

F0 GENERATION FOR SPEECH SYNTHESIS USING A MULTI-TIER APPROACH

Xuejing Sun

Department of Communication Sciences and Disorders, Northwestern University
2299 N. Campus Dr., Evanston, IL 60208, USA
sunxj@northwestern.edu

ABSTRACT

In this paper we propose a multi-tier phonetic model for intonation synthesis. Four tiers are defined aiming at different components of intonation. The parametric representation for each tier is realized through underlying pitch target. Regression trees are built for each tier upon appropriate feature set. Both numerical evaluation and informal listening test have yielded encouraging results. The synthesized sentences can be accessed at <http://mel.speech.northwestern.edu/sunxj/prosody.htm>

1. INTRODUCTION

Intonation modeling has both theoretical interests and practical applications. Many systems have been proposed in the past, ranging from high level phonological representations (e.g. ToBI [10]) to low-level phonetic models (e.g. Tilt model [13]). Partly due to the availability of large corpus, more phonetic models have emerged lately particularly for speech synthesis since they are better handled by statistical methods.

Among the existing phonetic models, we distinguish two lines of approaches – parametric vs. non-parametric. Parametric models (e.g. [5][7]) usually transform the original F0 values into some parametric forms, and after the parameters are predicted, the intonation can be re-synthesized from these parameters. In non-parametric models (e.g. [2][3][9][15]), several target F0 points from original curves are predicted directly.

Although seemingly simple, non-parametric models can usually achieve equivalent results compared with parametric models. This is because F0 values themselves are meaningful linguistic properties and thus predictable from linguistic features. However, even though input features at different levels are used, the models themselves do not have a clear hierarchy. This could be a problem since it is known that various components in intonation are caused by factors at different levels and have different perceptual weight. In addition, in non-parametric models the smoothness of predicted F0 curve is not treated inherently and must be realized through some post-processing techniques. On the other hand, parametric models usually are internally consistent and can selectively model certain intonation component based on theoretical assumptions or input features. However, transforming F0 values into some parametric representations has the risk of losing linguistic meaning and in turn predictive power, which has been observed in our own work and by others [6]. Certain parametric forms, such as polynomial functions, Fourier series, etc., may approximate intonation very

accurately, but mapping between their parametric space and the input linguistic features may be too complex for a statistical algorithm to learn.

In this paper we propose a multi-tier phonetic model built upon the work described in [12], but with significant extension. This model employs a hierarchical structure in attempting to model different components of intonation in different tiers. With this structure, we may build models for each tier using the best-suited feature set and method guided by existing linguistic knowledge. The model uses a simple parametric form to represent F0, where syllable is regarded as the basic intonational unit. The parametric representation ensures the smoothness of F0 curve at syllable level. It is believed that the parameters are linguistic meaningful and could be predicted well.

2. THE MODEL

2.1. Basic structure

The model consists of four tiers. Each tier covers different part of intonation. In the following we first describe the functionality of each tier and then illustrate their parametric representations.

Tier One

In this tier, we want to model intonation variations above the word level. This is similar to predicting pitch accent for each word, except that continuous F0 values need to be predicted in our work rather than discrete categorical labels. Specifically, we choose one F0 target point (referred as *anchor F0* hereafter) from each word (for multi-syllable words, the point is chosen from the syllable bearing the primary stress).

Tier Two

Here we consider intonation at word level, i.e., internal F0 variations in a multi-syllable word with respect to the anchor F0. We select one F0 point from each syllable that is not bearing primary stress to represent F0 variations within a word.

Tier Three

In last two tiers we have modeled intonation variations within or above word level by selecting a target F0 point from each syllable. The phonetic details at syllable level, however, are still missing. In this tier, we model F0 variations within a syllable by using simple pitch movements, e.g., whether the contour is a rise, fall, etc.

