Using Decision Trees within the Tilt Intonation Model to Predict F0 Contours

Kurt E. Dusterhoff, Alan W. Black, and Paul Taylor

Centre for Speech Technology Research, University of Edinburgh,
80 South Bridge, Edinburgh EH1 1HN
http://www.cstr.ed.ac.uk
email: {*kurt, awb, pault*}*@cstr.ed.ac.uk*

ABSTRACT

This paper presents an intonation generation system for use in a text-to-speech synthesis system. The intonation generation system uses classification trees to predict intonation event location and regression trees to predict parameters relating to the F0 shape for the predicted events. The decision trees model intonation within the Tilt intonation model, which provides a parameterized description of fundmaental frequency and an intuitive labelling scheme. The event location trees predict an event class (e.g. accent, boundary, none) for each syllable in an utterance based on local and global context (e.g. stress, phrasing, part of speech). The parameter prediction trees then provide the parameterized description of each intonation event based on similar context features. Informal results of the full system are presented together with results for the individual components.

1. INTRODUCTION

Most of the currently available speech synthesizers have some sort of intonation generation module. These range from using a simple declining F0 over a phrase to more complex statistical models of specific speech types (e.g. [3]). This paper presents an intonation generation system which uses classification and regression trees to predict the location and fundamental parameters of intonation events within the Tilt intonation model [8].

The Tilt model offers a parameterized description of F0 contours and an intuitive labelling system. The first stage of the intonation generation process is to predict intonation event location. This is acheived by building a classification tree which gives an intonation class for each syllable (e.g. accent, boundary, none) based on localized contextual information. Then, for each intonation event type (e.g. accents), the Tilt parameters are modelled using regression trees. These predicted parameters are used to determine the F0 shape of each predicted intonation event. Figure 1 illustrates the steps involved in generating an F0 contour using this method.

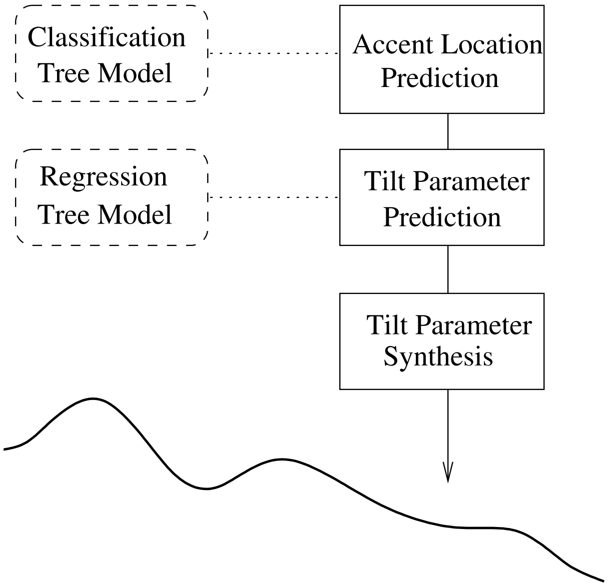

Figure 1: Illustration of F0 generation process

2. SPEECH DATABASES

The databases modelled cover three basic speech types: news commentary, isolated sentences, and instructional text. The models were built using the same methods for each of the databases described below.

The news commentary database is a portion of the Boston University Radio News Corpus [6], speaker F2B. The database consists of 114 paragraphs of news commentary (approximately 45 minutes) as delivered by a female speaker of American English. The database is labelled with segment, syllable, and word boundaries, and includes lexical stress markings. It is also labelled with intonation labels based on the Tilt intonation model [8].

The isolated sentence (TIMIT) database is a set of 450 phonetically balanced sentences, of which ten percent are questions. These sentences are spoken by a male American English speaker (KD-t), and are annotated in the same manner as the F2B database.

The instructional text database consists of 43 excerpts of text which describe exhibits in a museum (approximately 40 minutes). The text is read by a female Scottish English speaker (FHL) and is labelled with word boundaries and intonation labels. In the FHL database, syllable boundaries are estimated, segmental boundaries are not used, and stress is taken from dictionary entries - and is therefore approximate.

Each of the databases is tested in isolation. Cross-data training was avoided so that each individual speaker and style could be modelled without the difficulties caused by the different data types. Tilt parameter modelling was performed on each of the databases. The F2B database was also modelled for intonation event location.

3. EVENT LOCATION PREDICTION

The inventory of Tilt event types is quite simple. In its simplest form two types are supported, *a* for accents, and *b* for boundaries. Thus, unlike other intonation theories, in Tilt, accent placement is a matter of finding appropriate syllables to affix accent and the choice of the type of accent is effectively defined by the assignment of Tilt parameters.

To predict accents and boundaries we use a simple CART tree. We extract feature information for each syllable in a database. Not all of our databases used here for testing have the same labels but the following selection of features is typical:

- A single feature identifying if the syllable is the last stressed syllable in the phrase.
- The position/type of the syllable (single, initial, final or mid) for this and the preceding and following syllables.
- The stress value for this and the surrounding two syllables.
- The break value (minor, major or none) on the related word (and one before and after).
- The part of speech (and simple content/function marker) on the related word (and one before and after).

