UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

INTONATION MODELLING FOR THE NGUNI LANGUAGES

NATASHA GOVENDER

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

INTONATION MODELLING FOR THE NGUNI LANGUAGES

By

Natasha Govender

Submitted in partial fulfilment of the requirements for the degree

Master of Computer Science

in the

Faculty of Engineering, the Built Environment and Information Technology

at the

UNIVERSITY OF PRETORIA

Advisor: Professor Etienne Barnard

April 2006

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

INTONATION MODELLING FOR THE NGUNI LANGUAGES

BY NATASHA GOVENDER

PROFESSOR ETIENNE BARNARD

DEPARTMENT OF COMPUTER SCIENCE, MASTER OF COMPUTER SCIENCE

Although the complexity of prosody is widely recognised, there is a lack of widely-accepted descriptive standards for prosodic phenomena. This situation has become particularly noticeable with the development of increasingly capable text-to-speech (TTS) systems. Such systems require detailed prosodic models to sound natural. 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.

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 and isiXhosa as representatives of the Nguni languages.

Firstly, we need to extract physical measurements from the voice recordings of the Nguni family of languages. A number of pitch tracking algorithms have been developed; however, to our knowledge, these algorithms have not been evaluated formally on a Nguni language. In order to decide on an appropriate algorithm for further analysis, evaluations have been performed on two state-of-the-art algorithms namely the Praat pitch tracker and Yin (developed by Alain de Cheveingné). Praat's pitch tracker algorithm performs somewhat better than Yin in terms of gross and fine errors and we use this algorithm for the rest of our analysis.

For South African languages the task of building an intonation model is complicated by the lack of intonation resources available. 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 order to understand how the 'expected' intonation relates to the actual measured characteristics extracted, we developed two different statistical approaches to build intonation models for isiZulu and isiXhosa. The first is based on straightforward statistical techniques and the second uses a classifier. Both intonation models built produce fairly good accuracy for our isiZulu and isiXhosa sets of data. The neural network classifier used produces slightly better results for both sets of data than the statistical method. The classification model is also more robust and can easily learn from the training data. We show that it is possible to build fairly good intonation models for these languages using different approaches, and that intensity and fundamental frequency are comparable in predictive value for the ascribed tone.

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

TABLE OF CONTENTS

CHAPTER ONE

INTONATION MODELLING: HOW AND WHY

In Section 1.1 we describe what is meant by the term intonation, why intonation modelling is important and the difficulties experienced when building an intonation model for a language. In Section 1.2 we provide background information with regard to the main topics discussed in subsequent chapters.

1.1 INTRODUCTION

Intonation is a paradoxical aspect of human language [1]. It is universally used yet highly variable across languages. Every language possesses intonation and many of the linguistic and paralinguistic functions of intonation systems seem to be shared by languages of widely different origins. Despite the universal character of intonation, the specific features of a particular speaker's intonation system also 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. Intonation is achieved by varying the levels of stress, pitch and duration in the voice. An overview of intonation as observed in a variety of languages is provided in [1]. Since there is no universal agreement on the semantics of this domain, we use the terms prosody and intonation interchangeably.

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 further 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.

Attempting to create an intonation 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 intonation. 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. 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. A diagram depicting the role that an intonation model plays in a text-to-speech system is displayed in Figure 1.1.

For languages such as English or Japanese, for example, the ToBI marking system [4] 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 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 [5] and it's intricate tonal structure. On the other hand, little work of a quantitative nature has been published, and as Roux [6] 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

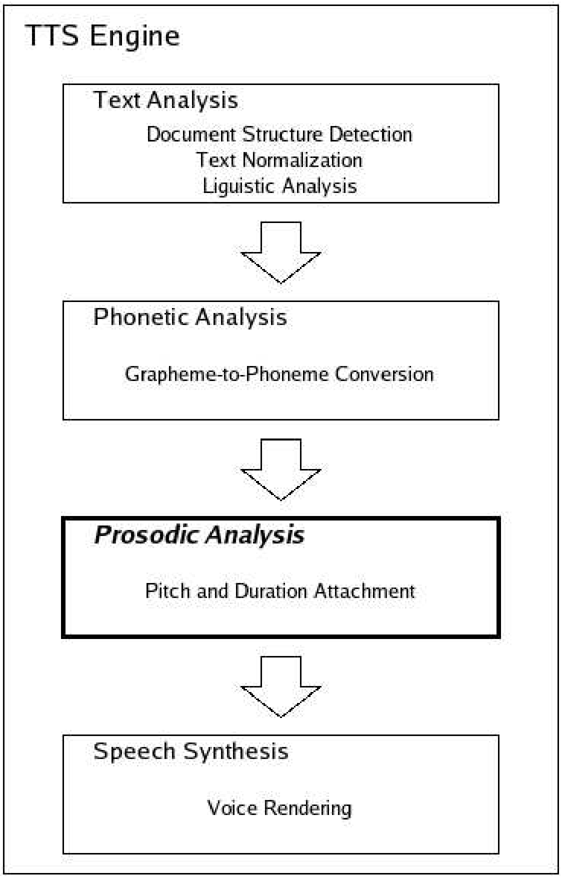

Figure 1.1: *Basic system architecture of a TTS system*

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. Tonal languages differ from other languages in three main ways:

- The length of the span of each tone melody is roughly the size of a word in a tone language, whereas in a non-tone language, its size ranges between that of a syntactic phrase and that of a sentence.
- The tone melody of an utterance in a tone language is composed of the tone melodies that are directly contributed by the lexical items in the utterance, and to a slightly lesser extent by the syntactic constructions present in the sentence, whereas the tone melody of an utterance in a non-tonal language is generally determined by the information structure of the sentence, so that it helps primarily to specify what information in the utterance is presupposed, what is taken as new, and so forth.
- Tone languages generally have phonological rules that modify the tone melodies depending on the tones found around them as well as the syntactic structure in which they occur. The changes that these tones rules give rise to may consist either of modifying the actual shape of the tone melody, or of shifting the relationship in time between the tone melody and the syllables that comprise the utterance.

Tonal languages give words a distinctive relative pitch e.g. high or low, or a distinctive pitch change e.g. level, rising, or falling which determines the meaning of the word. An example of such a word in isiZulu is shown in Table 1.1 below.

Word	Meaning
ithanga	thigh
ith**anga**	pumpkin

Table 1.1: *Same word with different meanings depending on the pitch cue*

1.2 BACKGROUND

In this section we provide background information on the main topics discussed in this thesis.

- Section 1.2.1 provides an overview of the different pitch tracking algorithms available and the corpora on which they have been evaluated.
- Section 1.2.2 discusses the current approaches used for building intonation models for languages.

- Section 1.2.3 describes the current linguistic knowledge on intonation that is available for isiZulu and isiXhosa.

1.2.1 PITCH TRACKING ALGORITHMS

To extract relevant features such as pitch, intensity and duration accurately from spoken utterances, a good pitch tracking algorithm is required.

A large number of pitch tracking algorithms have been developed to extract F0 from a speech waveform (see [7] for a review). The most comprehensive review remains that of W.Hess [8] that cites literally hundreds of methods. A few examples of recent approaches include instantaneous frequency methods [9–11], statistical learning and neural networks [12, 13] and auditory models [14, 15]. 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 [16].

A major weakness in this field is that it is often hard to judge the performance of the algorithms proposed in comparison with other methods because they are evaluated on different databases. A comparative evaluation of different fundamental frequency estimation algorithms was done by Alain de Cheveigné [17], using the same databases namely (English, Japanese and French) together with the accompanying laryngograph data. The Yin pitch tracking algorithm produced the most accurate results from the different algorithms evaluated [7]. This method combines autocorrelation and AMDF (Average Magnitude Difference Function) methods with a set of modifications to reduce errors. From the databases used in the evaluation, English and French are considered to be stress languages and Japanese is an intermediate between a stress and tone language generally referred to as a pitch accent or a tonal accented language.

On the other hand isiZulu and isiXhosa are considered to be tonal languages - defined in detail in Section 1.1 - and also possess unique characteristics such as the clicking sound present in isiZulu. For these reasons it is necessary for us to perform our own evaluations on pitch tracking algorithms using voice recordings from these two languages.

1.2.2 INTONATION MODELLING

Intonation modelling methods have been incorporated within the traditional fields of rhetoric and elocution for centuries. The most complete and accessible overview of modern prosodic analysis as embedded in mainstream linguistic theory is Ladd’s Intonational Phonology [18], which covers the precursors, current standard practise, and remaining unsolved issues of the highly influential autosegmental theory of intonation phonology.

The description of an intonation system is the result of a complex interaction between a language and an explicit or implicit theory. The difficulties experienced in describing an intonation model for a particular language are discussed in detail in Section 1.1.

A major obstacle is the remarkable absence of any consensus concerning the transcription of intonation. There are literally hundreds of approaches to modelling intonation from rule-based approaches to statistical techniques to using vector quantised models [19]. For an international transcription system to be effective, it needs to embody a certain number of hypotheses about what possible variations of prosodic features are significant across languages. There have been attempts to create such a system for modelling intonation. These are listed below:

- **INTSINT**: an *IN*ternational *T*ranscription *S*ystem for *INT*onation was developed during the preparation of a study of the intonation of twenty languages [1]. It aims to capture the surface distinctions used in different languages for building intonation patterns. It is based on a stylisation procedure of the fundamental frequency - or pitch - contour (F0) built up from interpolation between target points in which significant changes occur. It is then a system which is closely linked to the the phonetic realisation of the intonation contour, but at the same time is able to symbolise this contour in terms of a phonological representation. INTSINT aims therefore at the symbolisation of pitch levels or prosodic target points, each characterising a point in the fundamental frequency curve.

