Developing Intonation Corpora for isiXhosa and isiZulu

Natasha Govender, Etienne Barnard, Marelie Davel

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

Abstract

In order to bring the tools of statistical pattern recognition to bear on intonation modelling, we need tailor-made corpora in the languages of interest. We describe how two such corpora were developed (for isiZulu and isiXhosa, respectively). We also show how those corpora can be used without further interpretation to gain insight into matters such as overall pitch contours and gender differences, and discuss the additional steps that will be required to create truly generative models from these corpora.

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 [1]. In all of the above tasks, statistical data analysis have 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 the task of intonation modelling.

In the literature, a variety of different meanings have been associated with the term 'intonation'. In this paper we use the term in its broad sense, to refer to the *melodic pattern of an utterance*, either occuring 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*. These perceptual or abstract characteristics correspond to physical measurements such as fundamental frequency, intensity and duration in an often complex manner. An overview of intonation as observed in a variety of languages is provided in [2].

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. But 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 this paper we describe the process for developing a general-purpose intonation corpus (section 2) and describe two such corpora developed (section 3). We also demonstrate how these corpora can used to obtain useful information regarding some basic aspects of intonation in the studied languages.

2. Methodology

For the work described in this paper, our aim was to develop an annotated intonation corpus that would 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 – see [3]), 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 is described in detail below for corpora in two Nguni languages (isiZulu and isiXhosa), 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.

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 approximately 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 occuring in the larger text corpus, as illustrated in Table 1.

For isiXhosa 53 of the original 2300 sentences were se-

Table 1: Examples of bi-gram frequency counts

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

lected. For isiZulu 153 of the original 1700 sentences were selected for recording.

2.2. Recording of Sentences

The sentences selected by the text optimizer were recorded by a first language isiXhosa male and female speaker and a first language isiZulu male and 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.

2.3. Marking of Sentences

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 not influenced by the available audio data.

For each sentence the boundaries of every syllable was marked and transcribed using Praat [4]. An example of the syllable markings for a portion of an isiZulu sentence in Praat is illustrated in Figure 1.
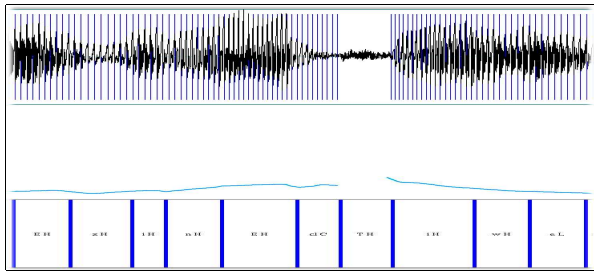

Figure 1: A portion of a signal extracted for an isiZulu sentence and the pitch contour

2.4. Extracting Intonation measurements

2.4.1. Fundamental Frequency

A large number of F0 tracking algorithms exist. Previous experiments have indicated that the Praat implementation of F0 extraction produced the best results for the studied languages [5]. The F0 extraction algorithm implemented by Praat performs an acoustic periodicity detection on the basis of an accurate auto-

correlation method [6]. This method is more accurate, noise-resistant and robust than methods based on cepstrum or combs, or the original autocorrelation method. This algorithm was selected to extract the pitch values from the isiXhosa and isiZulu recordings.

The fundamental frequency (F0) values were extracted at the syllable boundaries, i.e. they were initially 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.

In order to rectify this problem, the MOMEL (MOdelisation de MELodie) algorithm [7] 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 0.01 second, 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 Fig. 2. Note how the 'undefined' values provided by Praat (0's in the figure) have been removed in the MOMEL contour.
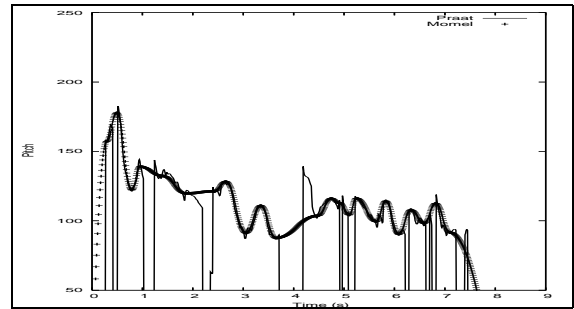

Figure 2: A graph showing the difference in the pitch values obtained between Praat and MOMEL

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.