Through experimenting we found that often a single complex feature could give better overall performance than a set of features even if they formally contained the same information, thus "last stressed syllable in phrase" was a stronger feature then stress, position in word and position in phrase. Similarly "first stressed syllable in word" was often found most predictive.

On the F2B database we achieved the following results on held out data for predicting accent placement. A

	none	accent	total	% correct
none	686	114	800	85.75
accent	102	341	443	76.98

Table 1: Accent prediction on F2B, overall 82.62% correct

further study we made was to combine the output of the decision tree with an n-gram of accent/non-accents and the use of a viterbi decoder to find the optimal selection of accentness globally over the whole utterance. However the additional information from the n-gram was never sufficient to improve the overall results even though it was clear it did solve some errors. On other databases this may make more of a difference. A trivial side effect of these tests allowed us to play with accent over/under prediction which was found to help marginally.

A similar CART was used to predict *b* events but the results are basically trivial as there was a one to one relationship between prosodic boundaries and the *b* event.

4. TILT PARAMETER PREDICTION

The intonation prediction experiments consist of a basic four step process. First, information about each utterance in a database is extracted. Regression trees are built for each Tilt parameter of each Tilt event type. These models are then used to generate Tilt descriptions of the fundamental frequency for each utterance.

4.1. Feature Extraction

For each utterance, a variety of information is extracted which may assist in modelling F0. The information was divided into five classes. These classes are described below.

The lexical stress (0 or 1) of a given syllable and the two syllables on either side make up the first class. The second class concerns the position of a given syllable within a phrase. Following [1], the features extracted in this class are:

- The number of syllables from the previous event (i.e. accent or boundary) and to the next.
- The number of syllables from the previous major phrase break and to the next.
- The number of stressed syllables from the previous major phrase break and to the next.
- The number of accented syllables from the previous major phrase break and to the next.
- The phrase break index (0-4) of the syllable in a window of two before and two after.

The third feature class contains information about the composition of the syllable and its place in a word. The composition-related features are the length of the onset and rhyme and a classification of onset and coda, following [10] and [7].

The fourth class is simlar to the lexical stress feature, but relates to intonation events. Two features are used here, one each for accent and boundary. A value is extracted (again 0 or 1) if the syllable is linked to an accent or a boundary.

The final class is more suprasegmental in nature than the other classes. Rather than being based on syllables, the features in this class are the event types of the event linked to a syllable, and the two events on either side, regardless of their location in terms of syllables. This view of the data was necessary because intonation events do not occur on every syllable, and a syllable-based window does not always contain information about any events.

4.2. Building Regression Trees

The regression trees used for the experimentes were built using the Wagon classification and regression tree tool [9] which uses standard CART techniques [4].

The trees consist of questions about features which are used to predict a particular parameter. Each node of the tree contains a question, a sub-tree for "yes" answers, and a sub-tree for "no" answers. The leaves of the trees contain mean and standard deviation values for the data points which are classified by the answer path required to reach a given leaf.

As noted above, the data is divided by both accent and parameter type. Thus, for example, there is a separate tree for the peak position parameter for accents from the peak position tree for rising boundaries. For each tree needed (one for each parameter for each accent type), a tree is begun by finding the best question that partitions the data such that the standard deviations within partitions is minimal. The tree is grown by continuing such question selection until a specified minimum number of data points is reached. The algorithm is greedy, in that it selects the best partition and question at a given time, rather than testing all possible combinations, which is computationally prohibative.

Previous experiments which have used this techniques ([5] [2]) have also benefitted from minor hand-optimising of the feature set for noise reduction. However, it is unclear whether the resulting, nominal improvement in correlation (2 percentage points) over a large corpus has any real effect.

5. RESULTS

The original results of experiments using regression trees to predict Tilt parameters showed promise for generating F0 for speaker F2B [5]. Their results show that this method produces results comparable to other similar studies using the same database. All of the results detailed in this section are for prediction based on hand-labelled data. Therefore, the initial results from Dusterhoff and Black [5] (RMSE of 33.9Hz and Correlation of 0.57) which uses similar data will be the departure point for comparisons.

Tables 3 through 5 show how the results of the intonation generation method described in the previous section compare with the original intonation of the databases. Each of the results shows a target and at least one experimental result. The targets are the result of comparing the smoothed F0 contours from which the original Tilt parameters are extracted with the F0 contours generated by the original Tilt parameters. In other words, this score is the score that would be given if the Tilt parameter prediction were 100% correct on all counts. Because the databases represent different voice types, dialects, and genders, it has been useful to consider the RMSE results in terms of their relation to the standard deviation of F0 in order to compare them with each other. Thus, a 34Hz RMSE may *look* like a large error, but if it is acheived on a voice with a large standard deviation (e.g. 53Hz), the error is relatively low. For the female speakers, the target RMSE score is roughly one-third of the standard deviation of F0. For the male speaker, the target RMSE is approximately one-seventh of the standard deviation (see table 2).

A base result, arrived at using the methods described in section 4, is shown in table 3.