Tier Four

In this tier, we focus on some segmental effects, such as tonal coarticulation and consonant perturbation, which are not modeled by Tier Three. Although exception exists, these

effects are usually local, with scope not beyond the syllable level.

2.2. Parametric representation

To model the tiers described above, we need to define the corresponding parametric representation. Similar to [12], we use the underlying pitch target [16][17] as the basic parametric form, and represent each tier with the parameters of pitch target. The pitch target analysis procedure is briefly described below, which is basically the same as that in [12].

First, for each syllable we define

$$T(t) = at + b \quad (1)$$

$$y(t) = \beta \exp(-\lambda t) + at + b \quad (2)$$

where $T(\cdot)$ represents the underlying pitch target, and $y(\cdot)$ represents the surface F0 contour. Coefficient β is a scaling parameter, and its value is the distance between F0 contour and the underlying pitch target when $t = 0$. Parameter λ is a positive number representing the rate of decay of the exponential part. Parameters a and b are the slope and intercept of the underlying pitch target.

Next, let (t_0, y_0) denote the first point on the F0 contour, and let (t_1, y_1) denote a point where underlying pitch target has been approached, we have:

$$y(t) = (y_0 - y_1 + at_1) \exp(-\lambda t) + at + y_1 - at_1 \quad (3)$$

The parameters of the model are estimated by nonlinear regression. When nonlinear regression fails, linear regression is performed. In practice, for the starting F0 value, we use an average of the first two F0 values in estimation because the first point sometimes can be aberrant. For (t_1, y_1) , we use the point in the middle of a segment, which seems to work best. Different from [12], in this study, we performed pitch target analysis on the vocalic part of a syllable rather than the vowel domain. This is closer to our notion that syllable is the basic intonational unit.

Tier One

For single-syllable words, we select the middle F0 point of its pitch target; for multi-syllable words, the middle F0 of pitch target of the syllable bearing primary stress is selected. It should be noted that the database used in this study only provides information on stress vs. non-stress. We arbitrarily select the first stress syllable as the primary stressed syllable, which is a suboptimal solution. In future work a dictionary lookup procedure may be used.

Tier Two

This tier covers all the syllables without primary stress in multi-syllable words. Since we want to model word internal F0 variation here, two phonetic realities are considered, the absolute F0 value of the syllable and the F0 difference with respect to the anchor F0 determined in Tier One. More specifically, for each syllable, we use middle F0 value of the pitch target and its difference to that of the syllable bearing primary stress.

Tier Three

For Tier Three, the slope of the underlying pitch target (a) is used to model the pitch movement at syllable level.

Tier Four

To model segmental effects, we can use either parameters β and λ in Eq. (2) or by some post-processing schemes. Since β and λ exhibit complex behaviors and are difficult to model, we opt to use interpolation or some simple rules.

3. EXPERIMENTS

3.1. The corpus

Training and testing data were taken from Boston University Radio Speech Corpus, speaker F2B. The database, which consists of 40 minutes speech read by a female professional announcer, is labeled using the ToBI [10] system. The database also contains text information, such as part-of-speech, and acoustic information such as phone duration. F0 curves were determined by the SHRP algorithm [11], and smoothed by de-step filter [1], five-point median and linear filters. No manual correction was performed on F0 values. The data set was split into training and testing sets with approximately a 9:1 ratio.

3.2. Building regression trees

“WAGON” [14], an implementation of standard CART (Classification and Regression Tree), was used to build regression trees. Although sometimes it may not be as accurate as a fine-tuned neural network especially for complex tasks, decision tree algorithm has several advantages: (1) faster training and testing; (2) less hand-tuning of the parameters; (3) results are human-readable and can strengthen our theoretical understanding; (4) the trained model can be integrated into existing speech synthesis systems very easily.

