Pitch Modelling for the Nguni Languages

Natasha Govender, Etienne Barnard, Marelie Davel

Human Language Technologies Research Group
Meraka Institute / University of Pretoria, Pretoria, South Africa

ABSTRACT

Although the complexity of prosody is widely recognised, the lack of widely-accepted descriptive standards for prosodic phenomena has meant that prosodic systems for most of the languages of the world have, at best, been described in impressionistic rule-based terms. For the languages of Southern Africa, the deficiencies in our modelling capabilities are acute. Little work of a quantitative nature has been published for the languages of the Nguni family (such as isiZulu and isiXhosa), and there are significant contradictions and imprecisions in the literature on this topic, which partially stems from the lack of quantitative, measurement-driven analysis.

This paper therefore embarks on a programme aimed at understanding the relationship between linguistic and physical variables of a prosodic nature in this family of languages. Firstly we undertake a set of experiments to select an appropriate pitch tracking algorithm for the the Nguni family of languages. We then use this pitch tracking algorithm to extract relevant data from speech recordings to build intonation corpora for isiZulu and isiXhosa. Using the extracted data in the intonation corpus, we show that it is possible to develop fairly accurate intonation models using a neural network classifier for isiZulu and isiXhosa.

KEYWORDS: prosody, Nguni languages, fundamental frequency, intensity, intonation corpus, intonation modelling, pitch tracking, autocorrelation, classification, tone

1 INTRODUCTION

Prosody is a paradoxical aspect of human language. It is universally used yet highly variable across languages. Every language possesses prosody and many of the linguistic and paralinguistic functions of prosody systems seem to be shared by languages of widely different origins [1]. Despite the universal character of prosody, the specific features of a particular speaker's prosody system depend strongly on the language, the dialect, and even the style, the mood and the attitude of the speaker.

In the literature, a variety of different meanings have been associated with the term 'intonation'. We use the term in its broad sense, to refer to the *melodic pattern of an utterance*, either occurring at word level (lexical intonation) or over larger sections of an utterance (supralexical or syntactic intonation). This 'pattern' represents the non-phonetic content of speech, and includes perceptual characteristics such as *tone*, *stress* and *rhythm.* A basic distinction is made between the perceptual attributes of sound, especially a speech sound, and the measurable physical properties that characterise it. These perceptual or abstract characteristics correspond to physical measurements such as fundamental frequency, intensity and duration in an often complex manner. Prosody is achieved by varying the levels of pitch, intensity and duration in the voice. An overview of intonation as observed in a variety of languages is provided in [1].

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 prosody of an utterance further complicates the relationship between measurable and linguistic variables.

Attempting to create an prosody model for any language is a complex task. This difficulty is exacerbated by the fact that there is little agreement about appropriate descriptive frameworks for modelling . The lack of widely-accepted descriptive standards for prosodic phenomena which can be used to describe all languages, has 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 [3]. 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.

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 the elucidation of autosegmental phonology [4] and

Email: Natasha Govender ngovender@csir.co.za, Etienne Barnard ebarnard@csir.co.za, Marelie Davel mdavel@csir.co.za

its intricate tonal structure. On the other hand, little work of a quantitative nature has been published, and as Roux [5] 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.

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. For these Southern African languages, the deficiencies in our modelling capabilities are acute.

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. We then use the information/knowledge gathered to build intonation models for isiZulu (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) and isiXhosa. isiZulu and isiXhosa are considered to be tonal languages i.e. a language in which pitch variations are used to indicate differences in meaning between words otherwise similar in sound.

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 pitch tracking algorithm for extracting fundamental frequency(F0) from isiZulu utterances. In Section 3 we describe the methodology used for developing a general-purpose intonation corpus and the various methods implemented to extract relevant features such as fundamental frequency, intensity and duration from the spoken utterances of these languages. In Section 4, in order to understand how the 'expected' intonation relates to the actual measured characteristics, we develop a neural network classifier to predict the tone for the isiZulu and isiXhosa utterances. In Section 5 we discuss our final conclusions from our experiments.

2 PITCH TRACKING ALGORITHMS

A number of pitch tracking algorithms have been developed; however, to our knowledge, these algorithms have not been evaluated formally on a Nguni language such as isiZulu. Although the expectation is that pitch extraction algorithms will not differ greatly between different languages, it is worthwhile to verify this assumption. In order to decide on an appropriate algorithm for further analysis, and to test the assumption that isiZulu utterances are served well by that algorithm, a number of analyses have been performed with two state-of-the-art algorithms namely the Praat pitch tracker [6] and Yin [7].

In Section 2.1 we explain some basic facts about the fundamental frequency of a speech signal. In Section 2.2 we define the methodology undertaken to select an appropriate algorithm for extracting fundamental frequency from isiZulu utterances. We also describe the various databases and algorithms used in the experiments. In Section 2.3 we display the results obtained from the experiments, and in Section 2.4 we summarise our conclusions from these experiments.

2.1 Fundamental Frequency

The fundamental frequency (F0) of a periodic signal is the inverse of it's period, which in turn is defined as the smallest positive member of the set of time shifts that leave the signal invariant [8]. 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.

2.2 Methodology

Yin [7] and the Praat [6] pitch tracker are two widely used algorithms for F0 extraction. These algorithms (Yin and the Praat tracker) are briefly described below.

- *Yin* is an implementation of the method developed by De Cheveigne [7]; 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 by Boersma [10]. This method tends to be more accurate, noise-resistant, and robust, than methods based on cepstra 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 [11].

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 and each file was manually inspected at the points where 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.

