THE TILT INTONATION MODEL

Paul Taylor

Centre for Speech Technology Research,
University of Edinburgh, 80 South Bridge, Edinburgh EH1 1HN, UK
http://www.cstr.ed.ac.uk Paul.Taylor@ed.ac.uk

ABSTRACT

The tilt intonation model facilitates automatic analysis and synthesis of intonation. The analysis algorithm detects intonational events in F0 contours and parameterises them in terms of the continuously varying *Tilt* parameters. We describe the analysis system and give results for speaker independent spontaneous dialogue speech. We then describe a synthesis algorithm which can generate F0 contours given a tilt parameterisation of an utterance. We give results showing how well the automatically produced contours match natural ones. The paper concludes with a discussion of the linguistic relevance of the tilt parameters and show that this is both a useful and natural way of representing intonation.

1. INTRODUCTION

The tilt intonation model is designed to facilitate automatic intonational processing for speech technology applications. The model represents intonation at a phonetic level as a sequence of parameterised intonational *events*. From such a representation, it is possible to encode the linguistically relevant information in an F0 contour, and then recreate the original F0 from this coding.

This paper gives a brief overview of the model itself, and the accompanying automatic analysis and synthesis algorithms which can map from and to F0 contours. A fuller description of the model and its underlying philosophy can be found in Taylor [11]

2. MODEL OVERVIEW

2.1. Intonational Events

In the Tilt model, intonation is characterised by a sequence of phonetic *intonational events*. By this we mean that events occur every so often and do not necessarily abut as in segmental representations.

Unlike segmental representations of speech, where one phone follows another, in intonation events occur every so often and need not abut.

There are two kinds of events, pitch accents, **a**, and boundary tones, **b**. Following from the standard type of representation in autosegmental phonology [5], one can think of the sequence of events as being an autonomous linguistic tier or level, in which each event is *associated* with a syllable.

Each event has a rise and fall component which can vary in size. Some events have a zero rise or zero fall component indicating that they only have a fall or only have a rise respectively. The "middle" of the event is defined as the end of the rise component or start of the fall component. Each event is characterised by the *tilt* parameters, which fully describe its f0 shape with a number of linguistically useful variables. Note that both pitch accents and boundary tones are characterised using the same set of parameters.

Amplitude: the size of the F0 excursion of the event.

Duration: in seconds from the start to the end of the event.

Tilt: a dimensionless parameter describing the shape of the event. Tilt is calculated from the relative sizes of the rise and fall components in the event. A value of +1 indicates the event is purely a rise, -1 indicates it is purely a fall. Any value between says that the event has both a rise and fall component, with a value of 0 indicating they are the same size.

F0 position: the F0 distance from the baseline (usually 0Hz) to the middle of the event.

Time position: where the event is located in time. There are two standard ways of describing this. When dealing with intonational representations in isolation, this parameter is often used to represent the time from the beginning of the utterance to the middle of the event. Alternatively this can be used to represent the *relative* time with respect to the associate syllable. The start of the vowel is normally taken as the reference point in the syllable and this parameter reflects the distance from that point to the middle of the event.

Section 4 describes how these parameters can be mapped into F0 contours. The next sections describe how these parameters can be automatically extracted from speech.

3. AUTOMATIC ANALYSIS

The model has been tested on three English databases. Although the model has primarily been used for English so far, the model has also been used for Korean and Japanese.

DCIEM Maptask This is a corpus of 216 dialogues collected by Canada's Defence and Civil Institute of Environmental Medicine (DCIEM)[1]. The speech consists of fully spontaneous dialogues and contains many disfluencies. The database has a particularly

Features	% c	% a	% major c	% major a
F0 and energy	57.7	26.6	69.6	46.3
Norm F0, energy	61.7	33.6	73.0	51.7
Norm F0, energy + d	65.6	43.8	76.7	56.1
Norm F0, energy + d + a	72.7	47.7	81.9	60.7

Table 1: Performance for different feature sets on the DCIEM corpus

Dataset	% c	% a	% major c	% major a
DCIEM	72.7	47.7	81.9	60.7
Radio News 1	68.9	49.2	n/a	n/a
Radio News 2	69.4	49.7	79.4	59.3
Switchboard	60.7	35.1	71.5	47.4

Table 2: Performance for different data sets

rich variety of types of utterance, e.g. it contains many questions, instructions, statements, confirmations back-channels etc. A subset of 25 dialogues (about 2 hours of speech) from a number of speakers was used here.

Boston Radio News Corpus This is a corpus of news reader speaker collected at Boston University [7]. A subset of 34 stories of about 48 minutes of one speaker was used for experiments here.