- **ToBI**: *Tone and Break Indices* was proposed for the transcribing of English [20]. There has been much interest in the possibility of adapting this system to other languages, although the authors have pointed out on several occasions that they do not believe it can be used directly for describing other languages or dialects, since, like a broad phonemic transcription, it presupposes that the inventory of tonal patterns of the languages is already established.

- **Tilt**: The Tilt intonation model facilitates automatic analysis and synthesis of intonation [21]. The analysis algorithm detects intonational events in F0 contours and parameterises them in terms of the continuously varying *tilt* parameters which represent the amount of fall and rise in the accent. A Tilt labelling for an utterance consists of an assignment of one of four basic intonational events: pitch accent, boundary tones, connections and silence. All events have a start parameter for the fundamental frequency at the start of the event (measured in Hertz). Pitch accents and boundary tones are also described by duration (seconds) and absolute amplitude (Hertz).

A number of other approaches for intonation modelling exists [1]. Since the existing intonation models for the Nguni languages are not sufficiently detailed for our purposes, we will inspect relevant intonation models built for tonal languages. We will specifically investigate the Chinese and Thai language models. The perceived tones which occur in these two languages are in some ways similar to those which are considered to occur in both the isiZulu and isiXhosa languages.

Chinese is described as possessing four lexical tones i.e. high, rising, low and falling although as Kratochvil [22] demonstrates an adequate characterisation of tonal phenomena in Chinese needs to account for both pitch and intensity variations. Thai is described as having five distinctive tones: high, mid, low, rising and falling.

To build a Thai intonation model, Sudaporn Luksaneeyanawin [23] studies the interplay of lexical tone and sentence intonation by the analysis of one word utterances. The tonal behaviour of the five tones in one word utterances under different meanings are studied and a systematic description of the tones or intonation system of Thai is postulated from the results. The fundamental frequency (F0) is extracted from the one word utterances using the PGR program module [24]. The extracted pitch values are used to obtain a clear, visual representation of the relative acoustic values of pitch, length and intensity.

The approach for the Chinese intonation model, as described by Paul Kratochvil [22], is different. Instead of using one word utterances, an informal monologue of a native speaker of the Beijing dialect of Chinese, was recorded. The words were analysed by hand from spectrograms using analytical procedures [25]. From each segment the fundamental frequency (F0), amplitude and duration were extracted at six equally distanced points within each segment. Various statistical procedures were then applied in the course of the analysis of the data. Apart from the normal descriptive procedures based on mean value and standard deviation calculations, they also used a discriminant analysis program for classifying and reclassifying data in relation to established classes on the basis of distinctive configurations of values.

The methodologies used for building intonation models for Chinese and Thai can be translated and used for creating intonation models for isiZulu and isiXhosa. However, it is difficult to decide which approach is better suited for the Nguni set of languages because there is no easy way to establish comparisons across languages. It is very often difficult to decide whether differences between descriptions are due to different theoretical and/or methodological approaches or whether they result from genuine differences between the systems constituting language specific prosodic parameters. Possible candidates for such parameters are listed below.

- *Lexical characteristics*: These are probably the most heavily theory-dependent of all prosodic characteristics. A typology is defined by two independent parameters: accent and tone. It is not at all evident what objective criteria might be used for establishing these parameters since, the fact that stress and tone are lexically distinctive in a given language does not necessarily imply any easy identifiable acoustic pattern.

- *Non-lexical characteristics*: As with dialectal variations, these can be grouped under three headings: rhythm and temporal organisation, sentence accent (and emphasis) and pitch patterns.

Since there is no concrete scientific method for deciding which of the numerous intonation models built for various languages is the best, the methodology to be used for building the intonation models for isiZulu and isiXhosa, would be a combination of techniques which have been proven to work well.

1.2.3 LINGUISTIC THEORY FOR ISIZULU AND ISIXHOSA

isiXhosa and isiZulu are traditionally regarded as tone languages proper, which implies that a specific tonal melody is assigned to a lexical representation *per se*. An interesting debate regarding the status and the phonological description of the two Nguni languages, isiZulu and isiXhosa as pure tone languages has emerged in recent years. Whilst Laughren [26] regards isiZulu as a tonal language and analyses it in autosegmental terms, Clark [27] argues for an analysis of isiZulu as a pitch-accent language. This latter approach is also found in the work of Goldsmith [28] describing the prosodics of the isiXhosa verbal system in accentual terms within a metrical framework. Claughton [29] on the other hand rejects the idea of isiXhosa being treated as a pitch-accented language and continues to view it as a tone language proper [6]. There is still much debate around this issue, but for the purposes of this thesis we will take the standpoint of Laughren and Claughton and consider isiZulu and isiXhosa as pure tonal languages.

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 [30] 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 [31] – or possibly only the first two. The tendency when describing the intonation for isiXhosa is to use two tones namely high and low [6, 28, 32].

However, in these modern formulations, the rules for assignment of tone levels to specific syllables are quite complex [5], 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.

1.3 OVERVIEW OF THESIS

Since very little work has been done in building intonation models for these languages, extraction of relevant features from the speech recordings is required. In Chapter 2, we discuss the selection of an appropriate pitch tracking algorithm for extracting relevant features from the recordings of these languages. In Chapter 3, we describe the methodology implemented for building intonation corpora for isiZulu and isiXhosa. Chapter 4 describes the two intonation models built for isiZulu and isiXhosa. In Chapter 5 we compare the accuracy of the two intonation models built. We summarise the contribution of this thesis, and discuss further applications and future work.

Chapter Two

Selection of a pitch tracking algorithm

2.1 INTRODUCTION

In order to extract physical measurements from the voice recordings of the Nguni family of languages as described in Chapter 3, an accurate pitch tracking algorithm is required. In these experiments we use voice recordings from the isiZulu language, as a representative of the Nguni family.

A number of pitch tracking algorithms have been developed as stated in Section 1.2; 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 [33] and Yin [7].

In Section 2.2 we explain some basic facts about the fundamental frequency of a speech signal and various ways in which it is extracted. In Section 2.3 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 2.4 we display the results obtained from the experiments, and in Section 2.5 we summarise our conclusions from these experiments.

2.2 FUNDAMENTAL FREQUENCY AND PITCH

We hear sounds as being different because they create different patterns of air pressure variation. When a guitar string is plucked, a similar pattern of air pressure variations is repeated at regular intervals. Each individual pattern is known as a pitch period and a waveform that consists of a number of pitch periods is said to be periodic. Speech waveforms are never absolutely periodic: the pitch periods are not all exactly the same length, and they vary slightly in shape. Nevertheless, when discussing speech waveforms, we shall refer to these as periodic while recognising that there will always be some deviations from absolute periodicity.

A periodic speech waveform is the acoustic correlate of vocal fold vibration that characterises most phonetically voiced sounds and that gives rise to the auditory quality of pitch. Since voiceless sounds are produced without vocal fold vibration, the corresponding waveform is nonrepetitive and there is no pitch.

In periodic waveforms, each pitch period corresponds to a single cycle of vocal fold vibration. Consequently, the time that each pitch period takes, or the duration of each pitch period, can be used to give an estimate of the rate of vocal fold vibration or the fundamental frequency (F0). From a listener's point of view, changes in the rate of vocal fold vibration are perceived as changes in pitch. When the vocal fold vibration/fundamental frequency increases, we hear a rising pitch; when it falls, we hear a falling pitch [34]. F0 correlates well, though not perfectly with the subjective experience of pitch. It is therefore common practice to use the terms F0 and pitch interchangeably.

In general, tones of differing pitch have different inherent perceived loudness. The sensitivity of the ear varies with the frequency and the quality of sound. Perceived pitch will therefore also change as intensity is increased and frequency is kept constant [35]; we will investigate those effects in Chapter 4.

2.3 METHODOLOGY

Yin [7] and the Praat [33] 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 characterised 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 20 milliseconds was used. Although Praat extracted the pitch values at the correct time intervals i.e. every 20 milliseconds, the time points at which it extracted the pitch did not correspond to those produced by Yin. To rectify this problem, linear interpolation was used to align the times that the pitch values were extracted by both algorithms.

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 were 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.

2.3.1 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. The speech recordings are available at http://recherche.ircam.fr/equipes/pcm/cheveign/data/nick/nick_speech.tgz and the laryngograph data at http://recherche.ircam.fr/equipes/pcm/cheveign/data/nick/nick_lx.tgz

- 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. The voice recordings are available at http://www.lam.jussieu.fr/Individu/Henrich/database.

- 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.

2.3.2 ALGORITHMS

Both algorithms used in the experiments use the basic autocorrelation function for pitch tracking, with each implementing their own set of modifications to the function.

Autocorrelation is useful for finding repeating patterns in a signal, such as determining the presence of a periodic signal which has been buried under noise, or identifying the fundamental frequency of a signal which doesn't actually contain that frequency component, but implies it with many harmonic frequencies.

