Fundamental frequency and tone in isiZulu: initial experiments

Natasha Govender, Etienne Barnard, Marelie Davel

Human Language Technologies Research Group
AAIICT / University of Pretoria, Pretoria, South Africa
ngovender@csir.co.za, ebarnard@csir.co.za, mdavel@csir.co.za

Abstract

Much theoretical work has been done on the tonal structure of languages in the Bantu family. However, most of these studies are not supported by physical measurements, or even a consistent model for mapping from linguistic constructs to such measurements. As a first step towards addressing this deficiency, we report on initial measurements regarding the relationship between fundamental frequency and linguistic tone in isiZulu. After choosing a suitable algorithm for pitch extraction, we have correlated a number of linguistically assigned tone values with measured values for fundamental frequency. These measurements indicate a fairly complex relationship between tone and pitch, and suggest that the commonly observed 'falling' tone in isiZulu may be a context-specific realization of the high tone.

1. Introduction

Intonation is a paradoxical aspect of human language [1]. It is universally used yet highly variable across languages; although humans naturally produce and perceive intonation as a rich channel of communication, it has to date not been a productive part of most automatic speech-processing systems. Even for well-studied languages (such as languages in the Indo-European family) much remains to be learnt. For example, the equivalent of an International Phonetic Alphabet for the unambiguous, language-independent description of intonation and other prosodic phenomena currently seems like a distant ideal, despite ongoing efforts to define such a system.

An analysis of intonation is complicated by the fact that measurable, physical quantities such as fundamental frequency, intensity, and duration depend in a complicated manner on linguistic variables such as tone, stress and quantity. Thus, the intuitive notion that tone is solely expressed in the fundamental frequency of an utterance, and stress in intensity or duration, does not hold up under closer inspection [2]. The interaction between lexical and non-lexical contributions to the intonation of an utterance further complicates the relationship between measurable and linguistic variables.

In this regard, the status of the Southern African languages in the Bantu family is quite interesting. On the one hand, intonation in these languages has attracted much attention because of its historical role in elucidation of autosegmental phonology [3]. On the other hand, these theoretical studies have not been matched by commensurate objective measurement of physical quantities, and even some basic issues on the status of tone in important languages within this group remain in dispute [4].

This leaves those who wish to develop technology for Bantu languages in a difficult situation. Whereas there is ample theoretical evidence that prosodic factors should receive significant attention in these languages, there is little by way of concrete models to guide one in this process. We have therefore embarked on a programme aimed at understanding the relationship between linguistic and physical variables of a prosodic nature in this family of languages. Our eventual aim is to produce a full account linking phonology, phonetics and objective measurements.

Given the large number of interacting variables, we approach this goal by studying closely related variables on adjacent linguistic levels. For the purposes of the current paper, we therefore investigate the relationship between the fundamental frequency (F0) and the phonetic tone levels that occur in a number of isiZulu utterances. (isiZulu is the largest family in the Nguni subfamily of the Bantu family of languages; it is also the most common first language of citizens of South Africa.)

In Section 2 below, we review some basic facts about the fundamental frequency of a speech signal, and then describe a set of experiments that was undertaken to select an appropriate algorithm for extracting F0 in isiZulu utterances. Section 3 contains our initial experimental results on the relationship between F0 and phonetic tone in isiZulu, and in Section 4 we summarize our plans to extend these initial measurements in order to develop a practically useful algorithm for the generation of F0 contours.

2. Accurately measuring fundamental frequency in isiZulu

2.1. Fundamental frequency and pitch

The fundamental frequency (F0) of a periodic signal is the inverse of it period, which in turn is defined as the smallest positive member of the set of time shifts that leave the signal invariant [5]. Speech waveforms are never absolutely periodic, so that *approximate* invariance has to be used in defining the fundamental frequency of a speech waveform. With an appropriate approximation, F0 correlates well with the subjective experience of pitch. It is therefore common practice to use the terms F0 and pitch interchangeably, and in the remainder of this paper we will do the same.

A number of algorithms have been developed to extract F0 from a speech waveform (see [6] for a review). These algorithms generally differ in the way they compute the degree of invariance in a signal, and in the ways that they use additional information (such as temporal smoothness) to adapt to the period-by-period changes that occur in speech. The development of algorithms that do this in an accurate and computationally efficient manner remains a topic of active research [7]. However, to our knowledge, these algorithms have not been evaluated formally on a Nguni language such as isiZulu. Although we do not expect that pitch extraction algorithms will differ greatly between different languages, it is worthwhile to verify this as-

sumption. In order to decide on an appropriate algorithm for our further analysis, and to test the assumption that isiZulu utterances are served well by that algorithm, we have therefore performed a number of analyses with two state-of-the-art algorithms.

