uniformity of de-emphasized variants would nevertheless be an interesting discovery.)

4. *The tonal classes may be expressed in other physical measurements than F0* – for example, Roux [11] has noted interesting correlations between tone and amplitude in isiXhosa, another language in the Nguni family. However, our initial experiments, summarized below, show that these physical measurements tend to be highly correlated, implying that similar conclusions would be reached with other physical measurements.

Understanding this apparent conflict is an important goal of our future work.

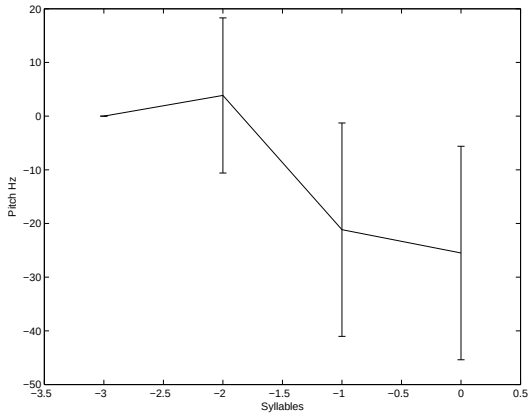

Figure 7: *Mean F0 of each syllable of 9 nouns spoken in sentence context.*

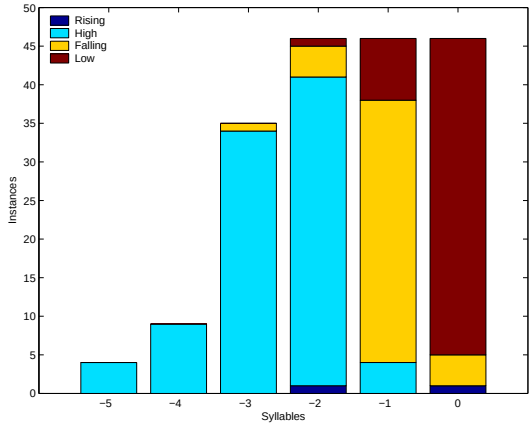

Figure 8: *Classification of each syllable for 46 nouns spoken in isolation.*

4. The physical correlates of abstract tone

In order to analyze the characteristics of lexical tone, we employed a very simple algorithm to relate pitch and tone in Section 3. However, there is little doubt that additional factors influence listeners' perception of tone[8, 11], and that the nature of these influences is more involved than in our simple model. In order to address this issue, we present initial statistical results on the mapping between abstract tone and physical variables such as amplitude and fundamental frequency in the current section.

4.1. Approach

As described in [7], we have collected a corpus of speech by one native male speaker and one native female speaker in each of the Nguni languages isiZulu and isiXhosa. In order to understand how the 'expected' intonation relates to the actual measured characteristics, the syllabic intonation was subsequently marked as either High (H) or Low (L) depending on how utterances were expected to be pronounced in the context of the sentence, without using the voice recordings as guide. (Because of the ambiguous theoretical status of falling tone in these languages, we chose not to mark it in these transcriptions.) These markings were performed by a first language isiXhosa speaker for the isiXhosa sentences and a first language isiZulu speaker

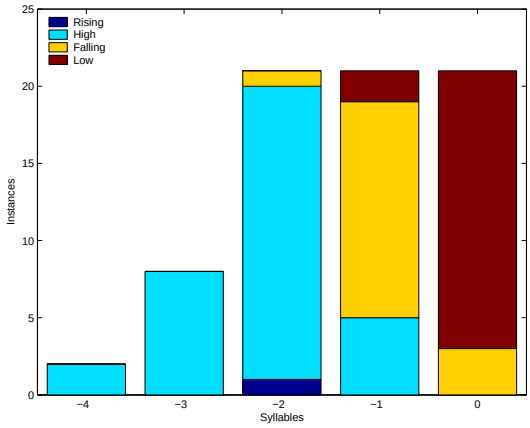

Figure 9: *Classification of each syllable of 21 adjectives spoken in isolation*

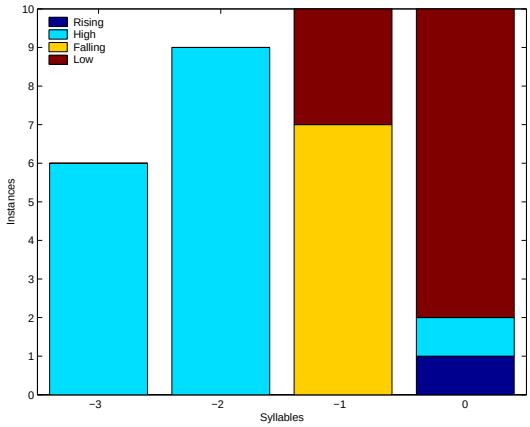

Figure 10: *Classification of each syllable of 9 nouns spoken in sentence context.*

for the isiZulu sentences. Note that different speakers were used for recording and transcription, i.e. these markings were not influenced by the available audio data. Our goal was to train an automatic classifier to assign either an 'H' or an 'L' to a segment, based on the tone assigned to the preceding segment and the measured F0 and intensity values of both the current and the preceding segments.