As table 4 shows, it is easier to predict the intonation of a database when it is for mostly declarative, isolated sentences that are spoken by a male speaker. The KD-t results are interesting in a number of areas. First, the target for KD-t is similar to the target for F2B, in terms of correlation. Therefore, we believe that the Tilt descriptions and the related tools handle the male and female voices equally well.

Speaker	Mean F0	σ F0	Target RMSE
F2B	163.5	42.2	14.5
KED	126.9	27.9	3.9
FHL	210.5	31.8	12.5

Table 2: F0 and Target RMSE information for three speakers

	Original Values	Predicted Values
RMSE	14.5	34.3
Correlation	0.93	0.6

Table 3: Comparison of F0 contours generated from original and predicted Tilt parameter values for F2B

	Original Values	Predicted Values
RMSE	3.9	9.1
Correlation	0.94	0.74

Table 4: Comparison of F0 contours generated from original and predicted Tilt parameter values for KD-t

The difference in RMSE targets reflects the difference in the speakers' pitch ranges. For KD-t, who has less natural variation in F0, it is necessary to prevent large variations in the generated contours, as they will likely sound out of place. This restriction was not true of F2B, who had a naturally larger range of possible F0 values.

The results of the parameter prediction experiment on this database show that, given a corpus of consistent data, it is possible to achieve a high correlation between the original and synthetic F0, while also keeping the RMSE down.

Table 5 shows that some databases are more difficult than others to model. The target correlation is noticeably lower than that of the other two speakers. This suggests that perhaps the labels are not of as high a quality, or perhaps that there is more movement in the non-event (connection) portions of the original F0, lowering the correlation score even if the events are accurately regenerated. Regardless of the cause for the lower target, it is important to recognize that a lower target will likely correspond to a lower result. Therefore, the results for FHL, while lower than for F2B and KED, are comparable to F2B's results. The resulting RMSE is less than twice the target (as compared with almost 2.5 times for F2B) and only slightly more than one-third σ F0. While these comparisons do not have any inherent meaning in themselves, they show that the FHL results are in the same range of success as the F2B results, while remaining lower than the KD-t results.

	Original Values	Predicted Values
RMSE	12.5	21.1
Correlation	0.87	0.53

Table 5: Comparison of F0 contours generated from original and predicted Tilt parameter values for FHL

6. DISCUSSION

This paper has presented an approach to generating fundamental frequency contours using decision trees within the paradigm of the Tilt Intonation Model. The event location prediction uses local information about syllables, words, and phrasing. The Tilt parameter prediction uses similar contextual data, but also exploits information about the phonetic content of syllables and sub-syllable constituents (e.g. rhymes, codas).

The integration of the event location and parameter prediction processes has been completed for one of the databases discussed, and is being tested on other speech databases for different tasks and speakers. This integrated approach produces reasonable intonation, and is better or at least equal to other approaches which we have investigated. We believe that this method is a promising step forward for full intonation prediction.

7. Acknowledgements

We gratefully acknowledge the support of the UK Engineering and Physical Science Research Council (grants GR/L53250 and GR/K54229) and Sun Microsystems.

REFERENCES

[1] A. Black and A. Hunt. Generating F_0 contours from ToBI labels using linear regression. In *ICSLP 96*, Philadelphia, Penn., 1996.

[2] A.W. Black, K. Dusterhoff, and P. Taylor. *Using the Tilt intonation model for speech synthesis: a data driven approach.* in press.

[3] A.W. Black, P. Taylor, and R. Caley. *The Festival Speech Synthesis System: system documentation.* The Centre for Speech Technology Research, University of Edinburgh, 1.3 edition, 1998. http://www.cstr.ed.ac.uk/projects/festival/manual-1.3.0/festival_toc.html.

[4] L. Breiman, J. Friedman, and R. Olshen. *Classification and Regression Trees*. Wadsworth and Brooks, Pacific Grove, CA, 1984.

[5] K. Dusterhoff and A. Black. Generating F_0 contours for speech synthesis using the Tilt intonation theory. In *Proceedings of ESCA Workshop on Intonation*, Athens, Greece, 1997.

[6] M. Ostendorf, P. Price, and S. Shattuck-Huffnagel. The Boston University Radio News Corpus. Technical Report ECS-95-001, Electrical, Computer and Systems Engineering Department, Boston University, 1995.

[7] P. Prieto, J. van Santen, and J. Hirschberg. Tonal alignment patterns in Spanish. *Journal of Phonetics*, 23(4):429–451, 1995.

[8] P. Taylor. The Tilt intonation model. In R. Mannell and J. Robert-Ribes, editors, *Proc. ICSLP 98*, volume 4, pages 1383–1386, 1998.

[9] P. Taylor, R. Caley, and A.W. Black. *The Edinburgh Speech Tools Library*. The Centre for Speech Technology Research, University of Edinburgh, 1.0.1 edition, 1998. http://www.cstr.ed.ac.uk/projects/speechtools.html.

[10] J.P.H. van Santen and J. Hirschberg. Segmental effects on timing and height of pitch contours. In *ICSLP*, volume 2, pages 719–722, Yokohama, 1994.