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 [12].
- 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 [13].
- DB3: Two male and two female speakers each pronounced between 42 and 55 French sentences for a total of 0.46 hours of speech [14].
- 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.

2.3 Results

We next present results on the suitability of the various pitch tracking algorithms for our purposes. Since we did not have laryngograph baseline values for DB3 and DB4, most of the results in this subsection are for DB1 and DB2 only.

2.3.1 Gross Errors

The average number of gross errors[1] per utterance was measured for the English and French databases, across all files, as well as the number of gross errors manually measured for each utterance in 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).

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

2.3.2 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.

2.3.3 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

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

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

in Table 4 are very close to those obtained in [7]; 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.

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

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.

2.4 Conclusion

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.

In the next section we discuss how the selected pitch tracker was used in the development of the intonation corpus.

3 DEVELOPING INTONATION CORPORA FOR ISIXHOSA AND ISIZULU

Attempting to create an intonation model for any language is a complex task. For South African languages this task is further complicated by the lack of intonation resources available. While intonation corpora exist for more researched languages such as French and English, there are no such corpora available for South African languages.

In Section 3.1 we describe the methodology used for developing a general-purpose intonation corpus. In Section 3.2 we describe the corpora developed and the various kinds of information that can be extracted. In Section 3.3 we report on some of the global measurements related to F0 that were extracted from our corpus; the more localised measurements, which are the main focus of this research, are described in subsequent sections.

3.1 Methodology

Our aim was to develop an annotated intonation corpus that will support further statistical research in intonation modelling. Corpus development was not guided by specific linguistic hypotheses (although the testing of such hypotheses is certainly supported by these corpora, as we describe in the rest of this paper), but rather was aimed at collecting natural read speech from a number of speakers, and annotating this data in ways that are meaningful from a pattern recognition perspective. The methodology used for building such corpora for two Nguni languages (isiZulu and isiXhosa) is described in detail below, illustrating the process from initially building the corpus of sentences, generating the voice recordings and tone markings, to extracting the fundamental frequency (F0), intensity and duration values.

3.1.1 Collection of Text Corpora

The first step consisted of the selection of an appropriate text corpus for recording purposes. Initially a large collection of text sentences was obtained from various isiXhosa and isiZulu websites. In total 2300 isiXhosa and 1700 isiZulu sentences were collected. These sentences were then verified as logically and grammatically correct by first language speakers of the respective languages.

From this larger corpus, we aimed to select those sentences that would provide the most value in terms of varying tone levels. Based on the assumption that a large variation in graphemic bigrams would result in a large variation of intonation phenomena, a subset of sentences was selected that provided large graphemic variability. This was done using a text optimiser that applies a greedy search algorithm, which selects each successive sentence as the sentence which adds the greatest set of additional bigrams to the pool of covered bigrams. The algorithm was initialised based on graphemic bigram frequencies occurring in the larger text corpus, as illustrated in Table 5.

isiXhosa bigram frequencies
d 12
j 6
r 1
h-i 84
m-i 57
y 91
p 16

Table 5: Examples of bigram frequency counts

For isiXhosa 53 of the original 2300 sentences were selected. For isiZulu 153 of the original 1700 sentences were selected for recording.

3.1.2 Recording of Sentences

The sentences selected by the text optimiser were recorded by one first language isiXhosa male and one female speaker and the isiZulu sentences by one first language isiZulu male and one female speaker, in a quiet office environment. All recordings were obtained at a recording rate of 16Khz, using the open source Audacity toolkit on a laptop computer, and a close-talking microphone.

3.1.3 Marking of Sentences

Tones can be understood as labels for perceptually salient levels or movements of F0 on syllables. Pitch levels and movements on accented and phrase-boundary syllables can exhibit a bewildering diversity, based on the speaker's characteristics, the nature of the speech event, and the utterance itself. For modelling purposes, it is useful to have an inventory of basic, abstract pitch types that could in principle serve as the base inventory for expression of linguistically significant contrast.

In order to understand how the 'expected' intonation relates to the actual measured characteristics, the syllabic intonation was 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. These markings were performed by a first language isiXhosa speaker for the isiXhosa sentences and a first language isiZulu speaker for the isiZulu sentences. Note that different speakers were used than during the recordings, i.e. these markings were independent of the recorded audio data.

For each sentence the boundaries of every syllable were marked and transcribed using Praat [6]. An example of the syllable markings for a portion of an isiZulu sentence in Praat is illustrated in Figure 1.

3.1.4 Extraction of Intonation Measurements

- **Fundamental Frequency**

  In Section 2.3 it was shown that Praat's implementation of a pitch tracking algorithm produced the best results for the studied languages [15]. Thus, this algorithm was selected to extract the pitch values from the isiXhosa and isiZulu voice recordings.

  The most fundamental distinction between sound types in speech is the voiced/unvoiced distinction. Voiced sounds including vowels, have in their time and frequency structure a roughly regular pattern that unvoiced sounds, such as constants like *s*, lack.

  The fundamental frequency (F0) values were extracted at the syllable boundaries, i.e. they were extracted at the start and end of each syllable in the sentence. However, the fact that unvoiced segments often occur at the beginning of a syllable meant that a large percentage of the values extracted for both isiXhosa and isiZulu were not defined in this way, and hence a large number of the pitch values extracted were undefined.