The autocorrelation function (ACF) of a discrete signal x_t may be defined as:

$$r_t(\tau) = \sum_{j=t+1}^{t+W} x_j x_j +_{\tau} \qquad (2.1)$$

where $r_t(\tau)$ is the autocorrelation function of lag τ calculated at time index *t*, and *W* is the integration window size.

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

- *Yin* is an implementation of the method developed by Alain De Cheveigné [7]; it combines the Autocorrelation Function and Average Magnitude Difference Function(AMDF) methods [12] with a set of modifications that reduce common errors of those algorithms. No post-processing is used.

- The *Praat* pitch tracker performs an acoustic periodicity detection on the basis of an accurate autocorrelation method, as described in Boersma [36]. 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/.

2.4 RESULTS

2.4.1 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 2.1. 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).

Database	Praat	Yin
English DB1	3.87	12.18
English DB2	0.23	10.27
French	49.67	65.87
isiZulu	0.80	1.30

Table 2.1: *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)*

2.4.2 ERRORS IN THE DETECTION OF VOICING

Tables 2.2 and 2.3 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.

Database	Praat	Yin
English DB1	0.05	26.68
English DB2	0.28	34.10
French	17.70	65.65

Table 2.2: *The average number of false positive voicing detections per utterance*

2.4.3 MEAN SQUARE ERROR

Table 2.4 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

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

Database	Praat	Yin
English DB1	75.39	10.92
English DB2	38.73	4.15
French	63.84	15.79

Table 2.3: *The average number of false negative voicing detections per utterance*

algorithm consistently more accurate than Yin. (The values reported in Table 2.4 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.

Database	% Mean Squared Error	
	Praat	Yin
English DB1	0.19	1.82
English DB2	0.08	1.88
French	0.39	1.08

Table 2.4: *The average mean squared error of both algorithms when compared with laryngograph measurements*

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 2.4, but broadly in line with those values.

2.5 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 errors. 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 analysing 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.

CHAPTER THREE

DEVELOPING INTONATION CORPORA FOR ISIXHOSA AND ISIZULU

3.1 INTRODUCTION

The description of an intonation system of a language or dialect is a particularly difficult task since intonation is paradoxically at the same time one of the most universal and one of the most language specific features of human language. Every language possesses intonation and many of the linguistic and paralinguistic functions of intonation systems seem to be shared by languages of widely different origins. Despite the universal character of intonation, the specific features of a particulars speaker's intonation system also depend strongly on the language, the dialect, and even the style, the mood and the attitude of the speaker.

Thus, attempting to create an intonation 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 intonation. It is widely agreed that the various abstract intonation characteristics (tone, stress and rhythm) interact with the syntactic and semantic characteristics of an utterance, and give rise to the physical measurements (fundamental frequency, amplitudes, durations) associated with intonation. The details of each of these processes have generally been tackled in ad-hoc fashion for each language of interest.

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 3.2 we describe the methodology used for developing a general-purpose intonation corpus. In Section 3.3 we describe the corpora developed and the various information that can be extracted. In Section 3.4 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 chapters.

3.2 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 dissertation), 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.

3.2.1 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 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 3.1.

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

Table 3.1: *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.

3.2.2 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.

3.2.3 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 marking 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 [33]. An example of the syllable markings for a portion of an isiZulu sentence in Praat is illustrated in Figure3.1.

3.2.4 EXTRACTION OF INTONATION MEASUREMENTS

3.2.4.1 FUNDAMENTAL FREQUENCY

In Chapter 2, it was shown that Praat's implementation of a pitch tracking algorithm produced the best results for the studied languages [37]. 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.

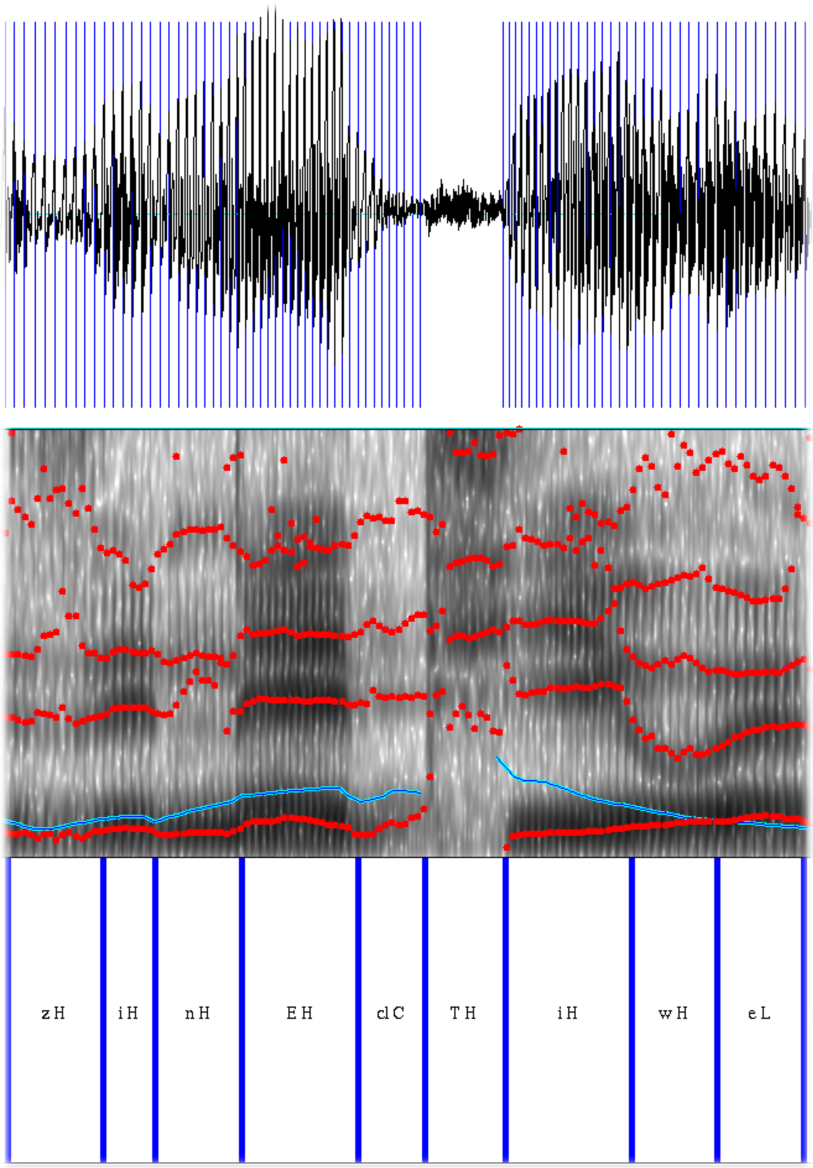

Figure 3.1: *A portion of a signal extracted for an isiZulu sentence and the pitch contour*

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) algorithm [38] 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 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 3.2. Note how the ’undefined’ values provided by Praat (zeros 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 3.3. 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*.

From the experiments the Non-Zero method was proven to obtain more accurate results. The experiments and the overall results are discussed in detail in Chapter 4.

3.2.4.2 INTENSITY

The intensity was calculated at each of the syllable boundaries, as the average squared value of the signal within a 5 millisecond window.

3.2.4.3 DURATION

To calculate duration of each syllable, the starting and ending times of the syllable were obtained from the hand labels, and subtracted.

3.3 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,

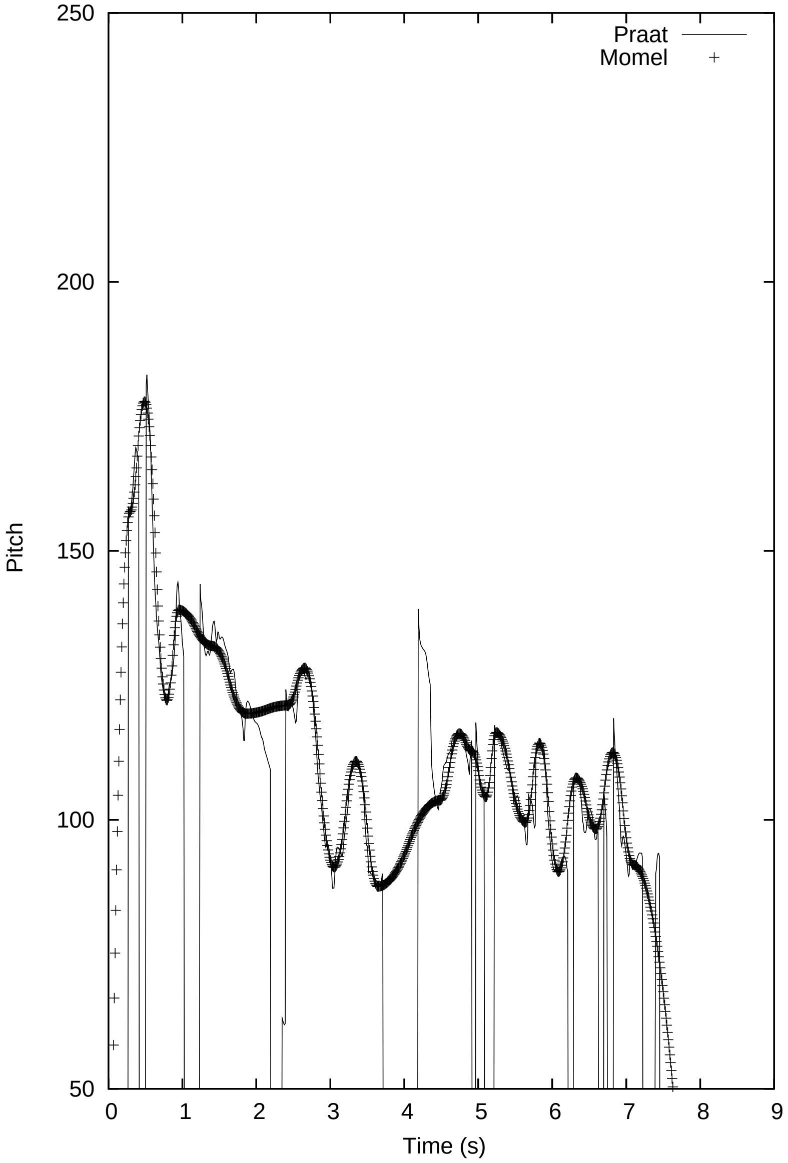

Figure 3.2: *A graph showing the difference in the pitch values obtained between Praat and MOMEL*

