

Automatic intonation modeling with INTSINT

J.A. Louw and E. Barnard

Human Language Technologies Research Group
CSIR / University of Pretoria, Pretoria, 0001

jalouw@csir.co.za ebarnard@up.ac.za

Abstract

Accurate intonation modeling has become a vital part of modern day speech synthesis systems. This is especially true for tonal languages such as isiZulu, where the intonation of an utterance not only influences the perceived naturalness of the synthetic voice, but may also influence its semantics.

In this work we explore the INTSINT intonation modeling algorithm and its application to an isiZulu speech synthesiser. For fundamental frequency an algorithm, MOMEL, for the automatic derivation of a representation as a sequence of target points is applied. A symbolic coding system for fundamental frequency patterns is implemented. We show that the model's level of phonetic representation has the potential to provide an interface between abstract cognitive representations and their physical manifestations, but requires more in-depth phonological information.

1. Introduction

Prosody is the feature of text-to-speech (TTS) systems which is most in need of improvement. An important aspect of prosody is intonation. Intonation is an aspect that can be reliably extracted from the speech signal, and that can be presented in an easily understandable form. In conjunction with appropriate timing information, fundamental frequency contours furnish most of the significant prosodic information contained in speech expressions. It is thus at the center of much current interest in speech analysis and speech modeling.

The body of research into manual and automatic intonation analysis systems and techniques has been growing rapidly in the last few years [1]. Spoken language understanding systems can benefit from the structural and pragmatic information which intonation often conveys. However, current trends in speech processing have increased the need for large corpora, since stochastic speech synthesis methods (including those used for intonation modeling) require a great deal of data to be effective. Manually labeling speech databases for intonation is recognized as difficult and time consuming. Automatic labeling can decrease both time and funds spent on building the databases from which theoretical models and viable applications can be built.

Automatic analysis of intonation is a crucial step towards the long-awaited automatic transcription of intonation. The goal of the research on intonation analysis detailed in this paper is to create a system which can automatically label speech with intonation information. This contribution evaluates a method for creating intonation models from recorded speech. The goal is to predict fundamental frequency contours, given

the orthography. The emphasis is on automatic data-driven techniques. Data-driven models can easily be adapted to different speakers, different text styles and different languages.

An important aspect of this work is the selection of universal (language independent) linguistic factors that are important for predicting observed intonation phenomena. These selected linguistic features (e.g. part-of-speech, type of punctuation) are then combined with prosodic features such as word boundary strength, word prominence and phone duration, which were themselves predicted by prosody models.

Phonetic models use a set of continuous parameters to describe intonation patterns observable in an F_0 contour [2]. An important goal is that the model should be capable of reconstructing F_0 contours faithfully when appropriate parameters are given. However, to make it functional, a phonetic model should also be linguistically meaningful. In fact, using certain functions, such as polynomial equations, to accurately represent F_0 contour is not a difficult task. What is more challenging is developing a model whose parameters are predictable from available linguistic information. To be more specific, the mapping from various linguistic factors, which could affect intonation, to the model parameters, or vice versa, is more critical.

2. Background

Below, we first classify intonation models into two major classes [3], and then provide details on a hybrid intonation model which is the focus of the current study.

2.1. Phonological Versus Phonetic Models

A *phonological* intonation model uses a phonological representation of F_0 . Such a representation is descriptive and discrete. It uses an inventory of abstract phonological categories, with each category having its own linguistic function. An example is the tonal tier of the ToBI labeling system [4]. ToBI specifies an inventory of tones: one set is used to mark accented syllables, while another set is used to mark phrase boundaries. Each tone marks a different type of accent or boundary.

A *phonetic* model is developed from acoustic (F_0) data. It attempts to describe F_0 movements and it is usually continuous in nature. Often, the description of F_0 movements is linked in some way to the linguistic level. The Tilt intonation model [2], for example, describes pitch accents and boundary tones via rising and falling quadratic functions. A pitch accent can be composed of a rising function, a falling function, or a rising followed by a falling function. Stretches of speech between

intonational events are described by straight-line interpolations. The amplitudes and durations of the rising and falling functions, combined with the position of the pitch accent/boundary tone in the (t, F_0) plane, together constitute the basis for the Tilt description of F_0 contours (5 continuously-valued parameters per event). The Fujisaki model [5] views an F_0 contour as the sum of a base F_0 value, phrase components and accent components (in the $\log F_0$ -domain). Phrase and accent components are generated by respectively passing impulse and step functions through second-order filters. The timings and amplitudes of the impulse and step functions constitute the phonetic representation of F_0 .