2.2. Methodology

Yin [6] and the Praat [8] pitch tracker are two widely used algorithms for F0 extraction. In order to compare these algorithms, F0 was extracted from a number of spoken utterances in three different languages, namely English, French and isiZulu. In the French and English databases, each (acoustic) utterance is accompanied by a laryngograph trace. The laryngograph measures the electrical resistance between electrodes on either side of the throat, and therefore provides a fairly accurate measurement of the fundamental frequency that was actually produced by the speaker. Hence, F0 as determined from the laryngograph data is used as ground truth when comparing the algorithms on the French and English databases.

Both Yin and the Praat algorithm are characterized by a number of tunable parameters. In order to make a fair comparison, the values recommended by the algorithm developers were used for all the parameters, except where the same parameter occurred in both algorithms: these were set to reasonable and equal values. In particular, the minimum allowable pitch frequency was set to 30 Hertz, the maximum to 2000 Hertz, and a window size of 0.02s was used.

Since the laryngograph data is itself a temporal waveform, F0 has to be extracted from the laryngograph before it can be used as baseline. Fortunately, both algorithms produced very similar results (as would be expected from the highly periodic nature of laryngograph data in voiced speech) and thus either could be used as the basis for the experiments. The pitch values extracted by Yin for all the laryngograph databases was consequently used as the basis for our comparisons.

Pitch extraction algorithms can fail in a number of ways. They can fail to detect periodicity when voicing is present, or assign pitch values to unvoiced regions of speech. In voiced speech, gross errors occur when the algorithm computes a completely wrong estimate of pitch (for example, pitch halving or pitch doubling), and fine errors reflect on the detailed computation of the pitch period. In order to understand these various classes of errors, we calculated a number of measures for each of the files in our corpus:

1. The number of gross errors for each file was calculated. This was defined as the number of times that the value obtained from the laryngograph differed from the corresponding value for the acoustic file by more than a set threshold. We used a threshold of 50 Hertz.
2. We also computed the number of false positive detections of pitch (when the laryngograph did not indicate voicing, but a pitch value was extracted from the acoustic waveform) and, conversely, the number of false negative detections.
3. The mean square error was calculated only across those pitch periods where both the laryngograph data and the acoustic data indicated the presence of voicing, and where no gross error occurred.

Since no laryngograph data was available for the isiZulu database, we computed the number of gross differences between the two methods (rather than the number of gross errors), and also computed the mean squared difference between the answers produced by the two algorithms. Finally, a manual process was used to decide which of the two algorithms was in error when gross differences occurred. That is, a random selection of files was made. Each file was manually inspected at the points were the fundamental frequency extracted by the two algorithms differed by more than the threshold value. At these points, the period (and hence the pitch) was calculated manually to decide which of the algorithms is in error.

2.3. Databases

Four databases were used in this study. These comprise a total of 1.16 hours of speech. The first three included a laryngograph waveform recorded together with the speech.

- DB1: Two male speakers of English produced a total 0.2 hours of speech.
- DB2: One male pronounced 150 English sentences for a total of 0.17 hours of speech. The database is available with the laryngograph data from http://www.festvox.org/examples/cstr_us_ked_timit.
- DB3: Two male and two female speakers each pronounced between 42 and 55 French sentences for a total of 0.46 hours of speech.
- DB4: An adult male whose first language is isiZulu produced the isiZulu voice recordings. He pronounced 150 sentences with a total of 0.33 hours of speech.

2.4. Algorithms

The compared algorithms (Yin and the Praat tracker) are briefly described below.

- *Yin* is an implementation of the method developed by De Cheveigne [6]; it combines autocorrelation and Average Magnitude Difference Function(AMDF) methods [9] with a set of modifications and post-processes that reduce common errors of those algorithms.
- The *Praat* pitch tracker performs an acoustic periodicity detection on the basis of an accurate autocorrelation method, as described in Boersma [10]. This method tends to be more accurate, noise-resistant, and robust, than methods based on cepstrum or combs, or the original autocorrelation methods. In order to estimate a signal's short term autocorrelation function on the basis of a windowed signal, this method divides the autocorrelation function of the windowed signal by the autocorrelation function of the window. It is available with the Praat toolkit at http://www.fon.hum.uva.nl/praat/.

