Statistical investigations into isiZulu intonation

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

Human Language Technologies Research Group
Meraka Institute / University of Pretoria, Pretoria, 0001
chrispieter@gmail.com,{vzimu,ebarnard,mdavel}@csir.co.za

Abstract

We study the relationship between lexical tone and measured values of fundamental frequency for several word classes in isiZulu, based on objective measurements from recordings of three speakers. For nouns and adjectives, both in isolation and in sentence context, a relatively invariant pattern of fundamental frequencies is observed. This observation is in apparent conflict with the range of tonal classes conventionally ascribed to the nouns and adjectives, and suggests the need for further research.

1. Prosodic models as statistical pattern recognition

Many sub-fields within patten recognition have seen the replacement of impressionistic, rule-based approaches with statistical methods – for example, Hidden Markov Models in speech recognition, or feature-based methods in computer vision. These statistical methods are often less intuitive than their rule-based predecessors, but succeed because they are based on measured data rather than subjective impressions.

A field where rule-based methods currently hold sway is the modelling of intonation in many of the world's languages. Although the complexity of prosody is widely recognized [6], 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.

In the current paper we detail initial results from a programme that we have inititiated in order to develop similarly 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, a language in the Nguni family. We also present tentative hypotheses on the modelling of nouns and adjectives, along with the measurements that bear on those hypotheses.

A wide-ranging overview over intonation in numerous languages is provided in [6]; 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 noticably 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 isiZulu

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 assgined 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.)

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 [5] 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[9] – or possibly only the first two. However, in these modern formulations, the rules for assignment of tone levels to specific syllables are quite complex [8], 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 or 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 number 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 postion 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. 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 or carefully controlled contexts. 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.

3. Methodology and subjective results

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. Our initial intent was to study the relationship between the observed F0 contours and the predicted tone assignments in order to understand the tone-to-F0 mapping; however, as we detail below, an unexpectedly simple pattern was observed in the F0 contours of certain word classes.

In particular, let us 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 lable as long as they are within approximately 20 Hz of the preceding syllbable; otherwise, the lable 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. 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' lable 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 could be characterised by a high-low-low pattern. The evidence in this regard conflicts significantly with the range of noun patterns reported, for example by Goldsmith [4] as well as Poulos [9] 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**

4. Measurements

In order to accurately categorize the observed tone values in the utterances, it is neccessary 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 seperate 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 ocurring in a sentence with a maximum of 4 syllables (9 instances). A rough representation of the pitch as computed by Praat is shown in figure 1; the figure indicates a typical high-falling-low pattern.

4.1. 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

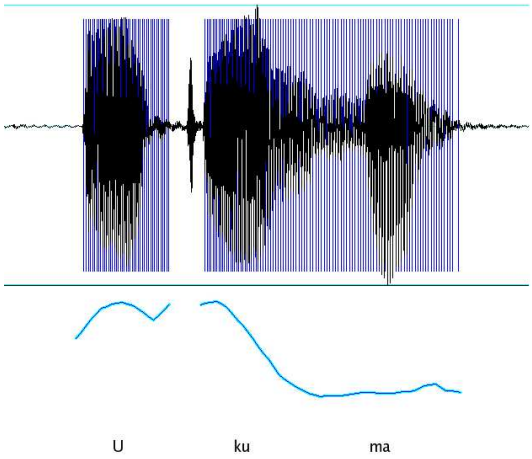

Figure 1: *F0 contour computed by Praat; the word used in this case is "ukuma".*

the Δf values, for each of the cases examined were then computed and can be seen in figures 2, 3 and 4.

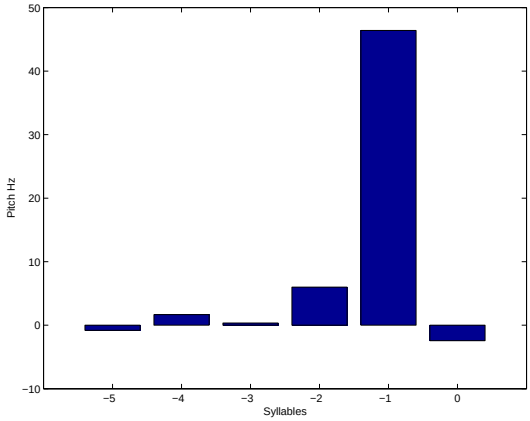

Figure 2: *Mean change in F0 across each syllable for 46 nouns spoken in isolation. In order to group words with different numbers of syllables together, syllables are counted backwards from the end of the word.*

In figures 5, 6 and 7 the mean values in the centre 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 assigned to each syllable according to the algorithm described in Section 3.