Traber's representation of F_0 [6] merely consists of samples of a smoothed and interpolated F_0 contour. It does keep track of the syllabic structure of the utterance, but it has no other links to the linguistic level.

2.2. MOMEL/INTSINT

The method used in this work can be viewed as a hybrid phonetic/phonological model [7]. It starts with a low-level phonetic analysis technique known as MOMEL (MODélisation de MELodie). Then a phonological description system, INTSINT (INTERNATIONAL Transcription System for INTonation), is derived from the results of phonetic analysis.

2.2.1. MOMEL

The MOMEL algorithm aims to analyze and synthesize F_0 curves automatically. An F_0 curve is modeled as the superposition of two components: a micro-prosodic component caused by the characteristics of the individual phonematic segments of the utterance and a macro-prosodic component reflecting the choice of intonation pattern for the utterance [1]. The MOMEL algorithm extracts the macro-prosodic component from the F_0 curve and models it as a series of quadratic splines.

There are four basic stages:

1. preprocessing of F_0

All values more than a given ratio higher than both their immediate neighbours are set to 0. This preprocessing has essentially the effect of eliminating erratic F_0 values.

2. estimation of target-candidates

The following steps are followed iteratively for each instant x .

- Within an analysis window centered on x , values of F_0 (including values for unvoiced zones) are neutralised if they are outside a range of thresholds and are subsequently treated as missing values.
- A quadratic regression is applied within the window to all non-neutralised values.
- All values of F_0 which are more than a given distance below the value of F_0 estimated by the regression are neutralised. Steps b and c are iterated until no new values are neutralised.
- For each instant x a target point is calculated from the regression equation. If the target is outside the

current analysis window or if the target lies outside the F_0 thresholds, then the target is treated as a missing value.

Steps b, c and d are repeated for each instant x , resulting in one estimated target point (or a missing value) for each original value of F_0 .

3. partition of candidates

The sequence of target candidates is partitioned by means of another moving window, in which the average value of the targets in the first half of the window is compared to the average value in the second half. The boundaries of the partition are then taken as those values which correspond to a local maximum for this distance and which is greater than the overall average value of the distances.

4. reduction of candidates

Within each segment of the partition, outlying candidates are eliminated. The mean value of the remaining targets in each segment is then calculated as the final candidate for that segment.

2.2.2. INTSINT

INTSINT describes intonation with a limited set of abstract tonal symbols, which is designed such that separate inventories of pitch patterns for different language are not required. The input to the INTSINT system is a series of target points, which is estimated from the acoustic low-level MOMEL modeling technique.

The abstract symbols defined to represent the target points are:

- **T** – Top
- **M** – Mid
- **B** – Bottom
- **H** – Higher
- **S** – Same
- **L** – Lower
- **U** – Up-stepped
- **D** – Down-stepped

Figure 1 shows the abstract symbols used in the INTSINT labeling system.

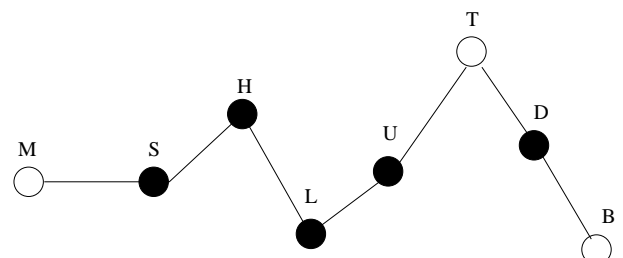


Figure 1: INTSINT labeling system.

Among these symbols, tones T, M, B are regarded as absolute tones, which refer to the speaker's overall pitch range.

Tones H, S, L, U, D are relative with respect to the value of the preceding target point. The relative tones are further distinguished between non-iterative H, S, L and iterative U, D tones – the latter can occur repeatedly, whereas the former cannot. An automatic coding scheme is used to relate the target points and the abstract symbols through a set of rules.

1. The highest and lowest target values in the utterance are coded as T and B, respectively.
2. The first target point, as well as any which follow a silent pause of more than a certain length in duration, is coded M (unless already coded T or B).
3. All other target points are coded with relative tones. A target which is less than a given threshold from the previous target is coded S. Otherwise it is coded H, L, U or D according to its configuration with respect to the preceding and following target points as in Figure 1. Where there is no relevant following target point the point is coded as either S, H or L depending on the previous target.
4. The statistical value of each category of target points is then calculated: for absolute tones the mean value is taken, for relative tones a linear regression on the preceding target is calculated.
5. Any target points originally coded H or L can be recoded as T, U, B or D if this improves the statistical model
6. Steps 4 and 5 are then repeated until no more points are recoded.