2.5. Results

2.5.1. Gross Errors

The average number of gross errors[1] measured for the English and French databases, across all files, as well as the number of gross errors manually measured for each on the isiZulu database are reported in Table 1. Across all three languages, the Praat algorithm tends to make fewer gross errors (possibly because of the more sophisticated post-processing done by Praat as part of its tracking algorithm). Alternatively, these differences may be a consequence of the relatively conservative voicing detection algorithm used by Praat (see below).

[1] Note that the number of errors is not comparable across databases, as this number is correlated with utterance length

Table 1: *Mean number of gross errors per utterance for Praat and Yin across all databases, as computed from a comparison with laryngograph data(English or French) or manual inspection(isiZulu)*

Database	Praat	Yin
English DB1	3.868	12.181
English DB2	0.227	10.267
French	49.674	65.873
isiZulu	0.8	1.3

2.5.2. *Errors in the detection of voicing*

Tables 2 and 3 contain the average number of false positive and false negative detections of voicing, respectively, for the various databases. These results indicate that the two algorithms have different thresholds for voicing detection - Praat makes fewer positive errors, at the cost of additional missed detections.

Table 2: *The average number of false positive voicing detections per utterance*

Database	Praat	Yin
English DB1	0.0533	26.68
English DB2	0.2828	34.101
French	17.699	65.650

Table 3: *The average number of false negative voicing detections per utterance*

Database	Praat	Yin
English DB1	75.393	10.919
English DB2	38.727	4.147
French	63.843	15.789

2.5.3. *Mean Square Error*

Table 4 contains the mean square errors obtained for the English and French databases, expressed as a percentage of the measured F0 values. Both algorithms are highly accurate, with the Praat algorithm consistently more accurate than Yin. (The values reported in Table 4 are very close to those obtained in [6]; the small observed differences are most likely the result of differences in our experimental protocols.) As with the gross errors, the relative superiority of Praat may either be the result of intrinsic algorithmic factors, or the more conservative voicing detection used in Praat.

The mean squared difference between the values obtained with the two algorithms on the isiZulu database (for which we did not have a laryngograph-derived baseline) was 0.115%. This difference is somewhat smaller than would be expected from the values in Table 4, but broadly in line with those values.

2.6. Conclusion: algorithms for the determination of F0

Both Yin and the Praat pitch tracker perform very well on the databases studied here; however, the Praat algorithm performs somewhat better than Yin in terms of gross and fine error. The main negative aspect of the Praat algorithm is that it is more prone to missing frames in which voicing was actually present. This disadvantage may weigh heavily in applications such as speech recognition, but is relatively unimportant for our purposes of analyzing the relationship between F0 and tone. Praat will therefore be used in the rest of our work. Also, the numerical results reported above, as well as our subjective inspection of the computed values, confirm that the performance on isiZulu data is very comparable to that on the other two languages. This gives us confidence that the algorithm will perform well on our isiZulu data.

Table 4: *The average mean squared error of both algorithms when compared with laryngograph measurements*

Database	% Mean Squared Error	
	Praat	Yin
English DB1	0.193	1.819
English DB2	0.081	1.884
French	0.387	1.076

3. The relationship between pitch and tone: Initial measurements for an isiZulu speaker

3.1. Methodology

For this part of the experiment, a first-language isiZulu speaker investigated the transcriptions of the isiZulu speech recorded in DB4, and manually marked each of the syllables in each word as *high* or *low*. In order not to bias this process, the marking was done without listening to the voice recordings; the transcriber was asked to mark each syllable as he would produce it *in the sentence context provided*.

All utterances were segmented at the phonemic level, and labelled. For each syllable, the initial and final F0 values were calculated with Praat. These values were then subjected to a number of analyses, in order to arrive at an understanding of the relationship between the measured pitch values and the tone predicted by the transcriber.

An example of the extracted the pitch contour and the segmented phoneme boundaries is shown in Figure 1. The nuclear vowel of each syllable is marked H (high) or L(low).