5. 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

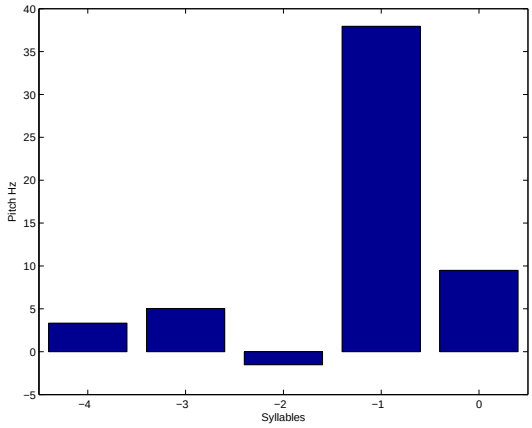

Figure 3: *Mean change in F0 across each syllable for 21 adjectives spoken in isolation.*

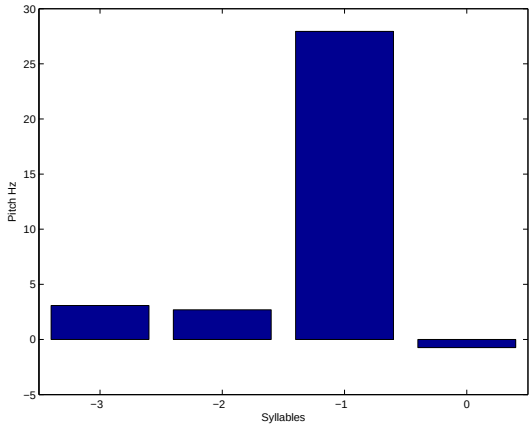

Figure 4: *Mean change in F0 across each syllable for 9 nouns spoken in sentence context.*

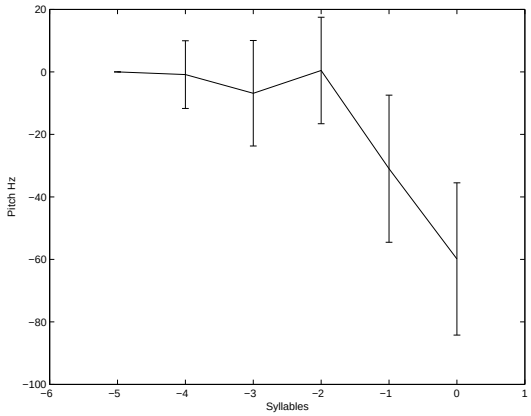

Figure 5: *Mean F0 of each syllable for 46 nouns spoken in isolation; error bars represent the mean plus or minus one standard deviation.*