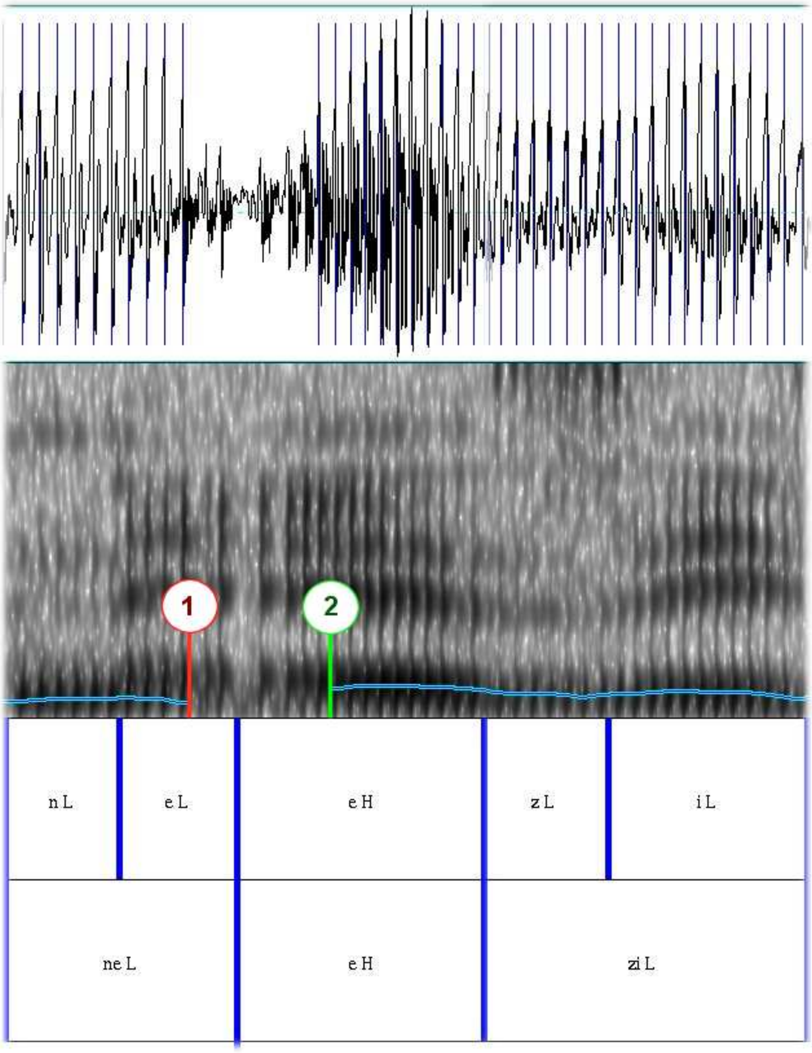

Figure 3.3: *A portion of an isiZulu signal depicting the points at which the pitch values would be extracted using the non-zero method*

- 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 3.2 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	Marking	starting F0	ending F0	Intensity	Duration
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 3.2: *An example of an annotated isiZulu data item*

3.4 OBSERVATIONS

3.4.1 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 3.4 for the two isiZulu speakers, and in Figure 3.5 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.

3.4.2 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.

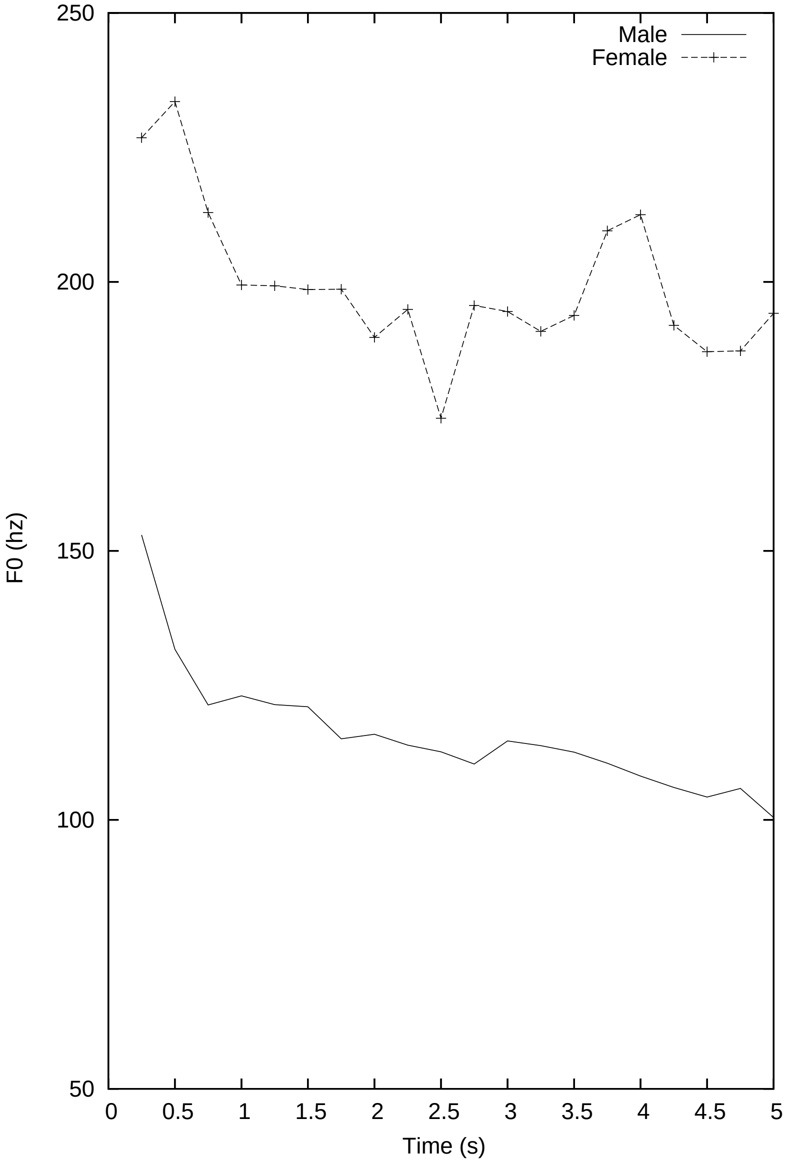

Figure 3.4: *Mean pitch as a function of the time since the start of an utterance, for isiZulu*

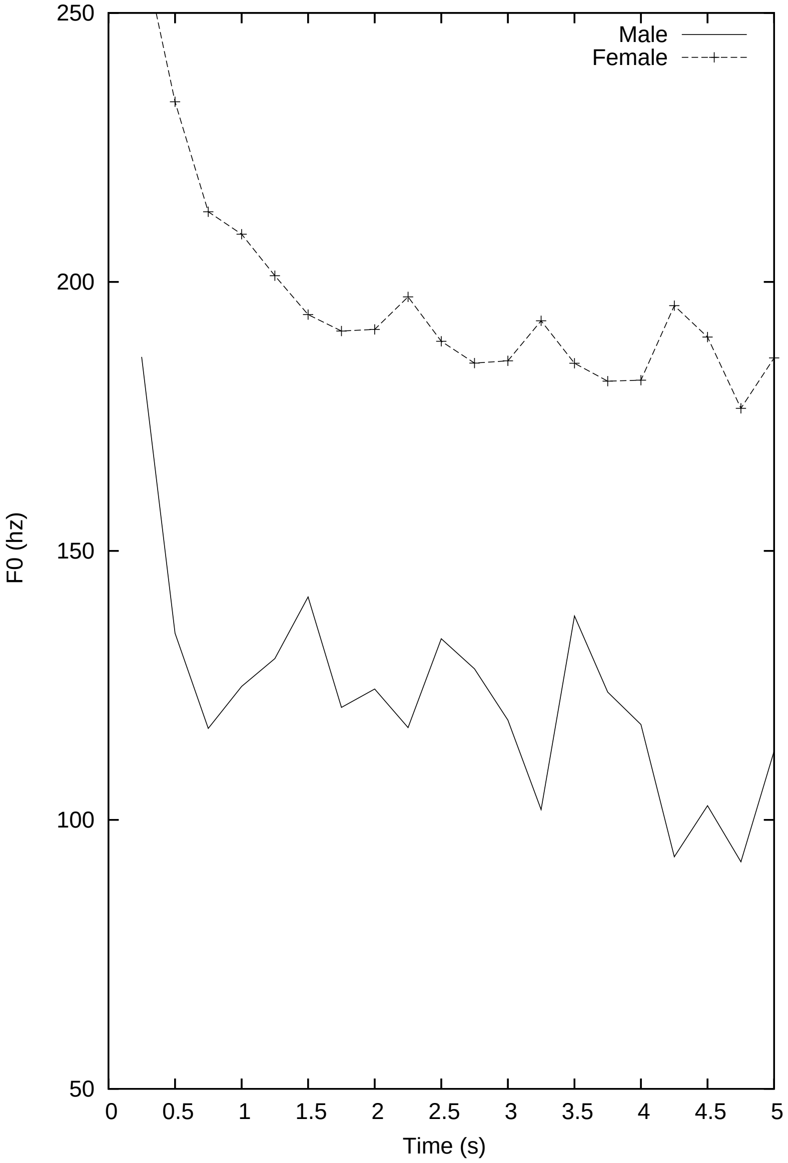

Figure 3.5: *Mean pitch as a function of the time since the start of an utterance, for isiXhosa*

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 3.3: 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.

