GENERATING F_0 CONTOURS FOR SPEECH SYNTHESIS USING THE TILT INTONATION THEORY

Kurt Dusterhoff and Alan W Black

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

ABSTRACT

This paper presents a method for generating F_0 contours for a speech synthesis system using the Tilt intonation theory ([10], [9]). The Tilt theory offers an abstract description of natural F_0 contours which may be derived automatically from natural speech. Given a speech database labelled with Tilt events, this paper shows how that data may be used to train a model which can adequately predict Tilt parameters from features available in a text to speech system and hence produce natural sounding F_0 contours. After a short description of the Tilt theory, the database used and the necessary features used to generate the parameters are presented. For comparison, this work is contrasted with a previous similar experiment on the same database using the ToBI intonation labelling system [2]. The Tilt method not only produces better results (RMSE 32.5 and correlation 0.60) but as it offers automatic labelling of data, it promises the ability to more easily train from general speech databases.

1. BACKGROUND

In the task of rendering natural sounding speech from raw text, one of the many tasks is generating natural sounding intonation. A number of intonation theories have been utilised in various systems to try to do this task. As the quality of speech synthesis improves, a greater demand is put on the intonation system to produce more varied intonation tunes. Because of this demand, and the requirement to quickly and easily add new voices and new accents to our systems, intonation systems should be trainable, where appropriate, from natural speech data.

ToBI [7] offers a well-defined intonation phonology for labelled speech. It is probably still the most widely available standard labelling system. The ToBI labelling system itself does not define a mechanism to go from the labels to an F_0 contour, or the reverse. However there are both hand written rule systems (e.g. [1]) and statistically trained methods (e.g. [2]) which do this task.

The Tilt intonation theory has been shown to be a good representational method for natural F_0 contours [10] but prior to the work presented here it has not been shown that Tilt parameters could be predicted reliably from text input. Tilt and ToBI typify two major classes of intonation system. Tilt comes from a data-driven approach attempting to form an abstraction of the natural contour but maintaining mechanism to recreate it. ToBI takes a more linguistic or phonological approach specifying a small set of discrete labels which identify the intonational space of accents and tones.

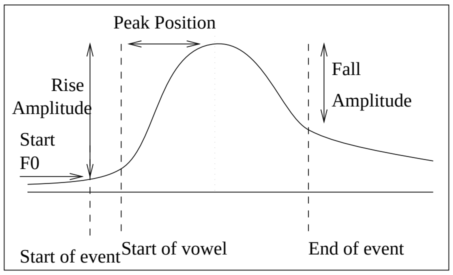
There are other intonation theories but we highlight ToBI as it has been used in a similar experiment to the one described below on the same database and hence offers a chance to directly compare these theories on the same task.

In order for an intonation theory to be suitable to be used in a speech synthesis system it must be possible to predict its parameters adequately from the information that is available from the raw text (or information that is automatically derivable from such text). Thus it is not just a good representation of the natural F_0 that is required but we must also be able to predict its parameters.

2. TILT

A Tilt labelling for an utterance consists of an assignment of one of four basic intonational events: pitch accents, boundary tones, connections, and silence (labelled a, b, c, sil). Each of the events includes a number of continuous parameters. 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 a *duration* (seconds), an absolute *amplitude* (Hertz), the *peak position* at which the rising portion of the event stops and the fall begins (measured in seconds from the start of the vowel), and a *tilt* value representing the "tilt" of the accent (described below). Figure 1 shows how the parameters relate to an example pitch accent.

The tilt parameter represents the amount of fall and rise in the accent. The starting F_0 of an event acts as a point from which all other calculations may be made. The absolute amplitude from the starting F_0 to the peak is the first portion of the absolute amplitude parameter. The other portion is the absolute amplitude from the peak to the end of the event. Either of these portions may be zero, if the event is a simple rise or simple fall. The two ampli-



1. Tilt parameters

tude values are added together to form the absolute amplitude value. The tilt parameter is the difference of the amplitudes divided by their sum [10].

$$tilt = \frac{|A_{rise}| - |A_{fall}|}{|A_{rise}| + |A_{fall}|}$$

The tilt parameter has a range of -1 to 1, where -1 is pure fall, 1 is pure rise, and 0 contains equal portions of rise and fall.