Note that in forming the feature set, we tried to use features that are readily available in a TTS system. Therefore, we deliberately excluded some features, such as duration, even though they are useful for better intonation models. In the features discussed below, phrase break is probably the most difficult feature to obtain. To make things simpler, we only used major phrase break also known as intonational phrase break. In general, our feature sets are quite simple, and most of them are commonly used by other studies.

3.2.1. Tier One

The input features can be loosely classified into three groups:

- Syntactic, semantic, pragmatic features or other discourse information that determine pitch accent distribution: Currently, we only have part-of-speech in this group due to its availability.
- Positional features: the word position in a sentence and intonational phrase.
- Other features: vowel contained in the syllable; syllable stress; syllable position in a word. These features may not be as important as those in the first two groups with respect to the overall intonation pattern, but they could be helpful in predicting minute F0 variations.

Three regression trees were built based on these features or features derived from them. The input feature set for the first tree includes mainly the positional features. In the second tree, in addition to positional features we used part-of-speech of current, previous, and next word. In the third tree, we expanded the length of context window by incorporating part-of-speech of previous and next two words. To include which features from the third groups for each tree was mostly determined experimentally. The final F0 values were obtained by averaging the results predicted by three regression trees. Note that by building trees with different feature sets, we attempt to capture specific intonation patterns in each tree.

3.2.2. Tier Two

In this tier, we collected the following features for each syllable:

- The lexical stress (1 or 0) of the current, previous, and next syllable
- Syllable position in a word (initial, medial, final, and single) of the current, previous, and next syllable
- Number of syllables in the current word and previous word
- The position of the word in a sentence and an intonational phrase.

Compared with Tier One, a significant change in feature set is that more local context information is included while part-of-speech information is removed. Two regression trees were grown upon these input features. In the first tree, the F0 difference between the current syllable and the syllable bearing primary stress in the same word was predicted. The second tree was trained to predict the absolute F0 value of the current syllable. The predicted values from the first tree were converted to F0 values by adding them to the corresponding anchor F0 values obtained from Tier One. Then the final F0 values were estimated by taking the average of reconstructed F0 from the first tree and the F0 predicted by the second tree.

3.2.3. Tier Three

In this tier we want to predict the slope value of the pitch target for each syllable. The input features were derived from the following information:

- Vowel contained in this syllable
- The final voiced consonant of this syllable
- The lexical stress (1 or 0) of the current, previous, and next syllable
- Syllable position in a word (initial, medial, final, and single) of the current, previous, and next syllable
- Number of syllables in the current and previous word
- Part-of-speech of the current, previous, and next word
- The position of the syllable in an intonational phrase
- The position of the word in a sentence and an intonational phrase.

We built two regression trees upon the input features and averaged their outputs. The main difference between the two trees is whether to include part-of-speech in the feature set. This reflects our assumption that pitch movements at syllable level be affected by factors at different levels.

3.2.4. Tier Four

Directly connecting underlying pitch target as that in [12] ignores too much segmental information. In the current work, we employed a more complex interpolation scheme: For a given pitch target, a cubic-spline interpolation starting from the end of the previous target to the middle point of the current target is performed. It should be noted that this interpolation scheme is still too simple and even incorrect in some cases. Possible extensions include defining some rules based on existing linguistic knowledge. For example, we may raise the starting F0 value of a syllable to certain extent when the syllable starts with certain consonants.

3.3. Results

For objective evaluation of synthesized intonation, we adopt the commonly used root mean square error (RMSE) and

correlation coefficient (r). First, we want to see how good our parametric representation of intonation is. We compare the original F0 values and the model F0 reconstructed from the “true” parameter values using Eq. (2). Note that only F0 values at voiced portion are considered. Table 1 shows the comparison results as well as that obtained by Dusterhoff et al.[5] using Tilt model.

	RMSE	r
Original F0 – Model F0 (present)	6.5	0.99
Dusterhoff et al. [5]	14.5	0.93

Table 1: Comparison between original F0 and model F0

It can be seen that the current approach yields closer F0 values numerically, which confirms the eligibility of the parametric form.