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 3.3: *Average pitch variance values for male and female speakers*

3.5 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 more variable than those of males in one language but not the other.

CHAPTER FOUR

COMPUTATIONAL MODELS OF PROSODY

4.1 INTRODUCTION

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.

Two different statistical approaches were used to build such a model. The first is based on straightforward statistical techniques and the second uses a classifier. The process used to extract the relevant features from voice recordings is described in detail in Chapter 3.

In Section 4.2 we describe the methodology and results obtained from building an intonation model using statistical techniques. In Section 4.3 we describe the methodology and results obtained from building an intonation model using a neural network classifier.

4.2 INTONATION MODELLING USING A STATISTICAL APPROACH

4.2.1 INTRODUCTION

Statistical pattern recognition techniques have been applied to many natural language processing tasks, such as part-of-speech tagging, grapheme-to-phoneme prediction, base noun phrase chunking and prepositional-phrase attachment to name a few [39]. In all of the above tasks, statistical data analysis has resulted in successful models, in the process either complimenting or contradicting accepted linguistic practise. We are interested in the application of pattern recognition techniques to the task of intonation modelling.

In Section 4.2.2 we define the methodology used to select an algorithm for predicting tone. In Section 4.2.3 we describe the statistical methods implemented to improve the initial results obtained from the selected algorithm. In Section 4.2.4 we determine if intensity has an influence on the tone of a sentence.

4.2.2 METHODOLOGY

In Chapter 3, two algorithms were implemented, namely Momel and the Non-Zero method, to overcome the problem of extracting undefined pitch values for unvoiced segments of speech. We ran several experiments to determine which of the two algorithms produced more accurate pitch values. We used the pitch values extracted by these two methods as inputs into our two methods used to predict the tone of a segment i.e. if a segment should be assigned a tone of high or low. The points at which the starting and ending pitch values were extracted for each method are explained in Chapter 3. The average pitch for each segment was calculated using the average of the starting and ending pitch values. For both the methods the average pitch value was used to predict the tone of the segment.

The methods implemented for predicting tone are:

- **Word Average Predict**: This method calculates the pitch average of each word in a sentence using the starting and ending pitch values of that word. The average pitch for each individual syllable in the word is also calculated. The syllable's pitch average is then compared to the average pitch of the word it constitutes. If the syllable's pitch value is greater, it is assigned a tone of *high* else it is assigned a tone of *low*.

- **Relative Predict**: This method predicts the tone of each syllable based on the pitch value and assigned tone of the preceding syllable. The first syllable of a sentence is assigned an unknown tone *x*. The average pitch of the next syllable is then inspected. If it is higher than the first syllable's average pitch, it is assigned a tone of high and the first syllable is assigned a tone of low. If the pitch of the second syllable is lower, it is then assigned a tone of low and the first syllable a tone of high. For each syllable thereafter, it's average pitch is compared to the previous syllable's pitch average and depending on whether it is higher or lower, it is assigned a tone of high or low respectively. This process is reset after each sentence.

These experiments were conducted for both the isiZulu and the isiXhosa corpora. Accuracy of each method was calculated by comparing the predicted tone to those originally assigned to each syllable by first language isiZulu and isiXhosa speakers as described in Chapter 3. Tables 4.1 and 4.2 display the results obtained from predicting high and low tones using the pitch values extracted by Momel and the Non-Zero method respectively.

The results indicate that the pitch values extracted using the Non-Zero method produces better results than the pitch values extracted using Momel, for both prediction methods. The relative predict method produces more accurate results than the word average predict method.

Database	% Word Average Predict	%Relative Predict
Female isiXhosa	54.12	65.20
Male isiXhosa	53.10	63.30
Male isiZulu	52.10	58.31

Table 4.1: *Accuracy of tone prediction using pitch values extracted using Momel*

Database	% Word Average Predict	%Relative Predict
Female isiXhosa	57.02	67.50
Male isiXhosa	57.65	65.27
Male isiZulu	53.11	60.32

Table 4.2: *Accuracy of tone prediction using pitch values extracted using the Non-Zero method*

4.2.3 IMPROVEMENTS

The results obtained using the relative predict algorithm does produce better results than the word average predict method, but these results can be improved. A few parameters were introduced before predicting the tone of syllable. We use the pitch values extracted by the Non-Zero method and the relative predict algorithm.

4.2.3.1 REMOVING PAUSES

Pauses occur between sentences, phrases or words. This occurs when the speaker pauses to breathe or for emphasis. When the sentences were transcribed into their phoneme and word level using the Praat software, the pauses between words were also marked. An example of this type of marking is shown in Figure 4.1.

Pitch values that were extracted from the pause boundaries were now removed when predicting tone. For the isiZulu corpus, clicks were also removed. The results obtained when the pauses were removed, and clicks for the isiZulu corpus, are displayed in the Table 4.3.

Database	% Relative Predict(without pauses)
Female isiXhosa	67.90
Male isiXhosa	66.95
Male isiZulu	64.33

Table 4.3: *Results obtained when pauses were removed*

By removing the pauses in both databases and the clicking sound in the isiZulu corpus, there is a slight improvement in accuracy. This may be due to that fact that pauses have little or no impact on

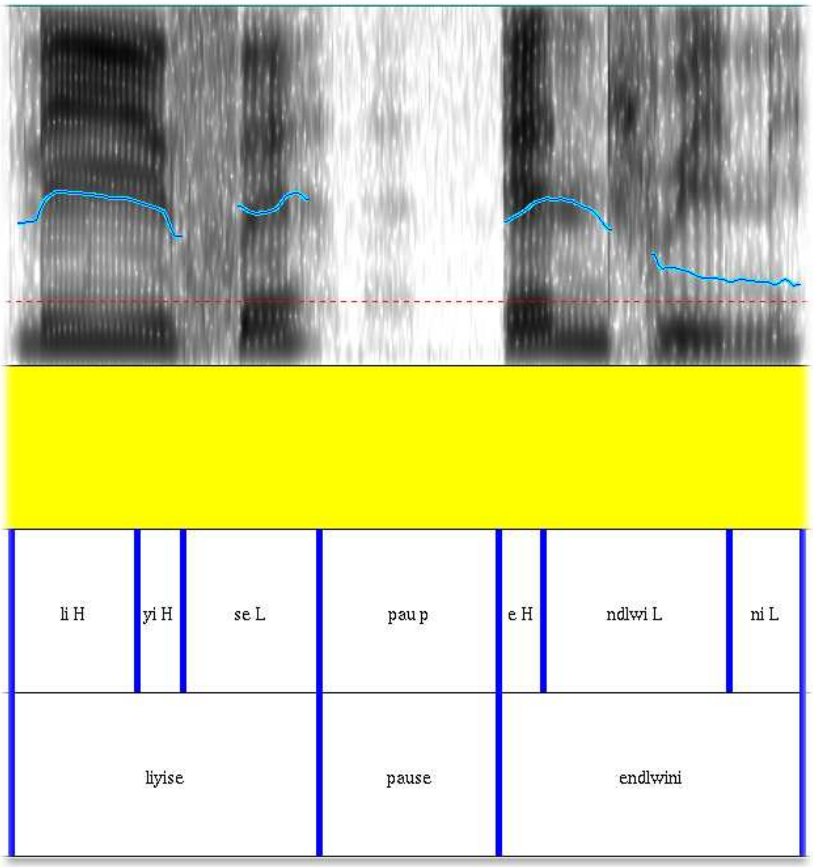

Figure 4.1: *A portion of a signal extracted for an isiXhosa sentence showing a pause in the sentence.*

the tone of the syllable or word following it.

4.2.3.2 THRESHOLD

A threshold parameter was implemented when comparing the pitch values of two syllables. If the pitch value of the second syllable is higher than the pitch value of the preceding syllable, it is assigned a tone of high or vice versa, as explained in Section 4.2.2. With a threshold implemented, this process is altered. If the pitch value of the second syllable is within a certain range of the previous syllable's pitch value, it is assigned the same tone as the previous syllable regardless of whether the pitch value is higher or lower. A syllable would be assigned a different tone if its pitch value falls outside the given range and the general rule would then apply.