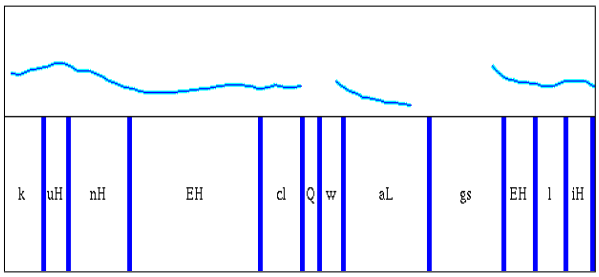

Figure 1: *A portion of a signal extracted from an isiZulu speech recording*

3.2. Results

As expected, the mapping between transcribed tone and measured F0 is fairly complex, and not simply a matter of F0 being

large for high tones and small for low tones. The most salient observations from our initial analysis were the following:

- The strongest overall determinant of the pitch values in a segment is the position of the syllable in the sentence, because of the general tendency of pitch to decline throughout an utterance (as is the case in many of the languages of the world [1]). This trend can be seen in Figure 1, and average F0 values of syllables marked as high are plotted as a function of the position of the syllable in the utterance in Figure 2.

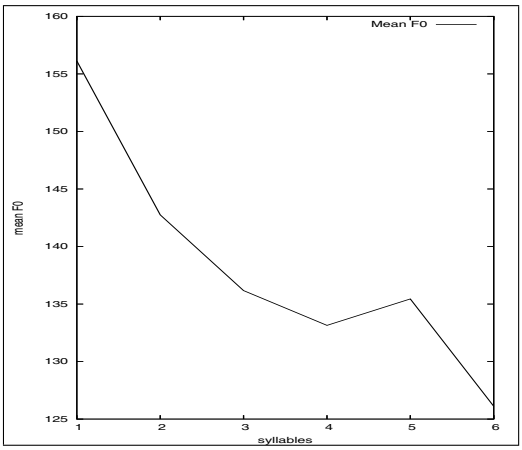

Figure 2: *Mean F0 for the first 6 syllables in a sentence*

- Relative to this trend line, the syllables marked as high do indeed have larger F0 values on average (in our corpus, the F0 of syllables marked as 'H' is, on average, 8% higher than that of 'L' syllables at the same location in an utterance).
- If the tone pattern *HHL* occurs in three consecutive syllables, the second of these syllables is fairly consistently produced with falling pitch. If 'falling' is defined as a syllable in which pitch decreases by at least 10% during the syllable nucleus, we observe a falling pitch in 78% of 'H' syllables preceded by an 'H' and followed by an 'L', whereas fewer than 30% of the 'H' syllables in other contexts have falling pitch.

4. Conclusion and future work

We have found that both Yin and the Praat pitch tracking algorithm are highly accurate over several languages – including isiZulu, which is our primary focus. The Praat algorithm is a little more accurate (though also a little more conservative in detecting voicing), and was used in our analysis.

Our initial exploration of isiZulu suggested a number of regularities. Potentially the most significant finding is the suggestion that the 'falling' pitch contour is a context-specific realization of a high tone; this would resolve some of the uncertainty surrounding the status of the falling tone in the 'tonology' of isiZulu [11]. We are in the process of analyzing a larger corpus of utterances to further investigate this and related issues; in order to construct a model of intonation, we would also like to relate the observed tone levels to predictable quantities of a lexical and supra-lexical nature. Other topics under investigation include an analysis of the inter-speaker variability of these observations, and also comparisons to closely related languages such as isiXhosa, and more distantly related languages such as Sepedi.

5. Acknowledgements

This work was supported by the African Advanced Institute for Information and Communications Technologies (AAIICT). We would like the thank the authors of the various databases used in the experiment: Nathalie Henrich for the French database and Nick Campbell for the English database (DB1).

6. References

[1] D. Hirst and A. D. Cristo, *Intonation Systems*. Cambridge University Press, 1998.

[2] D. B. Fry, "Experiments in the perception of stress," *Language and Speech*, pp. 120–152, 1958.

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

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

[5] A. de Cheveigne and H. Kawahara, "Comparative evaluation of f0 estimation algorithms," in *EuroSpeech*, pp. 2459–2462, 2001.

[6] A. de Cheveigne and H. Kawahara, "Yin, a fundamental freqency estimator for speech and music," in *Journal of Acoustical Society of America*, pp. 1917–1930, 2002.

[7] X. Li, J. Malkin, and J. Bilmes, "Graphical model approach to pitch tracking," *Interspeech:8th International Conference on Spoken Language Processing*, pp. 1101–1104, 2004.

[8] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, pp. 341–345, 2001.

[9] E. Barnard, R. Cole, M. Vea, and F. Alleva, "Pitch detection with a neural-net classifier," in *IEEE Transaction on Signal Processing*, 1991.

[10] P. 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*, pp. 97–110, 1993.

[11] G. Poulos and C. T. Msimang, *A Linguistic Anaylsis of Zulu*. Via Afrika, 1998.