Switchboard Switchboard is a corpus of about 2000 spontaneous speech dialogues collected live over the US telephone network [4]. Experiments reported here are based on a 1 hour subset, chosen (by researchers at ICSI, Berkeley) to achieve maximum acoustic and phonetic variability across the corpus. Within this hour there are about 100 different speakers from all parts of the United States. 50 minutes were used for training and 10 for testing.

Hand Labelling The databases were hand labelled to produce intonational transcriptions. The labellers were instructed to locate pitch accents and boundaries within each utterance, in accordance with the intonational event model described above. The size of events varies considerably, and it is felt that this in some way is related to the linguistic importance of the event, in that large events will carry more linguistic information than smaller ones. Preliminary experiments had shown that the event detector was less successful at recognising small events and hence the small accents were marked with a diacritic *minor* which allowed the testing programs to measure to what degree mis-recognition on this class occurred.

Comparing Transcriptions and Labelling Consistency To compare intonational transcriptions it is not enough to use symbol comparison algorithms such as the dynamic programming label alignment technique normally employed for measuring word accuracy in speech recognisers. This is because virtually any intonational transcription will align with any other as they are usually all alternating sequences of event and non-event. Hence we adapted the dynamic programming label alignment technique to have the additional constraint that two events had to overlap by 50% in time to be classed as the same.

Using this technique, we measured the amount of agreement in our hand labellers by having all of them label a portion of the test set and then cross comparing transcriptions. The pairwise scores for all the labellers were 81.6% correct with 60.4% accuracy. When ignoring the accents marked with the minor diacritic, the agreement is 88.6% correct with 74.8% accuracy, showing that a large number of errors were caused by minor accents. Looking at the types of events separately, the agreement for pitch accents is 81.6% correct 58.1% accuracy and the agreement for boundaries is 83.3% correct and 64.1% accuracy.

3.1. Event Detection

The first stage in automatic analysis is to find the events from the waveform. We achieve this by using a HMM based recogniser which effectively segments an utterance into event and non-event sections.

Waveforms are parameterised into F0, energy and their delta and delta-delta (acceleration) complements.

A three state left-to-right continuous density HMM was used to model accents, boundaries, non-event speech and silence. Each model was training using the standard Baum-Welch training algorithm, with the hand labelled data being used to provide segmentation boundaries during training. Embedded training of the type used in standard speech recognition, where the HMMs effectively decide their own segmentation during training, was also tried, but produced poor models in comparison. The training algorithm was used to iteratively increase the number of Gaussian components in each mixture. Eight to sixteen mixture components gave the best results.

The trained HMMs can be used to detect events by running them with the Viterbi search algorithm and a bigram intonational event sequence model over the parameterised data of a test utterance.

Table 1 shows results for a number of different feature sets. Using the delta coefficients increased recognition performance considerably, showing that the trajectories of F0 and energy is an important indicator of event presence. Table 2 shows event detection results for the 3 datasets. Results are shown for the whole test sets and for the major events only. It is clear that major accents are recognised better, with improved accuracy and correct scores for all the data sets. Switchboard performance is significantly worse than the other two datasets, and this is thought to be mostly due to the relatively poor acoustic conditions of this database which result in pitch tracking errors and general noise in the recordings.

While there is obviously room for improvement for all the datasets, it should be remembered that the DCIEM and Switchboard results are for speaker independent fully spontaneous conversational speech, with no prior information (such as a segmental transcription) being available. In light of this, the results are encouraging.

3.2. Tilt Parameterisation

The next stage in the process is to derive the tilt parameters for each of the events found by the event detector. This stage uses a algorithm which examines each event and fits rise or fall shapes by minimising the error between the original contour and the fitted shape. The result of this process is that each event is now described as a rise shape, a fall shape or a rise followed by a fall shape. This parameterisation produces a representation in terms of a previous model, known as the rise/fall/connection (RFC model) [10]. The tilt model can be thought of as a further stage to the RFC model, in that it takes RFC parameters and from them produces a more usable, higher level and compact intonational representation. The RFC model and shape fitting algorithm is more fully described in Taylor [9] and [10] .)