Varying threshold values were experimented with across each database and for each speaker. The threshold that produces the most accurate result for each speaker is highly dependant on the characteristics of the individual's voice and the amount of pitch variation that is used. This experiment excluded the pitch values extracted from pauses as described in the previous section. As shown in Chapter 3 females vary their pitch more than males, so it is expected that the threshold that best distinguishes a high tone from a low tone for a female would be lower than that for a male speaker. Table 4.4 displays the results obtained for a female isiXhosa speaker and Table 4.5 for a male isiXhosa speaker.

Threshold	Accuracy
15	63.10
10	67.93
8	68.89
5	69.28
3	68.00

Table 4.4: *Results of incorporating a threshold while predicting tone for a female isiXhosa speaker*

Threshold	Accuracy
25	70.00
20	74.43
18	74.24
15	72.12
10	68.35

Table 4.5: *Results of incorporating a threshold while predicting tone for a male isiXhosa speaker*

For a female isiXhosa speaker, the threshold that best distinguishes a high tone from a low tone is 5 hertz and this improves the accuracy from 67.9% to 69.28%.

For a male isiXhosa speaker, a threshold of 20 hertz produces the most accurate results and increases substantially from 66.95% to 74.43%.

The same experiment was conducted on the male isiZulu voice recordings. The results are displayed in Table 4.6.

Threshold	Accuracy
20	53.35
10	67.15
8	68.92
6	70.12
5	68.56
2	68.01

Table 4.6: *Results of incorporating a threshold while predicting tone for a male isiZulu speaker*

For a male isiZulu speaker, the threshold that produces the best results is 6 hertz. It improves the prediction results from 64.327% to 70.12%.

4.2.3.3 INCORRECT PITCH VALUES

A pitch tracking algorithm can sometimes extract incorrect pitch values for certain voice segments if it cannot properly follow the pitch contour. This results in pitch values being extracted that are either very low or very high and are clearly out of an individual's speaking range.

To determine the speaking range of each individual, the average pitch value was calculated for each speaker using the entire database of voice recordings. From manual inspection it was determined that approximately 35% above the average pitch value constitutes the highest pitch and 35% below the average pitch value, the lowest pitch for the particular individual. The speaking range for each speaker is displayed in Table 4.7.

Database	Lowest Pitch (Hertz)	Highest Pitch (Hertz)
Female isiXhosa	83.25	400.99
Male isiXhosa	61.57	340.37
Male isiZulu	65.68	360.54

Table 4.7: *Speaker Range*

Any pitch values lying outside these ranges were then excluded when predicting tone. After filtering out these values,(and removing the pauses and implementing a threshold), the results when predicting tone are displayed in Table 4.8.

Removing pitch values that are clearly incorrect i.e. either extremely low or extremely high, improves the prediction accuracy.

Database	Accuracy
Female isiXhosa	71.15
Male isiXhosa	75.56
Male isiZulu	70.95

Table 4.8: *Results of tone prediction when incorrect pitch values were removed*

4.2.3.4 FALLING TONE

At the physical level, falling tones are present in the Nguni languages. A falling tone may be perceived as high tone by one individual or a low tone by another individual. A syllable that has a falling tone starts with a high pitch and ends on a much lower pitch.

We searched for falling tones in both corpora by looking at the difference between the starting and ending pitch values of each syllable. If the difference was greater than a pre-determined value, it was considered a falling tone. These syllables were then flagged as falling and disregarded when calculating accuracy because of their ambiguous nature.

Experiments were then conducted to determine which threshold could best predict a falling tone for each language. These experiments also included all the measures implemented in the previous sections. The results for the isiXhosa and isiZulu corpora are displayed in Tables 4.9 and 4.10 respectively.

Falling tone value	%isiXhosa Female	%isiXhosa Male
20	79.61	82.46
25	77.38	80.58
40	73.78	78.61
50	72.36	77.77

Table 4.9: *Results of removing falling tones from the isiXhosa database*

Falling tone value	%isiZulu Male
25	68.93
30	69.16
40	71.17
50	71.25
60	69.76

Table 4.10: *Results of removing falling tones from the isiZulu database*

By not taking into account falling tone for the isiXhosa database, the prediction result is improved significantly. There is only a slight improvement in the isiZulu database indicating that falling tones may be more prevalent in our isiXhosa recordings than in our isiZulu recordings.

4.2.4 INTENSITY

Thus far we have investigated the effects of only pitch on the tone of syllables. The perception of intonation is highly complex and involves a number of different attributes including pitch, intensity and duration. In this section, we use the extracted set of intensity-derived features to determine if intensity can be used successfully in predicting tone either on its own or in combination with pitch values.

The starting and ending intensity values were extracted for each syllable. The average intensity of each syllable was calculated using the average of the starting and ending intensity values of that syllable.

Initially the tone of a syllable was predicted based only on its average intensity. For a syllable to be assigned a high tone it should be greater than a pre-determined value, else it is assigned a tone of low. Several thresholds were used in the experiment. The results are displayed in Table 4.11.

Intensity Threshold	%isiXhosa female	%isiXhosa male	%isiZulu male
70	30.35	26.49	40.86
80	68.27	49.78	58.16
85	75.49	77.91	63.12

Table 4.11: *Results for predicting tone based only on extracted intensity values*

As shown in Table 4.11 using only the extracted intensity values to predict the tone of a segment, we obtain fairly good results. This indicates that intensity does have an influence on the intonation of a syllable.

We then wanted to investigate the implications of using both the extracted pitch and intensity values. The results are displayed in Table 4.12.

%isiXhosa female	%isiXhosa male	%isiZulu male
80.60	84.32	70.76

Table 4.12: *Results for predicting tone using both pitch and intensity values*

The results from the experiments conducted indicate that using the average pitch and intensity values individually does produce good results for predicting the tone of a segment. It is widely accepted that pitch/fundamental frequency is the best indicator of tone. Here we have shown that a combination of pitch and intensity values produces better results for prediction, indicating that intensity also has an influence on the intonation of a sentence.

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

Table 4.13: *Segment Tone Combinations*

4.3 INTONATION MODELLING USING CLASSIFIERS

4.3.1 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 4.13.

For each combination above, the differences between consecutive pitch values were calculated and plotted. The average pitch value of each segment was used. Figure 4.2 displays the results of a ’HH’ combination plotted against a ’HL’ combination and Figure 4.3 displays the results of a ’LL’ combination plotted against a ’LH’ combination for the isiXhosa corpus using only pitch values.

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. Figure 4.4 displays the results of a ’HL’ combination plotted against a ’HH’ combination and Figure 4.5 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 4.4 and Figure 4.5 that pitch values combined with the intensity values are also reasonably separable and should work well as input to a classifier.

In Section 4.3.2 we describe the classifier used in the experiments. In Section 4.3.3 we describe the methodology implemented to build the intonation model and the results obtained from the classifier.

4.3.2 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.

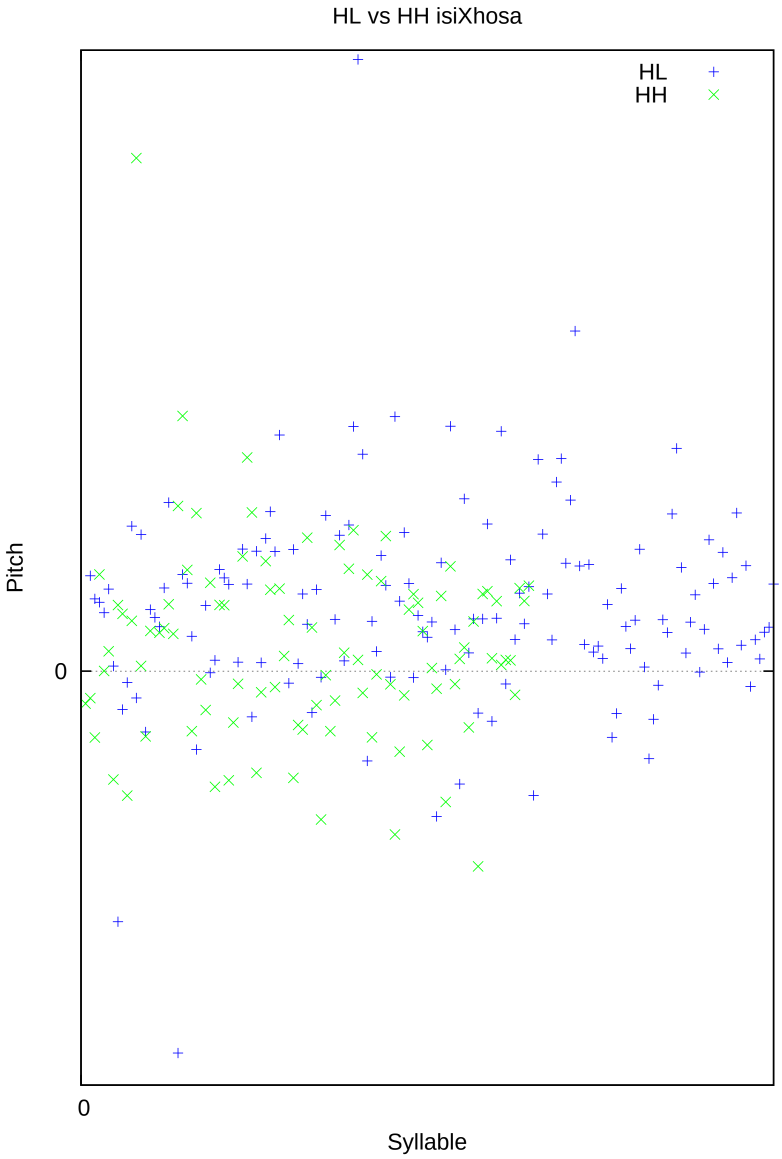

Figure 4.2: *HL consecutive segments plotted against HH consecutive segments using only pitch values for isiXhosa*