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

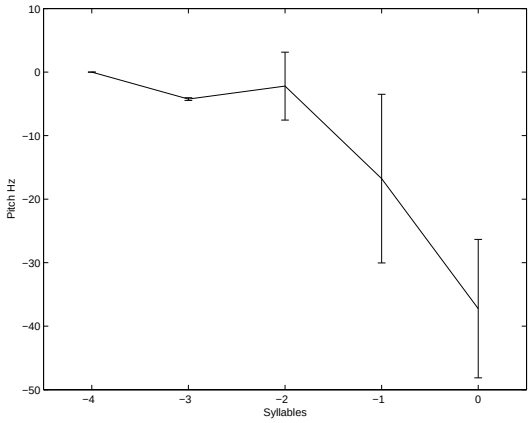

Figure 6: *Mean F0 of each syllable of 21 adjectives spoken in isolation*

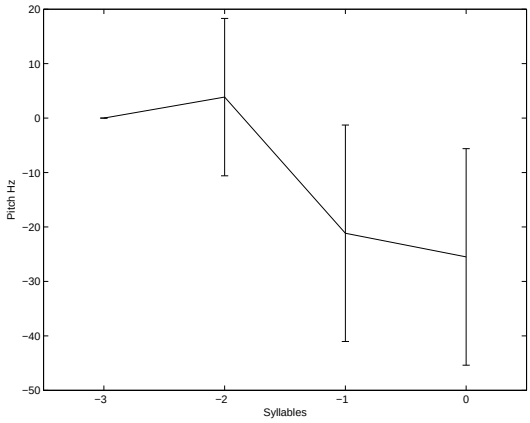

Figure 7: *Mean F0 of each syllable of 9 nouns spoken in sentence context.*

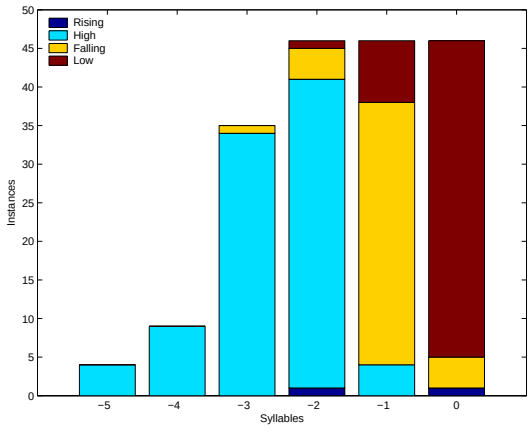

Figure 8: *Classification of each syllable for 46 nouns spoken in isolation.*

example Doke [5], Rycroft [11] and Poulos [9]. 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:

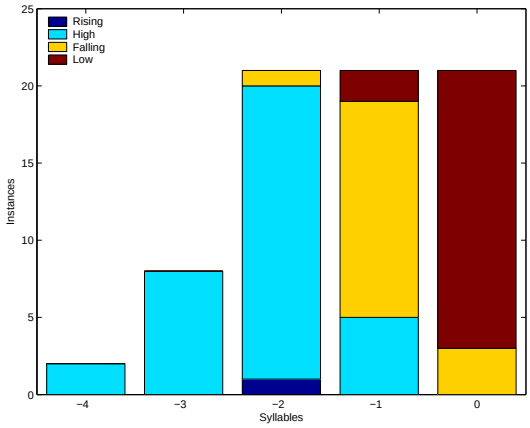

Figure 9: *Classification of each syllable of 21 adjectives spoken in isolation*

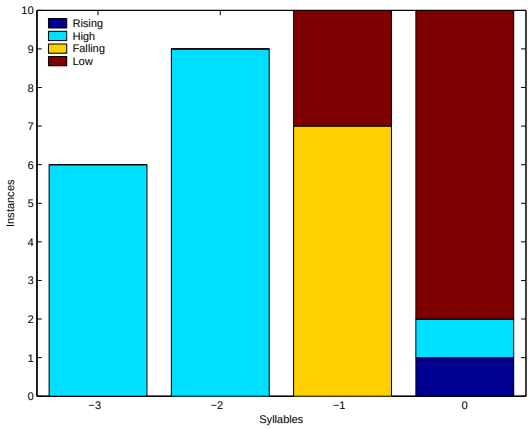

Figure 10: *Classification of each syllable of 9 nouns spoken in sentence context.*

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 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 [10] has noted interesting correlations between tone and amplitude in isiXhosa, another language in the Nguni family. A cursory analysis does not support this hypothesis, but further research is required.

Understanding this apparent conflict is an important goal of our future work, as discussed below.

One interesting exception encountered during our study deserves further mention. Five of the nouns in the database showed a pitch contour similar to that in figure 11. It can be observed that during the falling syllable "zi", there is a small mid-syllable rise, followed immediately by a continued drop in F0. This behaviour still remains unexplained and has been identified as another area of future research.

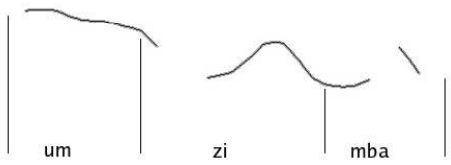

Figure 11: *F0 contour for "umzimba".*

6. Conclusion

By systematically measuring F0 values for several word classes, a significant simplification was observed for nouns and adjectives. In order to explain the conflict between this simplicity and the more complex descriptions found in the literature, we intend to (a) record additional speakers, (b) focus on nouns and adjectives that are assigned to different tonal classes by consensus in the literature, and (c) investigate the influence of additional factors such as emphasis or contrast. We will also study the behaviour of other variables such as amplitude and duration.

Although our initial explorations into other word classes have produced a number of interesting suggestions, those are less conclusive than the results presented in Section 4.1. We therefore continue to investigate those classes as well.

Finally, we intend to combine these observations into a mathematically robust framework, so as to improve the prosody of the isiZulu TTS system developed in our laboratory [7].

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

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

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

[7] J.A. Louw, M. Davel, and E. Barnard. A general purpose isiZulu TTS system. In *South African Journal of African Languages (accepted for publication)*, 2005.

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

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

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

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