The fitting algorithm produces a rise amplitude (A_{rise}), a rise duration (D_{rise}), a fall amplitude (A_{fall}) and a fall duration (D_{fall}). While these parameters accurately encode the F0 shape of the event, they are not ideal as the amplitudes and durations of the rise and fall components interact strongly with one another. A further set of transformations produces the tilt parameters as follows:

Amplitude tilt is given by

$$tilt_{amp} = \frac{|A_{rise}|-|A_{fall}|}{|A_{rise}|+|A_{fall}|} \quad (1)$$

and *duration tilt* is given by

$$tilt_{dur} = \frac{D_{rise}-D_{fall}}{D_{rise}+D_{fall}} \quad (2)$$

Empirical evidence has shown that these parameters are highly correlated to the extent that a single parameter can be used for both amplitude and durational tilt. This single value is calculated from the averages of both:

$$tilt = \frac{|A_{rise}|-|A_{fall}|}{2(|A_{rise}|+|A_{fall}|)} + \frac{D_{rise}-D_{fall}}{2(D_{rise}+D_{fall})} \quad (3)$$

amplitude and duration, are calculated in terms of the sum of the magnitudes of the rises and falls.

$$A_{event} = |A_{rise}| + |A_{fall}| \quad (4)$$

$$D_{event} = D_{rise} + D_{fall} \quad (5)$$

F0 position and time amplitude can be calculated directly.

4. AUTOMATIC SYNTHESIS

The Tilt parameters can be used to produce an F0 contour by first converting them back into RFC parameters and then using equations to generate actual contours.

The following equations produce RFC parameters from Tilt parameters:

$$A_{rise} = \frac{A_{event}(1+tilt)}{2} \quad (6)$$

$$A_{fall} = \frac{A_{event}(1-tilt)}{2} \quad (7)$$

$$D_{rise} = \frac{D_{event}(1+tilt)}{2} \quad (8)$$

$$D_{fall} = \frac{D_{event}(1-tilt)}{2} \quad (9)$$

which can be converted to F0 values as follows:

$$\begin{aligned} f_0(t) &= A_{abs} + A - 2.A.(t/D)^2 \quad 0 < t < D/2 \\ f_0(t) &= A_{abs} + 2.A.(1 - t/D)^2 \quad D/2 < t < D \end{aligned} \quad (10)$$

Between events, straight line interpolation is used to produce F0 values.

Representation	raw		smooth	
	F0 rmse	F0 ρ	F0 rmse	ρ F0
hand labelled tilt	14.58	0.647	7.14	0.829
automatic tilt	15.25	0.644	7.51	0.833

Table 3: Accuracy figures for Tilt synthesis. The first row shows the synthesis accuracy when the events are labelled by hand and the second shows the results for automatically labelled accents.

The accuracy of the tilt encoding and synthesis routines can be measured by performing a resynthesis experiment, where a natural F0 contour is compared to its resynthesized equivalent. Table 4 gives results for two types of synthesis test, RMS error and correlation on the DCIEM test set. Raw F0 contours contain glitches and segmental perturbations not modelled by the Tilt model, so a set of smoothed contours free of such effects were also created and compared.

5. DISCUSSION

In this section we discuss the above results and argue that the Tilt model is a more appropriate and powerful model for automatic speech processing purposes than models such as the tonal component of ToBI [8].

The high accuracy of the synthesis component is mainly due to the parameters of the model being directly interpretable in a synthesis sense. Application of the tilt and RFC equations will generate an F0 contour and hence in some sense no separate synthesis algorithm is required - the algorithm is part of the model itself. In this way the model is similar to the Fujisaki model [3] from which contours can also be directly synthesized.

The analysis process helped by the fact that only one type of pitch accent and boundary tone are used and so the fine distinctions of ToBI do not need to be made. Once the accent is located, the curve fitting equations will determine the RFC and then tilt parameters in a straightforward way.

When designing a model expressly for automatic analysis and synthesis purposes, one can be open to the criticism that the model goes too far in facilitating these needs at the expense of producing a linguistically useful description or parameterisation. Hence the usefulness of the tilt description system needs to be examined to show that it is not merely a data-reduction type of coding of an F0 contour.

In fact, when one follows this path and tries to examine the merits of *any* intonational description system (such as this model, Fujisaki or ToBI) one soon finds that this is extremely troublesome. The intonational literature is often vague on this issue and arguments advocating systems typically rely on them fixing theoretical problems with other systems or "naturally describing" the observed shapes in a set of contours. A large part of the problem lies in the semantic nature of intonation in that it is very difficult to say what sort of semantic effect a change in F0 produces, whereas in segmental phonology a change in voicing which converts a /p/ to a /b/ gives a more obvious semantic change, lending weight to the argument that /p/ and /b/ are two separate entities from a linguistic point of view. The minimal pair types of tests which phonologists have used to find the linguistically contrasting classes of sounds in a language (the phonemes) is difficult to apply to intonation, which leads to great difficulty in designing a intonational inventory of sounds. In summary, it is fair to say that the linguistic justification for any existing intonation systems are weak and hence we will attempt to describe the merits of the Tilt model on its own terms instead of having to justify it directly with respect to the tonal part of the ToBI model, which although maybe widely used, has never in any case been justified convincingly.