It is trivial to revert back to a MOMEL curve (which represents the intonation curve) from the INTSINT labels. Thus, if you have the INTSINT labels you can calculate the intonation curve.

3. Implementation

The MOMEL and INTSINT algorithms were implemented as modules into the Festival Speech Synthesis system [8]. The MOMEL algorithm was implemented straightforwardly as described in [7], but the implementation of the INTSINT labeling system required some more work.

The problem that arises with the implementation of the abstract phonological INTSINT labeling system stems from the fact that these labels are derived from a curve that was calculated from phonetic data. The temporal positions of the labels are defined by the inflection points of the quadratic splines calculated with the MOMEL algorithm. This implies that there is no direct connection between the INTSINT labels and the phonological data. Thus, the INTSINT labels must be time aligned with some form of linguistic feature that can be extracted from the orthographic data.

In this work the INTSINT labels were aligned with the syllables of the utterances. The method is as follows:

1. Add an INTSINT label to the mid point of each syllable in an utterance based on the rules as described in Section 2.2.2.
2. Calculate the resulting MOMEL curve from these labels.
3. Change the labels to minimize the difference between the MOMEL curves resulting from the syllable aligned labels and the original non-aligned labels.

The MOMEL curve calculated from the resulting phonologically aligned labels would then be the best approximation to the original MOMEL curve, which represents the intonation pattern of the utterance. From these labels *Classification and Regression Trees* (CART) [9] models were trained.

4. Experimental results

The dataset used in the experiments consisted of 152 isiZulu recordings. These recordings were phonetically transcribed by means of an automatic process and then checked by hand.

The experiments consisted of two parts:

1. Automatic intonation labeling as described in Sections 2.2 and 3.
2. Training and testing of intonation models build from the INTSINT-labeled data.

4.1. Automatic intonation labeling

The INTSINT labels were phonologically aligned to the middle vowel of the syllables in the utterances. Two experiments were conducted. In the first test all syllables were labeled with INTSINT markers, and in the second test only stressed syllables were labeled. The MOMEL curves resulting from these labels were then calculated and compared to the MOMEL curves calculated from the acoustic data. Table 1 shows the mean value and standard deviation of the root-mean-square (RMS) error calculated over the whole test set when compared to the original MOMEL curves.

| | All syllables | Stressed Syllables |
|--------------------|---------------|--------------------|
| Mean RMS error | 9.28 Hz | 20.67 Hz |
| σ RMS error | 2.01 Hz | 4.54 Hz |

Table 1: *INTSINT labeling accuracy obtained from labeling all syllables and only stressed syllables.*

Figure 2 shows an example of the MOMEL curve as calculated from the acoustics together with the F_0 curve, while Figures 3 and 4 give comparisons of the MOMEL curve as calculated from the acoustics versus the MOMEL curves aligned to all syllables and stressed syllables respectively.

4.2. INTSINT intonation models

Two CART intonation models were trained from the two sets of INTSINT labeled data. The *wagon* [10] CART training program was used for the training. A test set of 10% of the two labeled data sets was not included in the training data. A *stop* value of 50 and a *balance* value of 5 was chosen for the training of the trees.

Table 2 shows the intonation prediction results of a CART, trained on the all syllable labeled data, on the test set. The rows of the table gives the correct INTSINT labels and the columns gives the predicted INTSINT labels (for example, the *Same* label (S) was wrongly predicted as a *Bottom* label 18 times, and the prediction of S was correct in 22 of its 61 occurrences).

There were 301 syllables in total in the test set of which 106 were correctly predicted, giving a correct prediction rate of 35.22%.

| Label | B | D | H | L | M | S | T | U | Total | Correct | % Correct |
|-------|----|----|----|----|----|----|----|----|-------|---------|-----------|
| B | 35 | 0 | 0 | 5 | 3 | 0 | 0 | 0 | 43 | [35/43] | 81.395 |
| D | 6 | 2 | 2 | 15 | 2 | 5 | 0 | 2 | 34 | [2/34] | 5.882 |
| H | 3 | 4 | 4 | 5 | 4 | 7 | 1 | 1 | 29 | [4/29] | 13.793 |
| L | 15 | 6 | 3 | 8 | 2 | 10 | 0 | 3 | 47 | [8/47] | 17.021 |
| M | 0 | 0 | 0 | 0 | 18 | 0 | 1 | 0 | 19 | [18/19] | 94.737 |
| S | 18 | 5 | 1 | 10 | 2 | 22 | 1 | 2 | 61 | [22/61] | 36.066 |
| T | 0 | 0 | 3 | 2 | 4 | 3 | 14 | 0 | 26 | [14/26] | 53.846 |
| U | 5 | 5 | 3 | 13 | 4 | 9 | 0 | 3 | 42 | [3/42] | 7.143 |
| Total | 82 | 22 | 16 | 58 | 39 | 56 | 17 | 11 | | | |

Table 2: INTSINT labeling accuracy obtained from a CART trained on the all syllables data set.