  In order to rectify this problem, two different approaches were implemented and the more accurate of these was selected. The two approaches implemented were:

  – MOMEL (MOdelisation de MELodie) [16] was used to obtain a smoothed contour of the F0 values. MOMEL is an algorithm for the automatic modelling of fundamental frequency curves, factoring them into a macroprosodic and a microprosodic component. The macroprosodic component is modelled as a quadratic

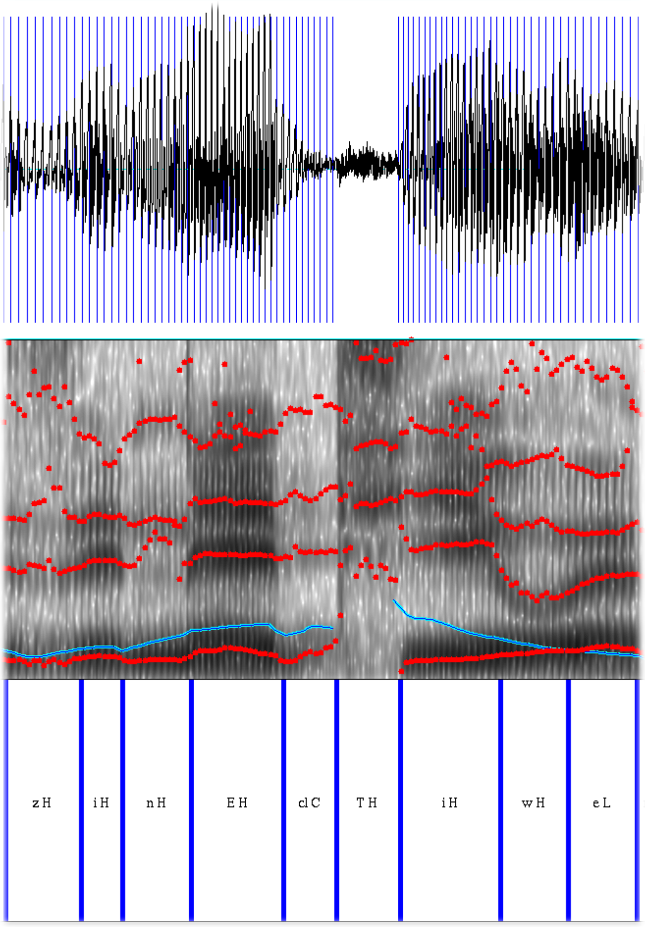

Figure 1: A portion of a signal extracted for an isiZulu sentence and the pitch contour

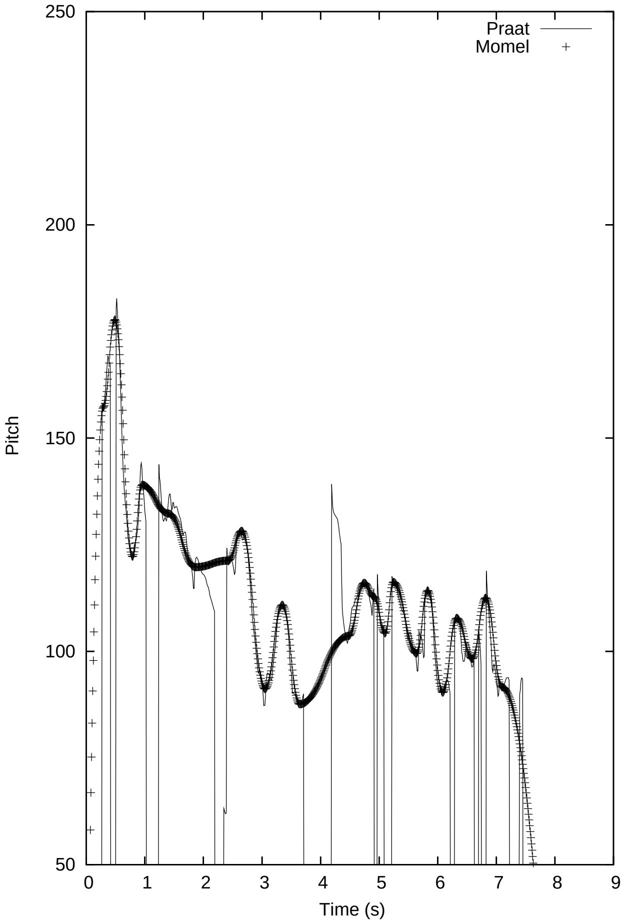

Figure 2: A graph showing the difference in the pitch values obtained between Praat and MOMEL

spline function i.e a continuous smooth sequence of segments of parabolas defined by a sequence of target points corresponding to points where the first derivative of the curve is nil.

The F0 for each recording was extracted at every 10 milliseconds, and MOMEL used to generate an interpolated F0 contour. The boundary times (i.e the starting time and ending time) for each syllable were then compared to the output and the corresponding F0 value extracted. This process is illustrated in Figure 2. Figure 2 displays the waveform, the spectrogram and the phone boundaries respectively. Within each phone boundary, the perceived tone of the phone appears as either high (H) or low (L), along with the actual phone. Note how the 'undefined' values provided by Praat (zero values in the figure) have been removed in the MOMEL contour.