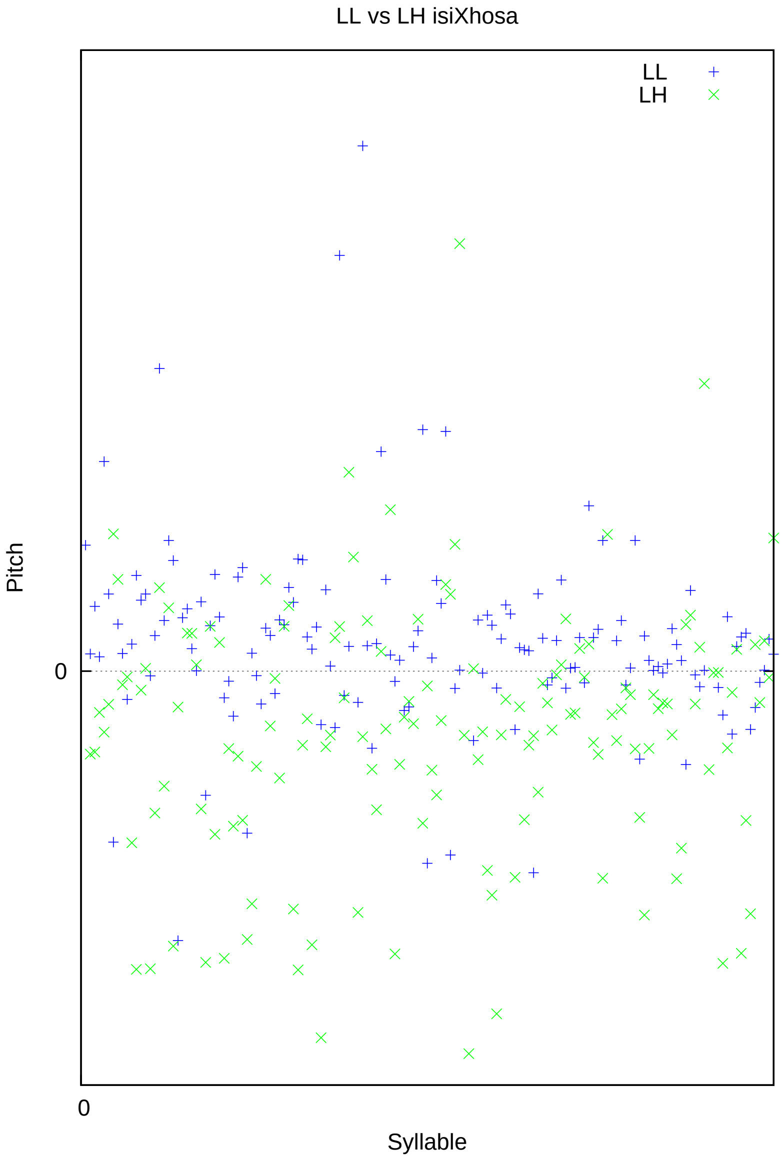

Figure 4.3: *LL consecutive segments plotted against LH consecutive segments using only pitch values for isiXhosa*

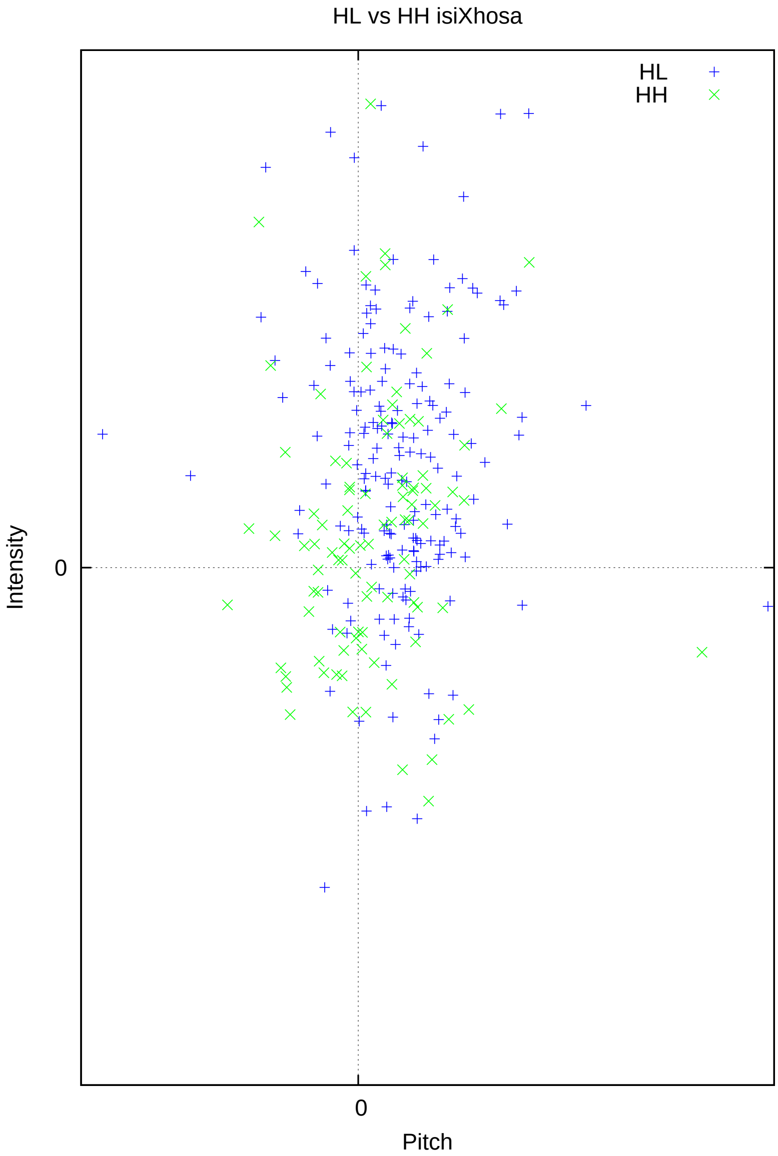

Figure 4.4: *HL consecutive segments plotted against HH consecutive segments using pitch and intensity values for isiXhosa*

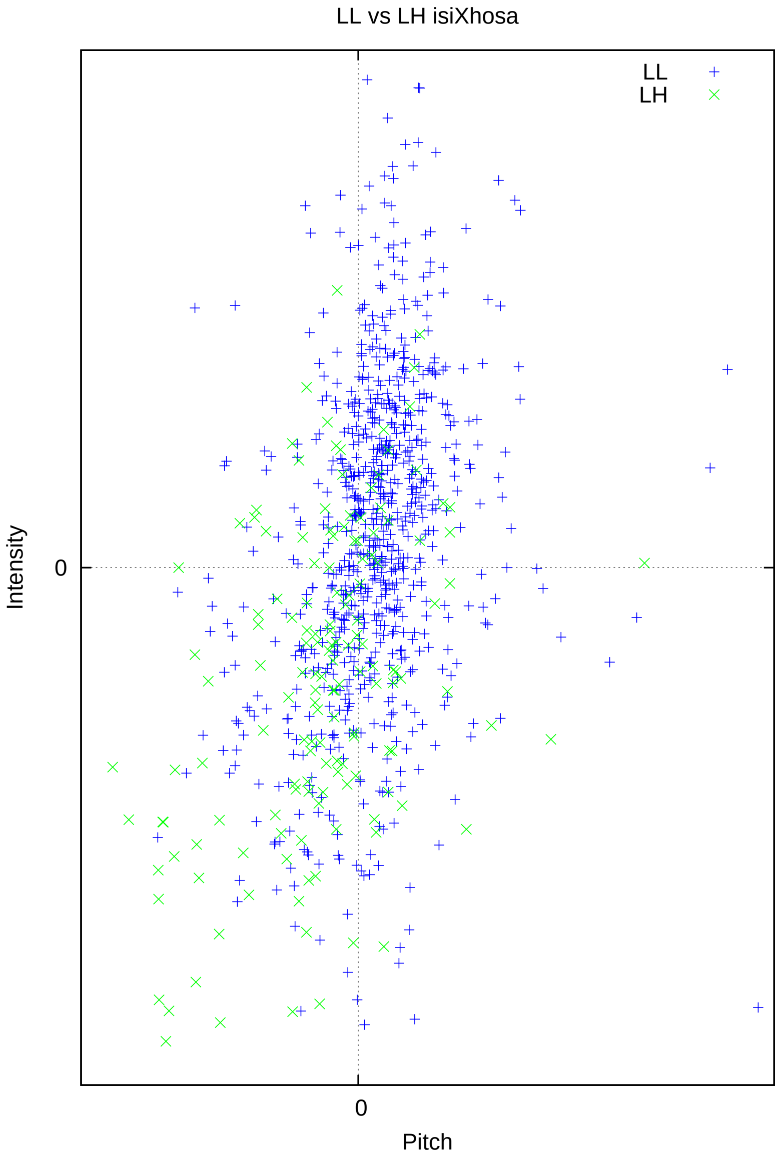

Figure 4.5: *LL consecutive segments plotted against LH consecutive segments using pitch and intensity values for isiXhosa*

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. This is the learning phase.