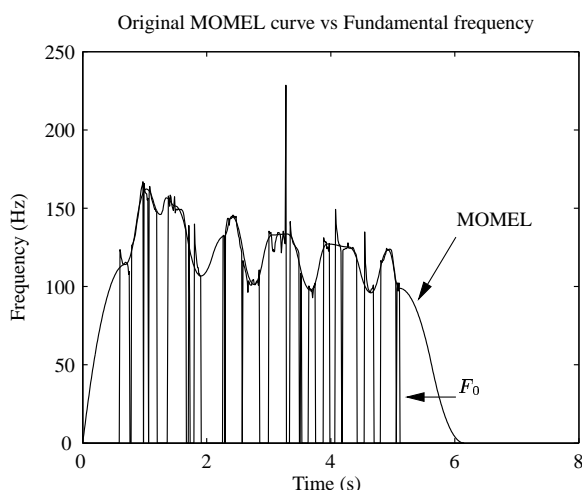


Figure 2: A comparison between the MOMEL curve as calculated from the acoustics and the F_0 curve.

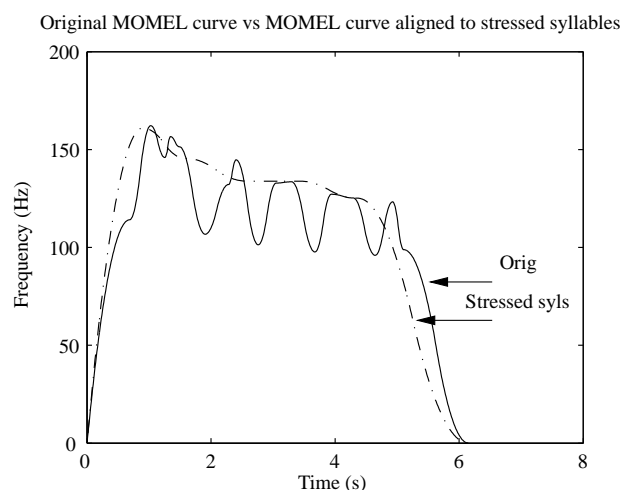


Figure 4: A comparison between the MOMEL curve as calculated from the acoustics and the MOMEL curve aligned to stressed syllables only.

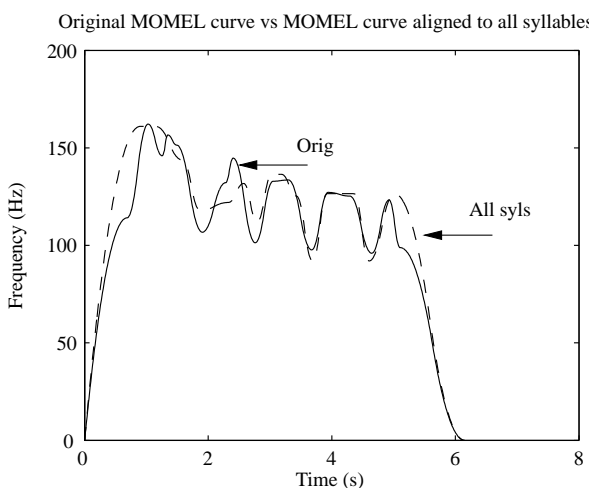


Figure 3: A comparison between the MOMEL curve as calculated from the acoustics and the MOMEL curve aligned to all syllables.

Table 3 shows the intonation prediction results of the CART, trained on only stressed syllables in the labeled data, as predicted on the test set. The convention is the same as in Table

2, except that – since only the stressed syllables can have an intonation event – there is an extra column for when there is no intonation label.

The same test set was used as in the previous experiment, thus of the 301 syllables a total of 227 (75.42%) were predicted correctly.