Table 2 presents comparison results between predicted F0, original F0, and model F0. Since all the previous reported results on the same database we know are obtained with human labeled accent labels, it is not easy to make a direct comparison. Nevertheless, in order to have a general idea, we built two regression trees for middle F0 and slope of pitch target with ToBI accent labels. Note that some studies’ results [3][7][15] are not included since they are obtained from different databases.

	RMSE	r
Original F0 – Predicted F0 (No ToBI)	36.2	0.66
Model F0 – Predicted F0 (No ToBI)	35.3	0.67
Original F0 – Predicted F0 (ToBI)	33.1	0.72
Model F0 – Predicted F0 (ToBI)	32.2	0.73
Original F0 – Predicted F0 ([2])	34.8	0.62
Model F0 – Predicted F0 ([5])	34.3	0.60
Original F0 – Predicted F0 ([9])	34.7	N/A

Table 2: Comparison between original F0, model F0, and predicted F0 from current approach and other models

Table 2 shows that in terms of RMSE, our approach yields better results with ToBI labels and worse results without ToBI labels. In terms of correlation coefficient, our results are consistently better. The numerical results are encouraging, which indicate that our system can achieve comparable performance even without accent labels. It should be noted that different ways of computing RMSE and correlation coefficient might affect the results. In this work, we calculated a RMSE and a correlation coefficient value for each file (in the current database, each file contains one paragraph, which usually consists of several sentences), and the final results are the average for all the files in the testing set. Other authors might have used different schemes.

Numerical results only give us a partial picture. To evaluate the results subjectively, we re-synthesized sentences using predicted F0 with PSOLA [8] technique implemented in Praat (<http://www.praat.org>) software. The original duration was used in the synthesis. Informal listening tests show that synthesized sentences without accent labels are satisfactory in most cases. Some uncommon intonation patterns were not captured, which we speculate a larger database would help. The intonation generated by the system trained with accent labels sounds fairly good.

4. DISCUSSION

The encouraging results from both numerical and informal perceptual evaluation have shown that the proposed approach is a viable direction for modeling intonation. To gain more insight, in the next we compare our model with others from several aspects.

First, by describing intonation in a multi-tier approach, we build the model step by step, in a roughly hierarchical way. In so doing, we can put more effort on the more important components (e.g. Tier One in our model). This structure allows us to use input features more efficiently and choose the best model for each tier, which could be viewed as an application of divide and conquer strategy. With this structure comparable results are obtained without using accent labels, which is attractive for developing TTS system since labeling or predicting accents is a nontrivial task.

Second, by using a parametric form, we can ensure the consistency and smoothness of the contour at the domain covered by the model, which is syllable in the current work. In non-parametric approaches (e.g. [2][3][9][15]), F0 contours are generated by connecting several target F0 values predicted either separately or together. Although problems can be alleviated by employing interpolation or post-smoothing techniques, the model itself has no inherent ability to prevent abnormal F0 trajectory.

Third, compared with other parametric models (e.g., [5][7][13]), the present parametric representation is very simple. The number of parameters needs to be predicted in this study is only two, which reflects our efforts in avoiding a complex model that may be less predictable and less efficient.

Finally, in our system we build two or three regression trees for each tier by splitting the input feature set (Tier One and Tier Three) or predicting different output values (Tier Two). This has both linguistic motivations and mathematical background. From a linguistic perspective, for example, in Tier One we want to use different subset of features to model certain intonation patterns. In Tier Two, predicting different F0 forms reflects our assumption that the F0 values of unstressed syllables depend both on the input features and the F0 of the syllable with primary stress within the same word. From a machine learning perspective, our methods are special forms of ensemble learning technique [4], which has proved to be superior to single learning machine. Experiments have also been conducted with more popular ensemble learning algorithms, such as bagging and boosting [4]. Considerably better numerical results were obtained. However, informal listening did not reveal audible perceptual difference. We speculate that ensemble learning should yield more robust intonation in a real TTS system, which can only be proved with more extensive and systematic perceptual tests.