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.

4.3.3 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 4.14, 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).

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

Table 4.14: *The number of utterances and syllables used for the training and testing of the classifiers.*

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

- starting pitch
- ending pitch
- average pitch
- 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 was 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 4.15 and Table 4.16 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 biasness in the classification, we boosted the number of 'HH' segment combination 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 4.15: *Classifier results for each individual feature for the isiZulu 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 4.17 and for isiXhosa in Table 4.18.

As shown in Table 4.19, we compared the classification accuracies achievable with the F0-derived features to those of the amplitude derived features.

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

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 4.16: *Classifier results for each individual feature for the isiXhosa database*

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

Table 4.17: *Classifier results for combination of features for the isiZulu database*

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 two intonation models 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 Roux [6] 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.

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

Table 4.18: *Classifier results for combination of features for the isiXhosa database*

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

Both intonation models built produce fairly good accuracy for our isiZulu and isiXhosa sets of data. Both intonation models use pitch values extracted by Praat using the Non-Zero method. The statistical model using the relative predict method to predict the tone of the segment using just the average pitch and intensity values. The classifier model uses four extracted F0-based features and four amplitude-based features (as mentioned in Section 4.3) for classification. Thus, the models are not entirely comparable because of the different feature sets they use to predict tone.

The neural network classifier, which uses eight extracted features, does produces overall better prediction results. The statistical model is very much aligned with the voice characteristics of an individual and it is also time consuming obtaining the various parameters for the model. The classification model is also more robust and can easily learn from the training data but the eight features also need to be extracted. We have shown that it is possible to build fairly good intonation models for these languages using different approaches.

CHAPTER FIVE

CONCLUSION

5.1 INTRODUCTION

As initially discussed in Chapter 1, the aim of this thesis was three-fold: (a) to obtain an effective pitch tracking algorithm for the Nguni set of languages; (b) to build an intonation corpus for isiZulu and isiXhosa; and (c) to obtain a better understanding of the relationship between abstract tone and physical measurables.

In this chapter we evaluate the extent to which we were able to achieve these goals and discuss further applications and future work.

5.2 CONTRIBUTION

The final aim of this thesis 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 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.

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.

We built intonation models for both languages using the two approaches described in Chapter 4. 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 better prediction results than the statistical approach for both languages, showing that a certain degree of non-linearity appears in the relationship between these quantities and tone. The publications arising from this research were the following:

- Fundamental frequency and tone in isiZulu: initial experiments [37]
- Developing intonation corpora for isiXhosa and isiZulu [40]
- Computational models of prosody in the Nguni languages [41]

5.3 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] Daniel Hirst and Albert Di Cristo, *Intonation Systems: A survey of twenty languages*, pp. 1–44, Cambridge University Press, 1998.

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

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

[4] M.E Beckman and J.B Pierrehumbert, "Intonational structure in Japanese and English," in *Phonology Yearbook*, 1986, vol. 3, pp. 255–309.

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

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

[7] Alain de Cheveigné and Hideki Kawahara, "Yin, a fundamental frequency estimator for speech and music," in *Journal of Acoustical Society of America*, 2002, pp. 1917–1930.

[8] W. Hess, *Pitch determination of speech signals*, Berlin: Springer-Verlag, 1983.

[9] T.Abe, T.Kobayashi, and S.Imai, "Harmonics tracking and pitch extraction based on instantaneous frequency," in *Proceedings of IEEE-ICASSP*, 1995, pp. 756–759.

[10] Y.Atake, T.Irino, H.Kawahara, J.lu, S.Nakamura, and K.Shikano, "Robust fundamental frequency estimation using instantaneous frequencies of harmonic components," in *Proceedings of ICLSP*, 2000, vol. 2, pp. 907–910.

[11] H.Kawahara, I.Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous frequency-based f0 extraction," in *Speech Communication*, 1999, vol. 27, pp. 187–207.

[12] E. Barnard, R.A. Cole, M.P. Vea, and F.A. Alleva, "Pitch detection with a neural-net classifier," in *IEEE Transaction on Signal Processing*, 1991, vol. 39, pp. 298–307.

[13] X.Rodet and B.Doval, “Maximum-likelihood harmonic matching for fundamental frequency estimation,” *Journal of the Acoustical Society of America*, vol. 92, pp. 2428–2429, 1992.

[14] H.Duifhuis, L.F. Willems, and R.J. Sluyter, “Measurement of pitch in speech: an implementation of Goldstein’s theory of pitch perception,” in *ICPhS*, 1991, pp. 218–221.

[15] A. de Cheveigné, “Speech f0 extraction based on Licklider’s pitch perception model,” in *Journal of Acoustical Society of America*, 1982, pp. 1568–1580.

[16] Xiao Li, Jonathan Malkin, and Jeff Bilmes, “Graphical model approach to pitch tracking,” in *Interspeech:8th International Conference on Spoken Language Processing*, 2004, pp. 1101–1104.

[17] Alain de Cheveigné and Hideki Kawahara, “Comparative evaluation of f0 estimation algorithms,” in *EuroSpeech*, 2001, pp. 2451–2454.

[18] R. D. Ladd, *Intonational Phonology*, Cambridge University Press, 1996.

[19] A.K. Syrdal, G.Moler, Kurt Dusterhof, A. Conkie, and A.W. Black, “Three methods of intonation modeling,” in *Proceedings of the Third ESCA Workshop on Speech Synthesis*, 1998, pp. 305–310.

[20] K.Silverman, M.Beckman, J.Petrelli, M.Ostendorf, C.Wightman, P.Price, J.Pierrehumbert, and J.Hirchberg, “Tobi: A standard for labeling English prosody,” in *Proceedings of the 1992 International Conference on Spoken Language Processing*, 1992, pp. 867–870.

[21] P. Taylor, “The Tilt Intonation Model,” *Journal of the Acoustical Society of America*, vol. 107, pp. 1697–1714, 2000.

[22] P. Kratochvil, *”Intonation in Beijing Chinese”,in Intonation Systems*, pp. 417–431, Cambridge University Press, 1998.

[23] S.Luksaneeyanawin, *”Intonation in Thai”,in Intonation Systems*, pp. 376–394, Cambridge University Press, 1998.

[24] B.Gold and L.Rabiner, “Parallel processing techniques for estimation of pitch periods of speech in the time domain,” *Journal of Acoustical Society of America*, vol. 46, pp. 442–448, 1969.

[25] P. Kratochvil, *”The case of the Third Tone”, in Wang Li memorial volumes, English volume*, pp. 253–276, Hong Kong: Joint Publishing Company, 1987.

[26] G.N. Clements, M. Laughren, and J.Goldsmith, *”Tone in Zulu nouns”, in Autosegmental studies in Bantu Tone*, Foris Publications, 1984.

[27] M.M. Clark, *”An Accentual account of the Zulu Noun”, in Autosegmental Studies on Pitch Accent*, pp. 51–79, Foris Publications, 1988.

[28] J.A. Goldsmith, K. Peterson, and J.Drogo, *”Tone and accent in the Xhosa verbal system”, in Current approaches to African linguistics 5*, pp. 157–177, Foris Publications, 1989.

[29] J.S. Claughton, “The tonology of Xhosa,” in *Doctoral Thesis, Rhodes University*, 1992, pp. 2459–2462.

[30] C.M. Doke, *Text-book of Zulu grammar*, pp. 1–28, Longmans, Green and Co, 1947.

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

[32] D.K. Rycroft, “Nguni tonal topology and common bantu,” in *African Language Studies XVII*, 1980, pp. 33–76.

[33] Paul Boersma, “Praat, a system for doing phonetics by computer,” in *Glot International*, 2001, pp. 341–345.

[34] Jonathan Harrington and Steve Cassidy, *Techniques in Speech Acoustics*, pp. 1–12, Kluwer Academic Publishers, 1999.

[35] Xuedong Huang, Alex Acero, and Hsiao-Wuen Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, pp. 19–49, Prentice Hall PTR, 2001.

[36] 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*, 1993, pp. 97–110.

[37] Natasha Govender, Etienne Barnard, and Marelie Davel, “Pitch and tone in isiZulu: Initial experiments,” in *Interspeech:9th International Conference on Spoken Language Processing*, 2005, pp. 1417–1420.

[38] 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.

[39] W. Daelemans, A. van den Bosch, and J. Zavrel, “Forgetting exceptions is harmful in language learning,” in *Machine Learning*, 1999, vol. 34, pp. 11–41.

[40] Natasha Govender, Etienne Barnard, and Marelie Davel, “Building intonation corpora for isiXhosa and isiZulu,” in *PRASA: Pattern Recognition Conference of South Africa*, 2005, pp. 129–132.

[41] Natasha Govender, Christiaan Kuun, Victor Zimu, Etienne Barnard, and Marelie Davel, “Computational models of prosody in the Nguni languages,” in *Multiling 2006: ISCA Tutorial and Research workshop on Multilingual Speech and Language Processing*, 2006, http://www.isca-speech.org/archive/ml06/ml06_015.html.