GENERATING F_0 CONTOURS FOR SPEECH SYNTHESIS FROM PROSODIC AND SYLLABIC CONTENT

Kurt Dusterhoff

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

ABSTRACT

This paper describes a method for generating F_0 contours from utterances labelled using the Tilt intonation theory. [8] [9] The method uses classification and regression trees (CART) to predict the five Tilt parameters: starting F_0, amplitude, duration, tilt, and peak position. The goal of the experiment is to predict the parameters such that natural intonation contours may be generated from them. Contours generated by this method from a test subset of an American English database have a correlation of 0.60 and a 32.5Hz RMS error when compared with smoothed versions of the original F_0 contour. These results are comparable to other F_0 generation methods which use ToBI intonation labels (0.62 and 34.8Hz, 33Hz).

1. INTRODUCTION

One of the tasks involved in speech synthesis is the generation of prosody. As synthetic speech quality advances, adjustments in pitch and duration represent steps toward more natural speech. The experiment discussed here presents one method of generating natural looking and sounding pitch contours for use in speech synthesis.

The experiment uses the Tilt intonation labelling theory to produce natural F_0 contours from prosodic and syllabic context. The labelling system provides four intonational events: accent, boundary, connection, and silence. The context features include accentedness, lexical stress, and segmental content information. It is reasonable to expect such information to be available at F_0 generation time.

First, a basic overview of the speech data and Tilt theory is presented. This is followed by a description of the experiment and the context features used. The results of the experiment are then discussed and related to other work in the field.

2. TILT THEORY

The Tilt intonation theory gives a data-driven abstraction of fundamental frequency contours, while also maintaining a method for reversing the abstraction to recreate the contours. Each intonation event (accent, boundary, silence, connection) is described by one or more continuous parameters, which are extracted from labelled contours. These parameters may then be used to regenerate an F_0 contour. The goal of the experiment presented here is to successfully predict the Tilt parameters such that natural intonation contours are generated from the predicted parameters.

The four intonation events are each described by at least one parameter: *starting F_0*. This is the fundamental frequency at the beginning of the event. Accents and boundaries are described by additional parameters. The *amplitude* parameter is the sum of the distance (in Hertz) between the starting F_0 and the F_0 at the peak of the accent (the *rise amplitude*) and the distance (in Hertz) between the peak and the F_0 at the end of the event (the *fall amplitude*). *Duration* is the length (in seconds) of the intonation event. The fourth parameter is *peak position*, the distance (in seconds) from the start of the first vowel of the event and the peak of the F_0 for that event. The final parameter is *tilt*. Tilt is the result of dividing the difference of the rise and fall amplitudes by the sum of the rise and fall amplitudes. [8].

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

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.

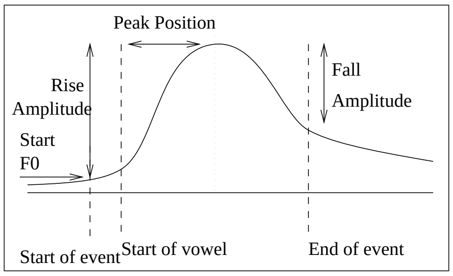

Figure 1: Tilt parameters

3. SPEECH DATABASE

The speech database from which all context features were extracted and on which all training and testing took place is from the Boston University radio news corpus, speaker f2b [4]. The corpus contains approximately 45 minutes of female American English news-reader speech. The data was divided such that one quarter of the utterances were held out as a test set. The database is labelled with segment, syllable, and word boundaries, including lexical stress location. In addition, the corpus is labelled with ToBI intonation labels. The database was labelled with Tilt intonation labels by automatically mapping the ToBI tone labels to Tilt event labels [3]. The Tilt parameters from these labels were used to train the decision trees for the experiment.

4. EXPERIMENT

For each syllable in the database which carries a Tilt label (a, b, c, sil), a set of 40 features was extracted for testing. The features include the number of syllables, stressed syllables, and accented syllables preceding 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 ([10],[7]); 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 classification (cf. [10]), [5]); Tilt event type (two event window, as well as two syllable), and phrase break values.

All of these features are available to the system at F_0 generation time during synthesis. Accent, a simple binary feature, is assumed to have been predicted prior to F_0 generation.

Once the features are extracted, a CART training algorithm [2] is used to create a decision tree for each parameter of each event type (twelve trees in total). These trees are used to predict the parameters of the Tilt events on a held out test set. The parameters are then used in the generation of intonation contours.

4.1. Accuracy measurement

The measurement of the accuracy of the F_0 generation process followed previous studies in using the root mean squared error (RMSE) to measure how far away the predicted F_0 is from its target (the original F_0 of an utterance) at any given time. Correlation is a more general measure which shows how well a contour changes in time with its target (i.e. do they both rise and fall in the same places). The RMSE is tied very closely to the speaker being tested (generally smaller for males than for females, due to overall pitch range), while correlation provides a more independent measure.

In addition to the overall measure, RMSE and correlation were used to judge the accuracy of the individual models used in the development of the decision trees. These individual results aided in the selection of optimized feature sets (i.e. not all features were used for predicting all parameter values) for each parameter, as well as deterimining which parameters were in need of improvement.