2.4.3. Duration

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

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,
- syllabification markings,
- the expected High/Low markings for each syllable,

- the values of the raw extracted pitch,

- the values of the smoothed extracted pitch using MO-MEL

- the extracted intensity values, and

- the extracted duration values.

The tables below illustrates a typical example of the values obtained for an isiZulu sentence.

Table 2: An example of an annotated isiZulu data item

Segment	Marking	F0	Intensity	Duration
i	H	155.26	59.91	0.13
s	L	167.60	71.50	0.146
i	L	174.48	83.12	0.158
m	L	132.07	83.34	0.123
o	L	123.36	82.73	0.184

In [3], these measurements are used to derive a number of conclusions regarding the relationships between tone markings and observed F0. In the remainder of this paper, our focus is on a number of more global measurements related to F0 that were obtained from the same corpus.

3.1. Declination in F0

In many of the languages of the world, F0 has a consistent tendency to decline within a phrase [2]. 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 0.025 seconds), as a function of the duration from the beginning of the utterances. These averages are shown in Fig. 3 for the two isiZulu speakers, and in Fig. 4 for the two isiXhosa speakers.
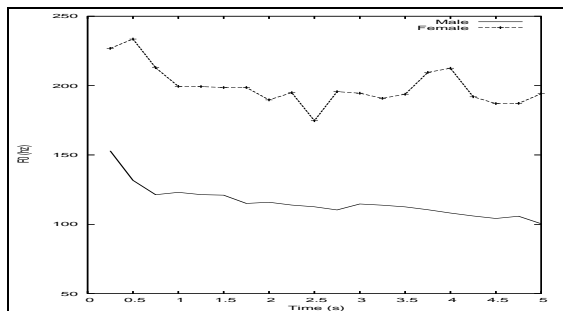

Figure 3: Mean pitch as a function of the time since the start of an utterance, for isiZulu

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

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
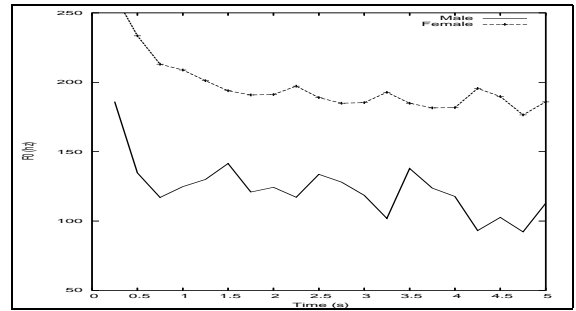

Figure 4: Mean pitch as a function of the time since the start of an utterance, for isiXhosa

of pitch variation is such a difference in the Nguni (and related) languages – specifically, we hypothesize 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 msec. increments. The results are shown in Table 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.

Table 3: Average pitch variance values for male and female speakers

Language	Male mean	Male variance	Female mean	Female variance
isiZulu	117.0	21.6	203.8	33.7
isiXhosa	122.9	38.3	197.0	36.0

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). In this paper, we have investigated a number of global charateristics 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.

Developing these corpora by collecting speech from more speakers is crucial in order to distinguish between speaker idiosyncracies, dialectal variations, and general statements within a language. In addition, it would be most interesting to collect similar corpora in related languages to see how widely some of these observations apply. Finally, the derivation of analytic models that allow us to compute reasonable contours for pitch,

loudness and duration from the written representation of an utterance remains a significant and worthwhile challenge.

5. References

[1] W. Daelemans, A. van den Bosch, and J. Zavrel, "Forgetting exceptions is harmful in language learning," *Machine Learning*, vol. 34, no. 1-3, pp. 11–41, 1999.

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

[3] C. Kuun, V. Zimu, E. Barnard, and M. Davel, "Statistical investigations into isizulu intonation," in *Proceedings of the Pattern Recognition Association of South Africa, Submitted*, 2005.

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

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

[6] 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.

[7] 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.