The fundamental frequency (F0) values were extracted at the syllable boundaries using the public-domain toolkit Praat [4]. (Actually, these values were extracted at the first and last speech frames classified as voiced by Praat.) The intensity was calculated at each of the syllable boundaries, as the average squared value of the signal within a 5 millisecond window, and the peak value within the syllable boundaries was also computed. Based on these values, we computed a set of F0-derived features and a set of intensity-derived features.

For each language, we trained two types of classifiers, depending on whether the previous state had been an 'H' or an 'L'. These classifiers were trained on training data as shown in Table 3, and evaluated on a separate set of test utterances (though from the same pair of speakers as the training data, since our goal was not to construct a speaker-independent tone-assignment algorithm). As shown in Table 4, we compared the classification accuracies achievable with the F0-derived features to those of

isiZulu		
	Training	Testing
Utterances	100	28
Syllables	2243	808
isiXhosa		
	Training	Testing
Utterances	50	15
Syllables	957	308

Table 3: *The number of utterances and syllables used for the training and testing of the classifiers.*

the amplitude derived features, and to the combination of these two types features.

isiZulu		
	High	Low
Pitch	67.71	74.44
Intensity	71.47	74.03
Both	77.74	76.48
isiXhosa		
	High	Low
Pitch	83.61	87.45
Intensity	85.25	83.40
Both	86.89	87.45

Table 4: *Accuracy obtained when classifying the tone of a syllable based on features derived from F0, intensity, or both measurements, for preceding High and Low tones, respectively.*

The results in 4 suggest a number of significant conclusions.

- Most importantly, we were able to construct reasonably accurate classifiers for all four subproblems (i.e. those designed for 'H' and 'L' preceding states, respectively, in both languages), despite the fact that the transcribers had produced their predictions without access to any acoustic data. This suggests that such surface-form tone assignments can be made with a fair amount of reliability.
- In all four cases, the F0-based features and the amplitude-based features produce comparable accuracy. This lends independent support to the hypothesis advanced in [11] regarding the substantial role of amplitude / intensity in the perception of tone – based on our analysis, amplitude may even be somewhat more important than F0 in this determination.
- Combining F0 and amplitude information produces lower classification accuracy than would be expected if these had been independent information sources. One can therefore conclude that the speakers tend to encode the same tonal information in both physical aspects, in a consistent manner.

A variety of factors may be responsible for the relatively better results obtained for isiXhosa in comparison with isiZulu, ranging from more significant dialectal differences between transcribers and speakers in isiZulu, through personal idiosyncrasies, to inherent languages differences. More data would be needed to distinguish between these possibilities.

5. Conclusion

We have described a data-driven approach that can be used to derive computational models for intonation in the Nguni languages. Our early experiments were focused on lexical and grammatical tone in isiZulu and isiXhosa. We have found that a very simple model accounts for the observed tonal patterns of nouns and adjectives in isiZulu, and that amplitude and fundamental frequency play comparable roles in conveying tonal content. Although our analysis of the tonal patterns was based on fundamental frequency by itself, the observed consistency between these two information sources implies that the same results would have been obtained if both had been considered.

The objective nature of our approach allows us to investigate the validity of our findings with larger speaker groups, additional word classes, and other languages in the Bantu family. These extensions are an important focus of our current work.

6. References

[1] M. E. Beckman and J. B. Pierrehumbert. Intonational structure in Japanese and English. In *Phonology Yearbook 3*, pages 255–309, 1986.

[2] A. Black, P. Taylor, and R. Caley. The Festival speech synthesis system, 1999. http://festvox.org/festival/.

[3] Paul Boersma. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *Proceedings of the Institute of Phonetic Sciences, University of Amsterdam*, pages 97–110, 1993.

[4] Paul Boersma. Praat, a system for doing phonetics by computer. *Glot International*, pages 341–345, 2001.

[5] G. N. Clements and J. Goldsmith. *Autosegmental studies in Bantu tone*. Foris Publication, 1984.

[6] C. M. Doke. *Text-book of Zulu grammar*. London: Longmans, Green and Co., 1947.

[7] N. Govender, E. Barnard, and M. Davel. Fundamental frequency and tone in isiZulu: initial experiments. In *Interspeech: 9th annual International Conference on Spoken Language Processing*, pages 255–309, 2005.

[8] Daniel Hirst and Albert Di Cristo. *Intonation Systems*. Cambridge University Press, 1998.

[9] M. Laughren, G. N. Clements, and J. Goldsmith. *Tone in Zulu Nouns. Autosegmental Studies in Bantu Tone*. Dordecht: Foris, 1984.

[10] George Poulos and Christian T. Msimang. *A Linguistic Anaylsis of Zulu*. Via Afrika, 1998.

[11] J. C. Roux. Xhosa: A tone or pitch-accent language? *South African Journal of Linguistics*, pages 33–50, 1998.

[12] D. K. Rycroft. Nguni tonal topology and common Bantu. *African Language Studies, XVII*, pages 33–76, 1980.