3. SPEECH DATABASE

For training and testing data, we used the Boston University FM Radio corpus, speaker f2b [4]. This database consists of 126 utterances of single speaker female American English news-reader speech (about 45 minutes). The utterances were divided into training and test sets, with the test set comprising one quarter of the utterances. The database is labelled with segment, syllable and word boundaries including lexical stress markings. The database is also already hand-labelled with ToBI intonation labels. For our work the database was additionally labelled with Tilt intonation labels. As our automatic Tilt event labeller is still under development, the Tilt events were derived from the existing ToBI labels. In addition, the database was also fully hand labelled with Tilt events.

4. EXPERIMENTS

For each syllable in the database bearing a Tilt event label, a set of 40 features was extracted. The features include the number of syllables, stressed syllables, and accented syllables proceeding and succeeding the syllable within the phrase; distance, in syllables, from the previous and to the next event; the number of non-major phrase breaks since the last major break; onset and rhyme length [11] [8]; percent of the syllable which is unvoiced; and position of the syllable within a word (e.g. initial, final, medial). The features also include, with a two-syllable window on either side, accentedness, lexical stress, onset and coda types (cf. [11]), Tilt event type and syllable break values. Specifically these are the features which are available at F_0 generation time during synthesis from raw text.

Once the features have been extracted, training sets are created on the basis of event type (accent, boundary, connection, silence) and individual models were built for each Tilt parameter (starting F_0, amplitude, duration, tilt, peak position).

A CART training algorithm [3] is used to develop a decision tree for each parameter, using an optimised subset of the features extracted. The twelve decision trees are used to generate an intonation description file composed of Tilt events and their parameters. The description files are processed to generate the final F_0 contours.

4.1. Measure of accuracy

Measuring the accuracy of a generated F_0 contour is not easy; small changes in the contour are perceptually important at some stages while similar changes elsewhere may be irrelevant. However in order to have some measure of accuracy we follow others ([6], [2]) and use the root mean squared error (RMSE) between the generated contour and the original (smoothed) contour. We also use the correlation between the generated contour and original. The RMSE magnitude is dependent of the F_0 range of the speaker (larger for females than for males) as well as the actual error, while the correlation is more independent. Note that for these examples the segment durations are the same in the generated examples as in the originals and hence voiced sections and unvoiced sections of the signal will always align. RMSE and correlation are only calculated during the voiced sections.

In addition to the overall comparison we also recorded the accuracy of each of the individual models we built on held out test data which helped us concentrate on particular areas for improvement (notably peak position).

Three experiments were carried out, varying the methods used to label the Tilt events. In all cases the continuous Tilt parameters are automatically derived from the Tilt events. In the first experiment the Tilt events were derived from the ToBI labels already in the database by a mostly trivial mapping. In the second experiment we used those same event labels but also include ToBI labels in the features we used to predict the parameters. In the final experiment we used hand labelled Tilt events, and like the first experiment, used no ToBI features in building the models.

4.2. Prediction using automatic Tilt labels

The first experiment consisted of generating parameter prediction models from the automatically labelled Tilt events. An optimised prediction model was created for each parameter (start F_0, amplitude, duration, tilt, peak position) of each event type. That is, not all features were used in each model and hand experimentation was used to find an optimal set of features. The CART method can deal with a certain amount of noise in the input features but will be misled by too much noise (even with cross validation). Tables 1 - 3 show the RMSE and correlation of each of the twelve optimised models.

	stf0	amp	dur	tilt	pk
RMSE	32.45	48.20	0.054	0.422	0.053
Corr	0.546	0.447	0.561	0.558	0.472

Table 1. RMSE and Correlation of accent models

The generated F_0 is generally similar to the smoothed original. As an overall measure of accu-