– The second approach, referred to as the Non-Zero method, extracts the first non-zero pitch value and the last non-zero pitch value for each syllable in a sentence. The first non-zero pitch value extracted is then used as the starting pitch and the last non-zero pitch value extracted as the ending pitch for that particular syllable. This is illustrated in Figure 3. In this figure, *1* denotes the point of the last non-zero pitch for syllable *ne* and *2* the point of the first non-zero pitch to be extracted for the next syllable *e*. A more sophisticated smoothing strategy could be employed here, but to limit the set of variables under consideration we have not investigtated this option

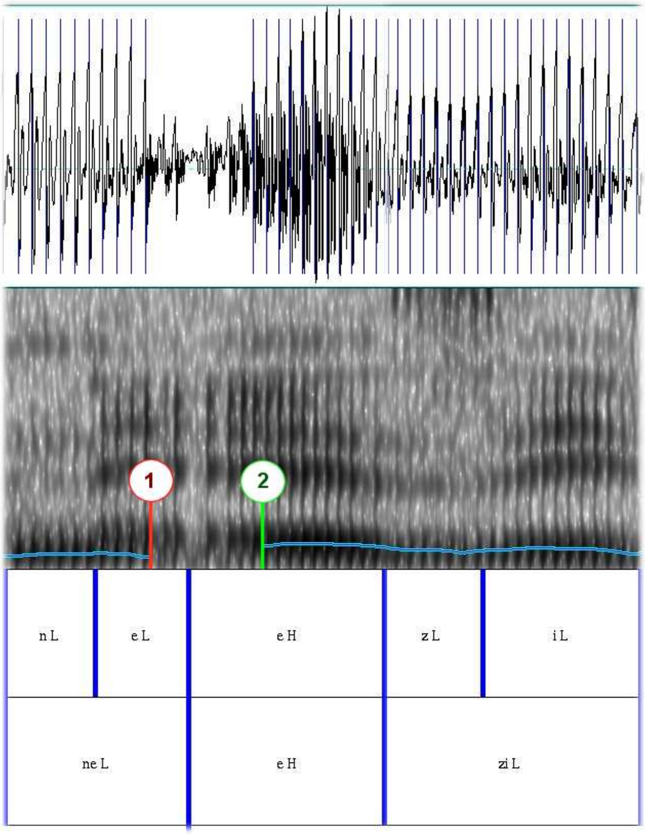

Figure 3: A portion of an isiZulu signal depicting the points at which the pitch values would be extracted using the non-zero method

From our experiments the Non-Zero method was proven to obtain more accurate results.

- **Intensity**
  The intensity was calculated at each of the syllable boundaries, as the average squared value of the signal within a 5 millisecond window.
- **Duration**
  To calculate duration of each syllable, the starting and ending times of the syllable were obtained from the hand labels, and subtracted.

3.2 Results

At this point the information contained in the intonation corpus includes:

- the actual voice recordings, grouped per speaker,
- the orthographic transcription per voice recording,
- syllabification markings,
- the expected High/Low markings for each syllable,
- the pitch values extracted using the Non-Zero method
- the extracted intensity values, and
- the extracted duration values.

Table 6 illustrates a typical example of the pitch values obtained using the Non-Zero method for an isiZulu sentence. The intensity and duration values extracted would be the same for both algorithms.

Segment	H/L	start F0	end F0	Intensity	sec
i	H	158.24	167.84	86.23	0.13
si	L	167.84	129.21	86.44	0.31
mo	L	129.21	132.95	84.93	0.30
so	L	143.82	131.37	82.21	0.23
ku	L	131.37	136.04	81.41	0.90

Table 6: An example of an annotated isiZulu data item

3.3 Observations

In Section4 we discuss the development of prosodic models using the developed corpora. In the remainder of this section we list some global measurements observed from the extracted data.

3.3.1 Declination in F0

In many of the languages of the world, F0 has a consistent tendency to decline within a phrase [1]. However, the extent of this declination varies significantly between different languages, for different speaking styles, and possibly also depends on factors such as the gender and age of the speaker.

We investigated the magnitude of this effect for our languages and speakers, by computing the average values across all utterances (in increments of 25 milliseconds), as a function of the duration from the beginning of the utterances. These averages are shown in Figure 4 for the two isiZulu speakers, and in Figure 5 for the two isiXhosa speakers.

We see that similar declinations occur in both languages, and that these declinations do not seem to differ systematically by speaker gender.

3.3.2 Pitch variability and speaker gender

The fact that F0 is generally higher for females than males is a simple consequence of anatomical tendencies; however, there are also gender differences in the production of prosody that are cultural in origin. Our subjective experience is that the extent of pitch variation is such a difference in the Nguni (and related) languages – specifically, we hypothesise that female speakers tend to produce wider variability in F0 than males.

In order to test this hypothesis, we define the *pitch variance* of a spoken utterance as the variance of the F0 values (as interpolated by MOMEL) observed when the utterance is sampled at 250 millisecond increments. The results are shown in Table 7: for our limited set of isiZulu speakers, the hypothesis is indeed confirmed, but for the limited set of isiXhosa speakers the same hypothesis does not hold.