The amplitude tilt parameter correlates well with the perceived prominence of a boundary or pitch accent. While the relationship is certainly not linear, in that effects such as pitch range need be taken into account, it is usually the case that increasing amplitude increases prominence. Duration seems to bear little direct semantic information, but rather seems to be a function of the segmental structure of the associate syllable. Syllables with short voiced regions will typically have short durations, ensuring that the most important part of the F0 movement occurs within a voiced region of speech. Time position (when measured relative to the syllable) carries important information and the same F0 shape can have a quite different effect depending on whether it occurs early or late in the syllable (this effect is well documented, e.g. [6]).

The parameters of the Fujisaki model bear significant resemblance to those just discussed. However, in the Fujisaki model every accent has more or less the same shape, differing only in duration or amplitude. this was found to be too restrictive for English where a great variety of pitch accent shapes occur. Hence the need for the tilt parameter which neatly encodes pitch accent shape in a single number. F0 position carries little information on its own, it is important in modelling the global shape of the contour, specifically effects such as down-drift and down-step.

While the tilt model shares many features of ToBI (most notably that intonation is modelled by sequences of events), it differs significantly in that ToBI uses discrete classes to model event variation whereas the tilt model uses continuous parameters. We have argued elsewhere [11] that the distinctions in accent type made in ToBI are problematic, in that distinctions are forced between very similar (if not identical) accents such as H* and L+H*, while many accents which are obviously different are lumped together in a single H* class. In the Tilt model, difficult categorical decisions are avoided, firstly because to make such decisions would lead to errors and secondly because there is little real evidence that such categories even exist.

Ultimately, the most important test of a model such as this is how well it performs in speech processing applications. Some benefits of the model are indirect, such as the ability to easily build an automatic analyser, which makes the model attractive from the point of view of saving time and effort in database labelling. But in addition the model has proved useful in synthesis and recognition applications. Dusterhoff and Black [2] describe a method for using CART to generate tilt parameters and then F0 contours from high level information in a text-to-speech system. Wright and Taylor [12] describe a system for automatically recognising the dialogue act of an utterance from an analysis of its intonation. While we can't argue that other intonational models couldn't perform these tasks also, we have at least proved that the tilt model can, and along with the general ease of analysis and synthesis we believe the model is a useful tool in facilitating automatic intonational processing.

Acknowledgements

Paul Taylor gratefully acknowledges the support of the UK Engineering and Physical Science Research Council, grant number GR/L53250.

6. REFERENCES

1. Ellen G. Bard, Catherine Sotillo, Anne H. Anderson, and M. M. Taylor. The DCIEM map task corpus: Spontaneous dialogues under sleep deprivation and drug treatment. In *Proc. of the ESCA-NATO Tutorial and Workshop on Speech under Stress, Lisbon*, 1995.
2. Kurt Dusterhoff and Alan Black. Generating intonation contours for speech synthesis using the tilt intonation theory. In *ESCA workshop on Intonation: Theory Models and Applications*, 1997.
3. Hiroya Fujisaki and S. Ohno. Comparison and assessment of models in the study of fundamental frequency contours of speech. In *ESCA workshop on Intonation: Theory Models and Applications*, 1997.
4. J. Godfrey, E. Holliman, and J McDaniel. SWITCHBOARD: Telephone speech corpus for research and development. In *ICASSP92*, pages 517–520, 1992.
5. John Goldsmith. *Autosegmental and Metrical Phonology*. Blackwell, 1989.
6. Klaus J. Kohler. The perception of accents: Peak height versus peak position. In Klaus J. Kohler, editor, *Studies in German Intonation*. Universität Kiel, 1991.
7. M. Ostendorf, P. J. Price, and S. Shattuck-Hufnagel. The Boston University radio news corpus. Technical Report ECS-95-001, Boston University, March 1995.
8. K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. ToBI: a standard for labelling English prosody. In *Proceedings of ICSLP92*, volume 2, pages 867–870, 1992.
9. Paul A. Taylor. Automatic recognition of intonation from f0 contours using the rise/fall/connection model. In *Proc. Eurospeech '93, Berlin*, 1993.
10. Paul A. Taylor. The rise/fall/connection model of intonation. *Speech Communication*, 15:169–186, 1995.
11. Paul A. Taylor. Analysis and synthesis of intonation using the tilt model. *forthcoming*, 1998.
12. H. Wright and P. A. Taylor. Modelling intonational structure using hidden markov models. In *ESCA workshop on Intonation: Theory Models and Applications*, 1997.