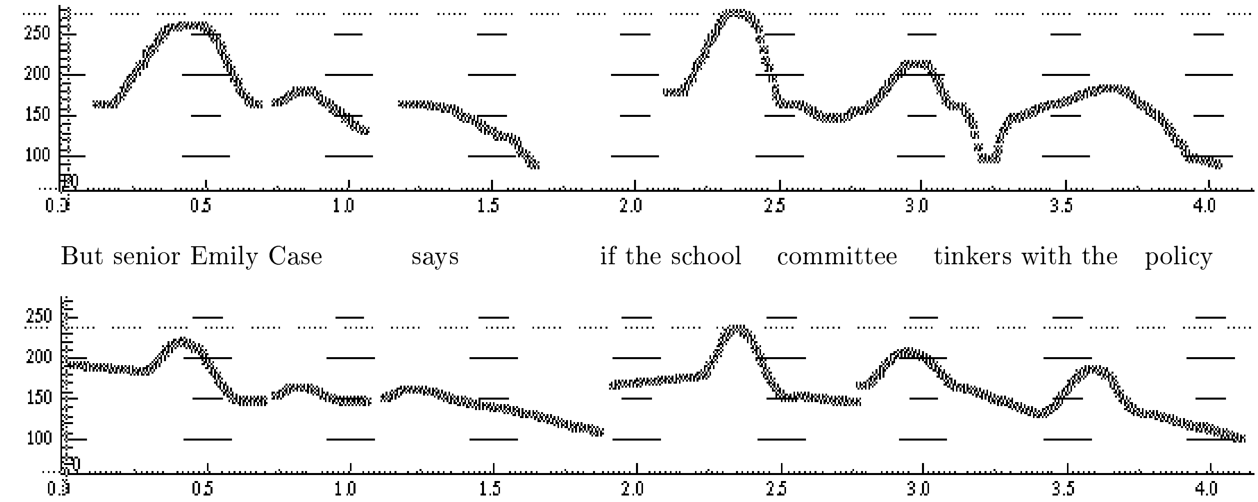


Figure 2: Original smoothed F_0 contour (above) and contour generated from predicted Tilt Parameters (below)

	stf0	amp	dur	tilt	pk
RMSE	28.74	40.92	0.066	0.517	0.079
Corr	0.530	0.408	0.778	0.768	0.695

Table 2. RMSE and Correlation of boundary models

racy for the 28 test utterance, we get an average RMSE of 32.5Hz and correlation of 0.60.

	c	sil
RMSE	31.54	28.17
Corr	0.441	0.810

Table 3. RMSE and Correlation of c and sil models

4.3. Prediction using ToBI labels

In the second experiment the same Tilt events derived from the ToBI labelling are used but this time we include the ToBI labels as features used to predict the individual parameter values. This is done to test what contribution the finer accent and boundary tone distinctions would make to our model.

The individual model results are similar to those of the first experiment. Six of the parameters score higher than the first set, while six score lower.

As with the results of the first experiment, the generated F_0 is generally similar to the smoothed original. For this experiment, we get slightly worse overall results with an average RMSE of 34.0Hz and lower correlation of 0.55 over the 28 test utterances.

4.4. Prediction using hand Tilt labels

Finally we hand labelled the Tilt events in the database. Note labelling Tilt events is much easier than labelling ToBI parameters as the type of accent and type of boundary tone does not need to be assigned, only its presence or absence.

As with experiment two, only half of the individual models show improvement over experiment one. In this experiment, we get RMSE of 33.9Hz and correlation of 0.57 over the 28 test utterances. These results fall in between those of the first and second experiments.

4.5. Comparison of results

Thus for our three experiments our results are outlined in Table 4

labels:	automatic	with ToBI	hand
RMSE	32.5	34.0	33.9
Corr	0.60	0.55	0.57

Table 4. Comparison of overall results

These results compare favourably with a statistically trained F_0 generation model for ToBI labelled data [2] (RMSE 34.8Hz and correlation 0.62) and [6] (RMSE 33Hz). Note these experiments were all carried out on the same database though they may have had different training and test sets.

Figure 2 shows an original smooth contour (above) from our test set and the generated contour using our prediction method (below) from the models created in experiment 1. Two points deserve comment. The difference at the phrase break in the middle of the example is due to our predicted contour being interpolated through unvoiced regions next to silence, unlike the original smoothed F_0. Hence the breaks appear greater in the above original. The second point concerns the accent around the words “the policy”. The original accent actually goes over the two words while the predicted one has a more restricted accent on the first syllable of “policy”, it is possible to hear the difference but it is not significant.

5. DISCUSSION

The goal of this study was to determine whether high quality F_0 contours can be generated using only information available from a text-to-speech system at synthesis time. The results of our contour generation experiments have shown that this is possible. The contours pictured in Figure 2 show the similarity between the generated contours and the originals. Informal listening tests confirm that they produce acceptable contours, though sometimes different from the originals (though usually unimportantly so).

The features used in the generation are all readily available during text-to-speech synthesis. Lexical stress, syllable, word, phonetic and phrasing information are routinely generated as part of the synthesis process before F_0 generation is required.