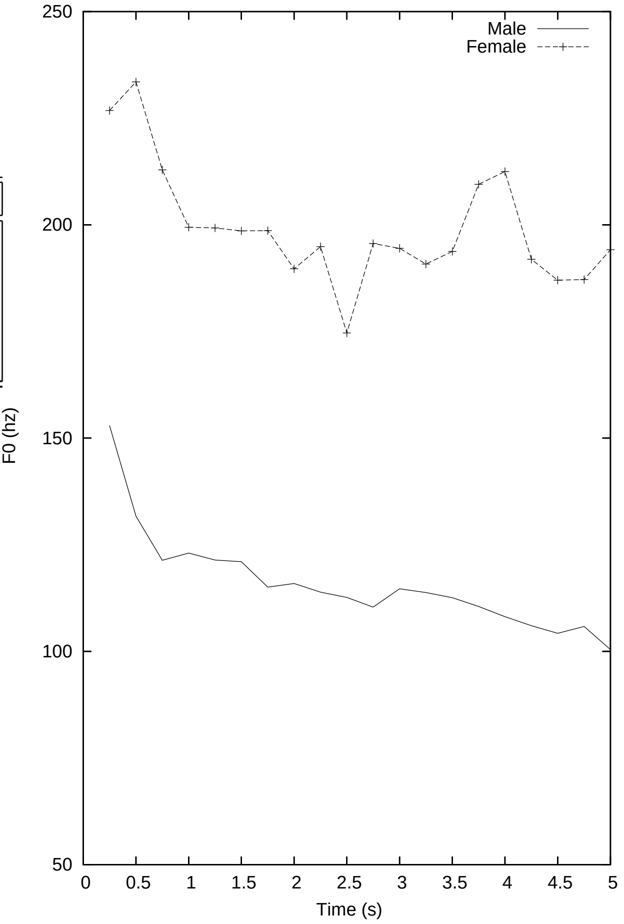

Figure 4: Mean pitch as a function of the time since the start of an utterance, for isiZulu

Language	Male mean	Male variance	Female mean	Female variance
isiZulu	117.10	21.60	203.80	33.70
isiXhosa	122.90	38.30	197.00	36.00

Table 7: Average pitch variance values for male and female speakers

3.4 Conclusion

We have motivated the need for intonation corpora in order to model spoken languages, and described a general approach to the development of such corpora. For the case of isiZulu and isiXhosa, we have developed limited corpora, consisting of one male speaker and one female speaker in each language. By applying standard tools from the field of pattern recognition – preprocessing, feature extraction, computation of statistical tendencies – it is possible to learn much from such corpora.

Our corpora are intended as a resource for various tasks, such as the development of models that relate tone to F0 (which is important for applications in speech recognition and speech synthesis). We have investigated a number of global characteristics of F0 that can be inferred from these corpora. In particular, we have seen that similar rates of pitch declination are observed in both isiZulu and isiXhosa for both genders. Also, female pitch values tend to be

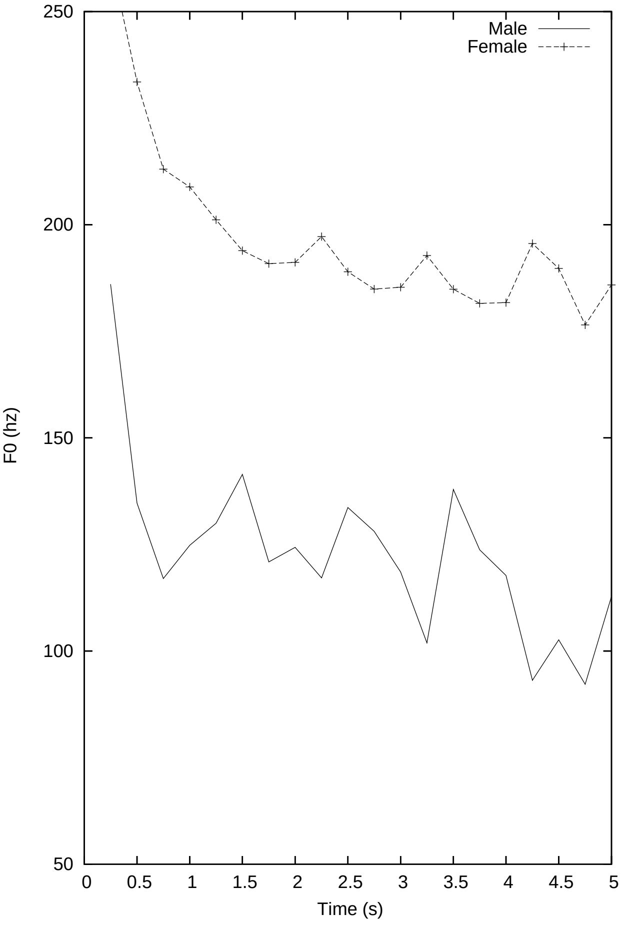

Figure 5: Mean pitch as a function of the time since the start of an utterance, for isiXhosa

more variable than those of males in one language but not the other. In the next section we utilise the newly developed corpora to investigate prosody from a computational perspective.

4 COMPUTATIONAL MODELS OF PROSODY

Thus far 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, we have developed statistical methods to build intonation models for isiZulu and isiXhosa. We choose to build a neural network classifier for our intonation model.

4.1 Introduction

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.

Data input into a classifier is required to be separable for the classifier to be able to learn and classify effectively. To determine if the pitch and intensity values extracted conformed to this expectation, scatter plots were produced for the various segment combinations which are described in Table 8.