5. Conclusion

From Table 1 and Figures 3 and 4 it is clear that the MOMEL curve resulting from the INTSINT labels aligned on all the syllables in an utterance gives a better approximation to the true MOMEL curve as calculated from the acoustics. From Table 3 we can see that the CART model trained from only the stressed syllables labels produces a higher correct prediction percentage. Closer inspection reveals, however, that the majority of correct predictions were for syllables with no intonation events. If we were to exclude the results of the syllables with no intonation events, the correct prediction percentage would drop from 75.42% to 19.56%. This result is not unexpected: by labeling each syllable with an intonation event, a richer model (with more variation) is produced; as long as this model can track the contour of the true intonation data, it should be able to provide a more detailed fit to that data. Encouragingly, that

| Label | 0 | B | D | H | L | M | S | T | U | Total | Correct | % Correct |
|-------|-----|----|---|---|----|---|---|----|---|-------|-----------|-----------|
| 0 | 209 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 209 | [209/209] | 100.000 |
| B | 0 | 4 | 0 | 0 | 2 | 0 | 0 | 1 | 0 | 7 | [4/7] | 57.143 |
| D | 0 | 8 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 13 | [0/13] | 0.000 |
| H | 0 | 7 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 9 | [0/9] | 0.000 |
| L | 0 | 8 | 0 | 0 | 8 | 0 | 0 | 0 | 0 | 16 | [8/16] | 50.000 |
| M | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 13 | 0 | 15 | [0/15] | 0.000 |
| S | 0 | 8 | 0 | 0 | 6 | 0 | 0 | 0 | 0 | 14 | [0/14] | 0.000 |
| T | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 6 | 0 | 7 | [6/7] | 85.714 |
| U | 0 | 4 | 0 | 0 | 7 | 0 | 0 | 0 | 0 | 11 | [0/11] | 0.000 |
| Total | 209 | 41 | 0 | 0 | 31 | 0 | 0 | 20 | 0 | | | |

Table 3: *INTSINT* labeling accuracy obtained from a CART trained on the stressed syllables data set.

is the case for the models considered here.

Although the INTSINT model is phonological in nature, the actual labels are derived from a MOMEL curve, which in turn is derived from phonetic data. Unfortunately, this link is still fairly weak, as indicated by the relatively low CART classification accuracies that we have observed. This restricts the overall accuracy of the system, and the intonation produced (even for a stress language such as English) is not particularly natural. Work on additional ways of incorporating phonological and maybe even semantic data into the model is therefore required.

6. Acknowledgements

This work was supported by the CSIR *Information Society Technologies Centre*, South Africa, as well as the *Local Language Speech Technology Initiative* (LLSTI).

7. References

- [1] Hirst, D. , “Automatic analysis of prosody for multilingual speech corpora,” *Improvements in Speech Synthesis* (Keller, E. , G.Bailly, J.Terken, and M.Huckvale, eds.), Wiley, 2001.
- [2] Taylor, P. , “Analysis and synthesis of intonation using the Tilt model,” *Journal of the Acoustical Society of America*, vol. 107, no. 3, 2000, pp. 1697–1714.
- [3] Garrido, J. , *Modelling Spanish intonation for text-to-speech applications*. PhD thesis, University Autònoma de Barcelona, 1996.
- [4] Beckman, M. and Elam, G. , *Guidelines for ToBI labelling, version 3*. Ohio State University, March 1997. www.ling.ohio-state.edu/research/phonetics/E_ToBI.
- [5] Möbius, B. , Pätzold, M. , and Hess, W. , “Analysis and synthesis of German F0 contours by means of Fujisaki’s model,” *Speech Communication*, vol. 13, 1993, pp. 53–61.
- [6] Traber, C. , “F0 generation with a database of natural F0 patterns and with a neural network,” *Talking Machines: Theories, Models and Designs* (Bailly, G. , Benoît, C. , and Sawallis, T. R. , eds.), pp. 287–304, Amsterdam, The Netherlands: Elsevier, 1992.
- [7] Hirst, D. , Cristo, A. D. , and Espesser, R. , “Levels of representation and levels of analysis for intonation,” *Prosody : Theory and Experiment* (Horne, M. , ed.), Dordrecht, The Netherlands: Kluwer, 2000.
- [8] Black, A. , *Speech Synthesis in Festival*. Language Technologies Institute, Carnegie Mellon University, Pittsburgh,

USA, 1.4.1 ed., May 2000. A practical course on making computers talk.

- [9] Breiman, L. , *Classification and Regression Trees*. Pacific Grove, CA: Wadsworth and Brooks, 1984.
- [10] Taylor, P. , Caley, R. , Black, A. , and King, S. , *Edinburgh Speech Tools Library*. University of Edinburgh, Edinburgh, Scotland, 1.2 ed., June 1999. System Documentation.