One aspect that we do not cover in this paper is the automatic assignment of accent and boundary event labels during the synthesis process, likewise [2] and [6] also assume a labelling from which they then generate the F_0. Although we have not yet done tests, predicting Tilt accents and boundaries seems a much easier task than predicting various ToBI pitch accent labels and end tones.

The work in [2] is very similar to this study and was the starting point for our experiments, in terms of the feature sets and the speech database. The phrasal features (e.g. phrase breaks, syllable distances) and stress and accentedness features from that study were all incorporated into our experiment. However, unlike [2], which predicts F_0 for every syllable in the database, we only predict parameters for events. Thus, if a number of syllables fall within a single event, they do not have individual values predicted for them. Our approach is more focused on generating intonation from an accent structure within an utterance than on deriving F_0 from the utterance itself.

The work in [6] is also similar to our approach. By taking similar phrasal and segmental characteristics into account, and working with an accent inventory of minimal size (4 accents, 3 intermediate boundaries, and 2 final boundaries), they generate good quality intonation contours. They also incorporate energy prediction into their experiments, with successful results.

One advantage our study has over [2] and [6] is that the intonation event inventory (accent, boundary, silence, connection) is very simple. Both of the previous studies were forced to collapse the large ToBI inventory into a smaller number of classes in order to achieve their results. The use of the Tilt labelling system eliminates this requirement.

5.1. Context features

The majority of the context features used in this study have been tested previously. [2] However we found that the features used in other work were inadequate for the prediction of peak alignment, so we adopted the use of rhyme length [8] (though non proportional, as supported in [11]), onset and coda classes [11], and onset length [5] (compare accent peak position correlation of 0.42 without use of these features to 0.51 when they are included).

Of these new inclusions, rhyme length was a central feature in all of the peak position predictions (accents, boundaries, with and without ToBI). Coda classification was useful in non-ToBI boundary peak alignment, as well as non-peak predictions (accent amplitude and boundary durations, without ToBI). Onset type was used in both accent peak alignments, but not in the boundary peak alignments. Onset length was among the optimised feature sets for only boundary peak position (with ToBI), but appeared in both ToBI and non-ToBI sets for accent duration and accent starting F_0. Thus, while not all of the features meant to aid peak alignment appeared in prediction models as expected, they did improve the prediction of the peak position, as well as contributing to other prediction models.

6. CONCLUSION

We feel we have adequately shown that an F_0 generation algorithm based on the Tilt theory of intonation can produce acceptable contours from models trained from databases of natural speech. Our results suggest at least equal or better accuracy on held out test data than other similar experiments on the same database. Our best scores are RMSE 32.5Hz and correlation of 0.60 while a ToBI based approach [2] gives 34.5Hz and 0.62 and the dynamical system model [6] produces an RMSE of 33Hz.

REFERENCES

[1] M. Anderson, J. Pierrehumbert, and M. Liberman. Synthesis by rule of English intonation patterns. In *Proceedings of ICASSP84*, pages 2.8.1–2.8.4, 1984.

[2] A. Black and A. Hunt. Generating F0 contours from ToBI labels using linear regression. In *ICSLP96*, volume 3, pages 1385–1388, Philadelphia, Penn., 1996.

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

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

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

[6] K. Ross and M. Ostendorf. A dynamical system model for generating F0 for synthesis. In *Proc. ESCA Workshop On Speech Synthesis*, pages 131–134, Mohonk, NY, 1994.

[7] 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.

[8] K. Silverman and J. Pierrehumbert. The timing of pre-nuclear high accents in English. In J. Kingston and M. Beckman, editors, *Between the Grammar and the Physics of Speech*, number 1 in Papers in Laboratory Phonology, pages 72–106. Cambridge University Press, 1990.

[9] P. Taylor. The Rise/Fall/Connection model of intonation. *Speech Communication*, 15(1/2):169–186, 1994.

[10] P. Taylor and A.W. Black. Synthesizing conversational intonation from a linguistically rich input. In *Proc. ESCA Workshop on Speech Synthesis*, pages 175–178, Mohonk, NY, 1994.

[11] J.P.H. van Santen and J. Hirschberg. Segmental effects on timing and height of pitch contours. In *ICSLP94*, volume 2, pages 719–722, Yokohama, 1994.