4.2. Results

The first step in predicting the overall contour is to predict each parameter for each event type. Thus, the five Tilt parameters for accents are individually modelled. Table 1 shows the final results for each of the accent parameters. Table 2 shows the boundary results. Note the consistently higher correlation scores in the boundary results. These results reflect the nature of boundaries to fall into categories which are more distinctive than those of accents. Table 3 shows the start F_0 parameter for silences and connections. The lack of systematicity of connections is reflected in the comparably low correlation as compared to that of silences.

	stF_0	amp	dur	tilt	pk
RMSE	32.45	48.20	0.054	0.422	0.053
Corr	.5459	.4473	.5606	.5576	.4718

Table 1: RMSE and Correlation of "A" parameter models

	stF_0	amp	dur	tilt	pk
RMSE	28.74	40.92	0.066	0.517	0.079
Corr	.5301	.4084	.7779	.7679	.6947

Table 2: RMSE and Correlation of "B" parameter models

	conn	sil
RMSE	31.54	28.17
Corr	.4414	.8095

Table 3: RMSE and Correlation of start F_0(Hz) parameters for connections and silences

The generated F_0 from the optimized models is generally similar to the smoothed original. We get an RMSE of 32.5Hz and correlation of 0.60 averaged over the 28 test utterances. These results are comparable with [1], which uses similar features to predict F_0 from ToBI labels (34.8Hz and 0.62). They are also comparable to [6], which uses a dynamical state system incorporating energy and syllabic context (33Hz).

Figure 2 shows an original smoothed contour (above) and a contour generated from predicted Tilt parameters (below). Note that the large phrase break in the middle of the original contour is not duplicated in the generated contour. This is due to the interpolation through unvoiced segments adjacent to silence in the generated contour. In the original, unvoiced segments next to silence are treated as a part of that silence. The accent and boundary on the phrase "the policy" are also worthy of note. In the original contour, the accent spans the phrase, with a sharp drop of F_0 at the boundary. In the generated contour, the accent is restricted to the immediate vicinity of the syllable where the peak is located. While it is possible to hear the difference, the general nature of the utterance is unaffected.

5. CONCLUSION

The goal of this experiment was to generate natural F_0 contours from prosodic and syllabic context using the Tilt intonation theory. The results have shown that this is possible. The contours pictured in Figure 2 show great similarity between the original and synthetic contours. While informal listening tests show noticeable differences, these differences do

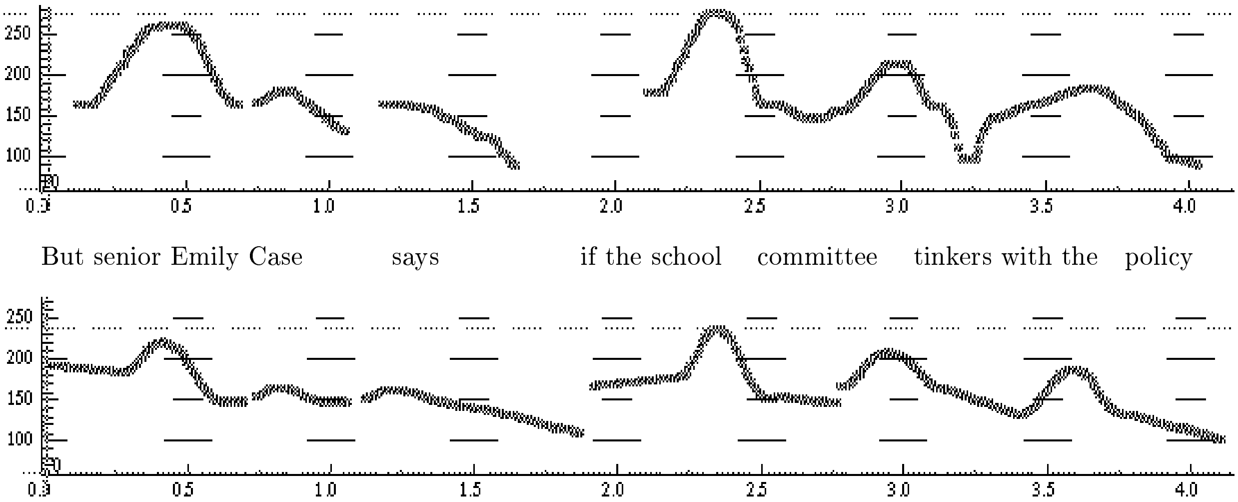

Figure 2: Original smoothed F_0 contour (above) and contour generated from predicted Tilt Parameters (below)

not appear to reflect any important distortion of the utterance.

The features used in predicting the Tilt parameters for this study are all available prior to F_0 synthesis. Most are routinely generated as a part of synthesis in modules other than intonation. Accent assignment is seen as a separate step to contour generation, and is outside the scope of this study.

The feature modelling approach taken here is very similar to that in [1]. The features used in this experiment include those in [1] and they are used on the same dataset. While the results are an improvement, our approach has yet to be tested using ToBI labels. Future experimentation will compare the results of the method described here as used on Tilt as well as ToBI labelled data.

The large number of features tested in this experiment exceeds that of previous experimentation. The twenty-four features from [1] are tested here, in addition to seventeen new features. In [6], a variety of segmental, syllabic, and phrasal features are combined with energy to produce similar results. However, the wide range of features used in the experiment may also serve to test specific hypotheses found in the literature. Several hypotheses dealing with peak alignment (cf. [7] [10] [5]) contributed to the list of tested features. Further testing of individual features under a variety of conditions is possible, and may constitute further experimentation.

This work represents an improvement upon previous work. The results suggest accuracy matching or bettering other studies on the same corpus. The average score over 28 test utterances of RMSE of 32.5Hz and correlation of 0.60 compare favorably with a ToBI-based approach [1] (34.5Hz and 0.62) and the dynamical system model [6] (33Hz).

REFERENCES

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

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

[3] 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.

[4] 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.

[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 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.

[8] 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.

[9] Paul Taylor. The Rise/Fall/Connection model of intonation. *Speech Communication*, 15, 1994.

[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.