Original State	Next State	Designation
High	High	HH
	Low	HL
Low	High	LH
	Low	LL

Table 8: Segment Tone Combinations

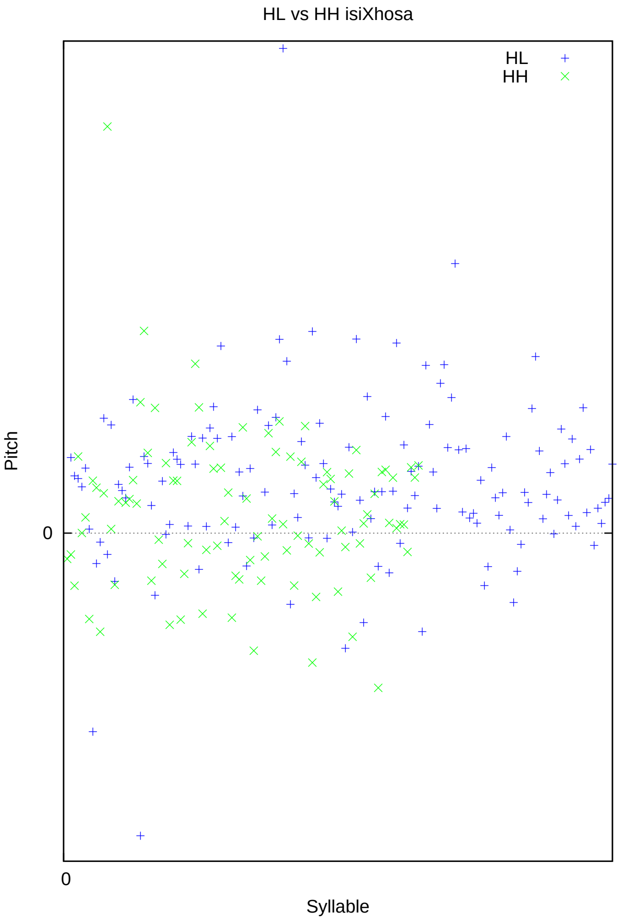

Figure 6: HL consecutive segments plotted against HH consecutive segments using only the difference between consecutive pitch values for isiXhosa

The average pitch value of each segment was used. For each combination above, the differences between consecutive pitch values were calculated and plotted. Figure 6 displays the results of a 'HH' combination plotted against a 'HL' combination and Figure 7 displays the results of a 'LL' combination plotted against a 'LH' combination for the isiXhosa corpus using only pitch values. Using all possible 'H' and 'L' combinations would have produced interesting results but was beyond the scope and purpose of this investigation.

From the graph, we can deduce that there is a reasonable degree of separability between the two combinations based on pitch alone.

Scatter plots were also produced for each segment combination using the difference between consecutive pitch values and consecutive intensity values. The intensity value used for each syllable, was the high-

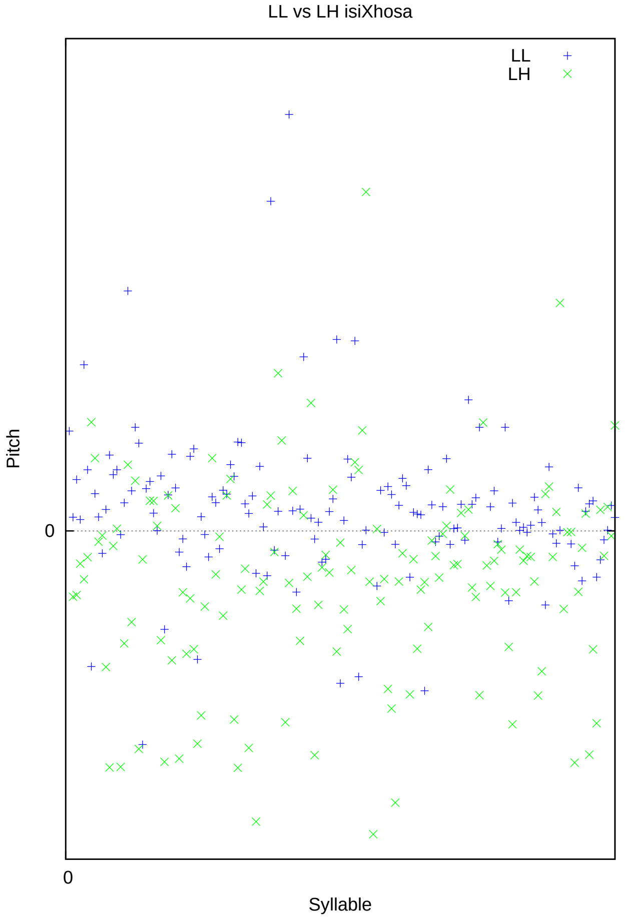

Figure 7: LL consecutive segments plotted against LH consecutive segments using only the difference between consecutive pitch values for isiXhosa

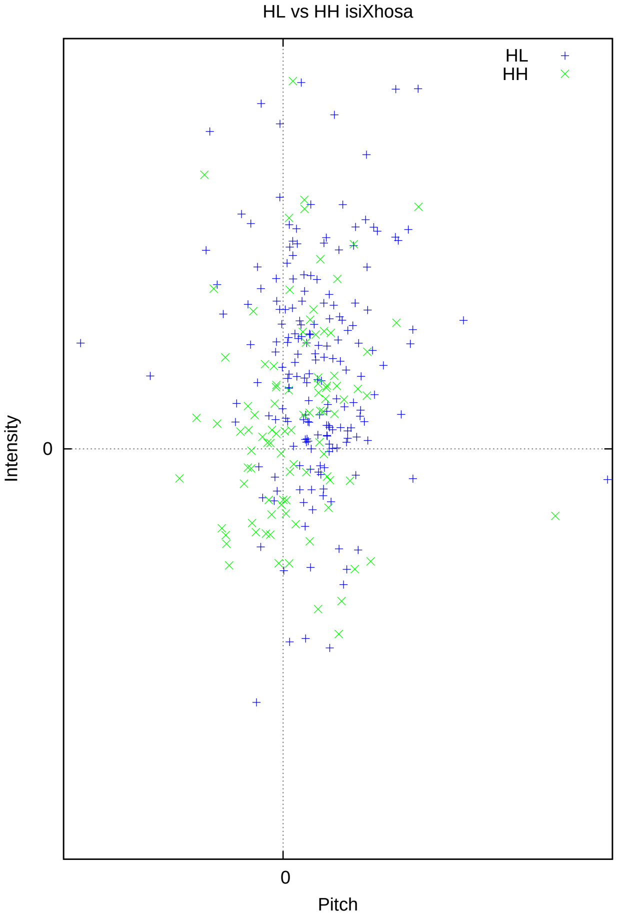

Figure 8: HL consecutive segments plotted against HH consecutive segments using the pitch and intensity values extracted for isiXhosa

est intensity value extracted for that syllable. Figure 8 displays the results of a ‘HL’ combination plotted against a ‘HH’ combination and Figure 9 displays the results of a ‘LL’ combination plotted against a ‘LH’ combination for the isiXhosa corpus using both pitch and intensity values.