5. CONCLUSIONS

The preliminary results of the present study have shown that modeling intonation via a multi-tier approach is promising. With a simple feature set, bypassing the accent layer, our method achieved comparable results to those reported in previous studies, in which accent labels are often used. This encourages us to explore further along this directions. Future studies should include experiments with better feature set for each tier and more sophisticated learning algorithms.

6. ACKNOWLEDGEMENT

The author wishes to thank Yi Xu for helpful comments on the manuscript. This study was supported in part by NIH grant DC03902.

7. REFERENCES

- [1] Bagshaw, P. C. "Automatic prosody analysis," *Ph.D. Thesis*, University of Edinburgh, Scotland, UK, 1994.
- [2] Black, A. and Hunt, A. "Generating F0 contours from ToBI labels using linear regression," *Proc. of ICSLP*, Philadelphia, Penn., 1996.
- [3] Buhmann, J., Vereecken, H., Fackrell, J., Martens, J. and Coile, B. "Data driven intonation modelling of 6 languages," *Proc. of ICSLP*, Beijing, China, 3, pp. 179-182, 2000.
- [4] Dietterich T.G. "Machine learning research: Four current directions," *AI Magazine*, 18(4):97-136, 1999.
- [5] Dusterhoff, K. E., Black, A. W., and Taylor, P. A. "Using decision trees within the tilt intonation model to predict F0 contours", *Proc. of Eurospeech*, 1999.
- [6] <http://chardonnay.elis.rug.ac.be/en/research/tts.html#high>
- [7] Mohler, G. and Conkie, A. "Parametric modeling of intonation using vector quantization," *Proc. of Third International Workshop on Speech Synthesis*, 1998.
- [8] Moulines, E., and Charpentier, F. "Pitch-Synchronous Waveform Processing Techniques for Text-To-Speech Synthesis Using Diphones", *Speech Communication*, 9: 453-467, 1990.
- [9] Ross, K. and Ostendorf, M. "A dynamical system model for generating fundamental frequency for speech synthesis," *IEEE Trans. Speech and Audio Proc.* 7, 295-309, 1999.
- [10] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., Hirschberg, J., "ToBI: A standard for labelling English prosody", *Proc. of ICSLP*, Banff, Alberta, pp. 867-870, 1992.
- [11] Sun, X., "Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio," *Proc. of ICASSP*, Orlando, Florida, Vol. 1, pp 333-336, 2002.
- [12] Sun, X. "Predicting Underlying Pitch Targets for Intonation Modeling," *Proc. of the 4th ISCA Tutorial and Research Workshop on Speech Synthesis*, Perthshire, Scotland, pp. 143-148, 2001.
- [13] Taylor, P.A., "Analysis and synthesis of intonation using the Tilt model", *J. Acoust. Soc. Am.* 107, 1697-1714, 2000.
- [14] Taylor, P., Black, A., and Caley, R. *Introduction to the Edinburgh Speech Tools*, 1999. http://www.cstr.ed.ac.uk/projects/speech_tools/.
- [15] Traber, C. "F0 generation with a database of natural F0 patterns and with a neural network," in *Talking Machines: Theories, Models, and Designs*, G. Bailey, C. Benoit, and T. R. Wawallis, Eds. Amsterdam, The Netherlands: Elsevier, pp 287-304, 1992.
- [16] Xu, C. X., Xu, Y. and Luo, L-S. "A pitch target approximation model for F0 contours in Mandarin", *Proc. of ICPhS*, San Francisco. pp. 2359-2362, 1999.
- [17] Xu, Y. and Wang, E., "Pitch targets and their realization: Evidence from Mandarin Chinese", *Speech Communication* 33 (4), 319-337, 2001.