We can conclude from Figure 8 and Figure 9 that pitch values combined with the intensity values are also reasonably separable and should work well as input to a classifier.

In Section 4.2 we describe the classifier used in the experiments. In Section 4.3 we describe the methodology implemented to build the intonation model and the results obtained from the classifier.

4.2 Neural network classifier

A neural network classifier was selected to build our intonation model. Neural networks have found application in a wide variety of problems. These range from function representation to pattern recognition, which is what we consider here.

A neural network consists of units (neurons), arranged in layers, which convert an input vector into some output. Each unit takes an input, applies a (possibly nonlinear) function to it and then passes the output on to the next layer. Generally the networks are defined to be feed-forward: a unit feeds its output to all the units on the next layer, but there is no feedback to the previous layer. Weightings are applied to the signals passing from one unit to another, and it is these weightings which are tuned in the training phase to adapt a neural network to the particular problem at hand.

Neural networks learn by example. The neural network gathers representative data, and then invokes training algorithms to automatically learn the structure of the data which is essential to building our model.

4.3 Methodology

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 9, 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).

The features extracted for each segment as input into the classifier are:

- starting pitch
- ending pitch
- average pitch

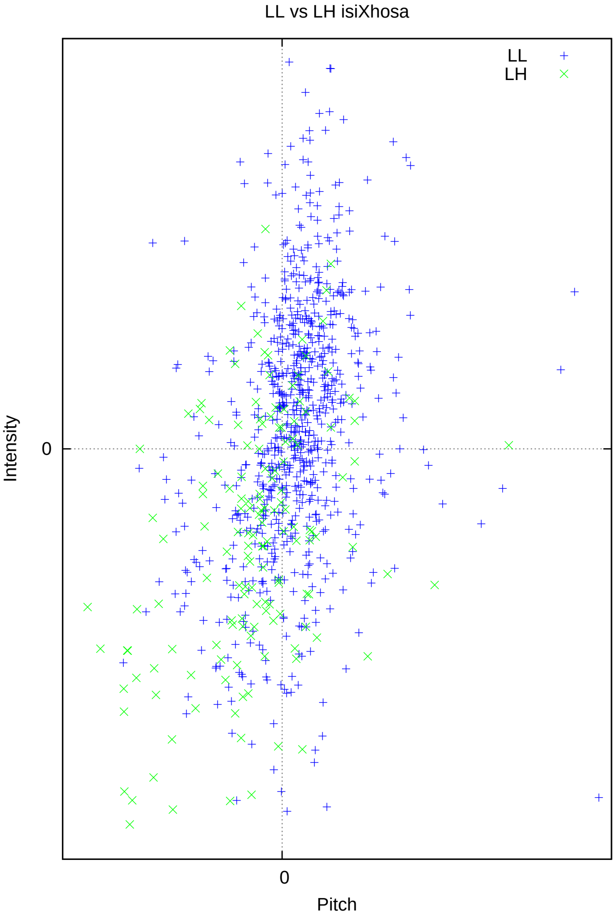

Figure 9: LL consecutive segments plotted against LH consecutive segments using the pitch and intensity values extarcted for isiXhosa

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

Table 9: The number of utterances and syllables used for the training and testing of the classifiers.

- difference between the starting and ending pitches values
- starting intensity
- ending intensity
- highest intensity value (within a segment)
- difference between the starting and ending intensity values

Initially we experimented with the number of hidden neurons for the neural network to determine which produced optimal results for our type of input. For the isiZulu database 16 hidden neurons produced the most accurate results for both sets of data (pitch and intensity) and for the isiXhosa database 10 hidden neurons were found to produce the most accurate results.

These values were then used as parameters in our classifier. With so many features extracted, it was important to determine which feature/s contributed the most to improving the accuracy of the classifier. For the initial experiment, each feature was individually trained and tested using the classifier. The features were then ranked according to their accuracy.

The results for the isiZulu and isiXhosa databases are displayed in Table 10 and Table 11 respectively. The 'High' and 'Low' columns indicates the accuracy of the two types of classifiers trained, depending on whether the previous state had been an 'H' or an 'L'. There were a larger number of 'HL' segment combinations than 'HH' segment combinations (approximately three times as many). To prevent any bias in the classification, we boosted the number of 'HH' segment combinations to test if the classifier was learning or simply guessing. In the latter case the results would not be better than chance which is 50%.

isiZulu		
Feature	% High	%Low
difference in intensity	70.53	70.96
difference in pitch	68.03	70.96
average pitch	63.01	70.96
ending pitch	57.99	68.71
ending intensity	57.99	68.30
starting intensity	57.99	68.10
starting pitch	56.40	67.89
highest intensity	56.40	58.90

Table 10: Classifier results for each individual feature for the isiZulu database

isiXhosa		
Feature	%High	%Low
ending pitch	85.25	83.00
starting intensity	83.61	83.00
starting pitch	83.60	83.23
difference in pitch	83.60	82.40
ending intensity	82.47	82.19
average pitch	81.97	81.78
difference intensity	73.49	80.23
highest intensity	70.49	74.03

Table 11: Classifier results for each individual feature for the isiXhosa database

Individually each feature does produce good results using the classifier. We then combined the first two features for each database to train and test on the classifier to determine if the combination would produce better results. Thereafter we added each feature on the list to the previous combination and so forth, finally using all eight features. The results are displayed for isiZulu in Table 12 and for isiXhosa in Table 13.

As shown in Table 14, we compared the classification accuracies achievable with the F0-derived features to those of the amplitude derived features.

isiZulu		
Combination	%High	%Low
(1) difference in pitch + intensity	71.47	74.23
(2) (1) + average pitch	73.35	75.87
(3) (2) + end pitch	73.67	77.3
(4) (3) + end intensity	75.24	77.71
(5) (4) + start intensity	75.64	77.71
(6) (5) + start pitch	75.64	78.21
(7) All features	77.74	78.32

Table 12: Classifier results for combination of features for the isiZulu database

isiXhosa		
Combination	%High	%Low
(1) end pitch + start intensity	85.25	84.62
(2) (1) + start pitch	85.25	86.23
(3) (2) + difference in pitch	85.7	86.23
(4) (3) + end intensity	85.7	86.23
(5) (4) + average pitch	85.7	86.62
(6) (5) + difference in intensity	81.97	86.23
(7) All features	86.89	86.32

Table 13: Classifier results for combination of features for the isiXhosa database

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.

4.4 Conclusion

From the intonation model built we can deduce that the F0-based features and the amplitude-based features produce comparable accuracy. This lends independent support to the hypothesis advanced in [5] 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. Both F0 and amplitude produce good results and a combination of the two features produces only a slight improvement on the individual results. 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 among these possibilities.

The neural network classifier, which uses eight extracted features, produces good prediction results. The classification model is also robust and can easily learn from the training data but the eight features also need to be extracted. We have demonstrated that it is possible to build a fairly good intonation model for these languages using a classifier.

isiZulu		
	% High	%Low
Pitch	67.71	74.44
Intensity	71.47	74.03
isiXhosa		
	%High	%Low
Pitch	83.61	87.45
Intensity	85.25	83.40

Table 14: Accuracy obtained when classifying the tone of a syllable based on features derived from F0, intensity, for preceding High and Low tones, respectively.

5 CONCLUSION

5.1 Contribution

The final aim of this paper was to build a model for the relationship between tone and measurables such as pitch and amplitude, for isiZulu and isiXhosa as representatives of the Nguni languages. For this to be achieved, there were a number of other experiments that initially needed to be completed. Firstly, we had to select an an appropriate pitch tracking algorithm for these languages, which to our knowledge was not done before. The selected algorithm needed to cater for the unique characteristics of these languages. Praat's pitch tracking algorithm which uses a modified version of the autocorrelation method produced the best results in our experiments for these languages.

Secondly, we had to develop an intonation corpus for isiZulu and isiXhosa. Praat's pitch tracking algorithm was then used to extract relevant features from the spoken utterances of isiZulu and isiXhosa. These features included fundamental frequency, amplitude and duration, which were used to build intonation corpora for these languages.

Thirdly, we wanted to obtain a better understanding of the relationship between abstract tone and physical measurables. From our experiments we concluded that pitch and intensity play comparable roles in the prediction of tone for a segment. A combination of both features does provide a slight improvement in the prediction results for both approaches. The neural network classifier produced fairly good prediction results for both languages, showing that a certain degree of non-linearity appears in the relationship between these quantities and tone.

5.2 Further application and future work

We would like to incorporate the intonation model built by the neural network classifier into a text to speech system and speech recognition system for both languages. We can then investigate if including this model into such systems makes a distinguishable difference to the quality of the system. We would also like to continue with our investigations in various ways which include:

- using larger speaker groups
- analysing different dialects within these two languages
- using other languages in the Bantu family

Finally, we have found that pitch and amplitude play comparable roles in determining abstract tone; it would be interesting to investigate whether duration is also involved in this relationship.

REFERENCES

[1] D. Hirst and A. D. Cristo. *Intonation Systems: A survey of twenty languages*, pp. 1–44. Cambridge University Press, 1998.

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

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

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

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

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

[7] A. de Cheveigné and H. Kawahara. "YIN, a fundamental frequency estimator for speech and music". In *Journal of Acoustical Society of America*, pp. 1917–1930. 2002.

[8] A. de Cheveigné and H. Kawahara. "Comparative evaluation of F0 estimation algorithms". In *EuroSpeech*, pp. 2451–2454. 2001.

[9] E. Barnard, R. Cole, M. Vea and F. Alleva. "Pitch detection with a neural-net classifier". In *IEEE Transaction on Signal Processing*, vol. 39, pp. 298–307. 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] P. Boersma. "praat".
URL `http://www.fon.hum.uva.nl/praat/`.

[12] N. Campbell. "English".
URL `http://recherche.ircam.fr/equipes/pcm-/cheveign/data/f0/databases.html`.

[13] A.Black, P.Taylor and R.Caley. "Festival". URL `http://www.festvox.org-/examples/cstr_us_ked_timit`.

[14] N. Henrich. "French".
URL `http://recherche.ircam.fr/equipes/pcm-/cheveign/data/f0/databases.html`.

[15] N. Govender, E. Barnard and M. Davel. "Pitch and tone in isiZulu: Initial Experiments". In *Interspeech:9th International Conference on Spoken Language Processing*, pp. 1417–1420. 2005.

[16] D. Hirst and R. Espesser. "Automatic Modelling of Fundamental Frequency Using a Quadratic Spline Function". *Travaux de l'Institut de Phonetique d'Aix en-Provence*, vol. 15, pp. 75–85, 1993.