INTERSPEECH 2009 BRIGHTON

Optimization of T-Tilt F0 Modeling

Ausdang Thangthai, Anocha Rugchatjaroen, Nattanun Thatphithakkul, Ananlada Chotimongkol, Chai Wutiwiwatchai

Human Language Technology Laboratory,
National Electronics and Computer Technology Center (NECTEC), Thailand
{ausdang.tha, anocha.rug, nattanun.tha, ananlada.cho, chai.wut}@nectec.or.th

Abstract

This paper investigates on the improvement of T-Tilt modeling, a modified Tilt model specifically designed for F0 modeling in tonal languages. The model has proved to work well for F0 analysis but suffers from text-to-F0 prediction. To optimize, the T-Tilt event is restricted to span over the whole syllable unit which helps reduce the number of parameters significantly. F0 interpolation and smoothing processes often performed in preprocessing are avoided to prevent modeling errors. F0 shape pre-classification and parameter clustering are introduced for better modeling. Evaluation results using the optimized model show the significant improvement for both F0 analysis and prediction.

Index Terms: T-Tilt, optimization, F0 analysis and prediction

10.21437/Interspeech.2009-187

1. Introduction

Intonation is the speech melody that is important for general and expressive text-to-speech systems to make the synthesized speech sound like a human. It is a crucial part of tonal languages because each syllable uses a tone expressed by a specific F0 contour to distinguish the meaning of words. Another challenge of intonation modeling might need to be remodeled for individual speakers with different speaking styles. It depends on many factors such as genres, dialects and genders.

In recent years, many papers have investigated on parameterization approaches that can be done fully automatic and is easily adaptable to new languages and/or new speakers. The general parameterization approach consists of two processes. First, F0 contours of training samples are fitted to mathematic formulations containing a set of parameters, namely, model fitting. Second, an intonation model is built using machine leaning techniques. The model concerns finding a mapping function from linguistic features extracted from the input text to each parameter [1-3] or all parameters [4, 5] in the formulations.

For tonal languages, there are two challenges in parameterization, namely, F0 shape variations and training data sparseness. Regarding the first challenge, Xu [6] found that the F0 shape of lexical tone in isolated syllables seems well defined and quite stable, while the nature of tonal context are largely varies. Traditional approaches attempting to capture all the variations using too many parameters might lead to model over-fitting. Moreover, F0 contours of speech utterances are often disconnected due to unvoiced parts. F0 transitions are often filled using straight line interpolation with median smoothing [1-3, 5, 7-9]. This process unintentionally produces tonal variations and distorted tonal F0 shapes. Pablo et al. [4] mentioned to avoid filling the F0 transition between disconnected voiced portions.

Regarding the second challenge, comparing to non-tonal languages, the number of tones and the F0 shape variation in each tone increases the size of data required for building an accurate F0 model. Some clustering as well as machine learning techniques were proposed to alleviate this problem e.g. vector-quantization [5], k-mean [9], and decision tree [1,2,4].

Recently, the T-Tilt model [2] has been invented as one of effective models for F0 modeling in tone languages. Besides its advantage in constructing with a fully automatic method, it showed a high performance in F0 contour analysis. However, as those found in other models, its problem was the weak relationship between textual features and model parameters in the F0 synthesis process. In this paper, we introduce three ways to optimize the T-Tilt model. First, instead of having a Tilt event at any position in a syllable, we restricted that the Tilt curve always expands over the whole syllable. Second, we proposed a method to remove fault F0 movements that came from the interpolation and smoothing process normally performed in F0 modeling [2, 5, 9]. Lastly, we reduced the F0 shape variations by classifying the shape into eight groups such as rise, fall, rise-fall, fall-rise, etc. Classification and regression trees (CART) were built independently for each group. Furthermore, the k-mean clustering technique was applied to solve data sparseness in each shape group.

This paper is organized as follows. Section 2 reviews the T-Tilt model. Section 3 presents an overview of the proposed model. Section 4 explains details of experiment data and evaluations. Finally, Section 5 provides the conclusion to this work.

2. T-Tilt Modeling

The T-Tilt model [2] is one of parameterization approaches that was modified from the Tilt intonation model [1] to better work for tonal languages. In tonal languages, the F0 movement has often been modeled on the basis of syllable units. The T-Tilt model consists of eight continuous-value parameters forming the F0 contour of a syllable including

- *start_f0* : the F0 at the starting point of the syllable,
- *start_tilt* : the starting time of the Tilt in the syllable,
- *event_amp* : the summation of absolute rising (A_{rise}) and falling (A_{fall}) amplitudes (negative for the valley F0 shape),
- *event_dur* : the summation of rising(D_{rise}) and falling(D_{fall}) duration,
- *peak_pos* : the duration distance between the starting point of the syllable to the peak of the Tilt,
- *shape_type* : the type of F0 shape,
- *tTilt_amp* and *tTilt_dur* : the difference of rising and falling amplitudes and durations divided by their summation.

Equations 1 and 2 represent the hill shape and Equations 3 and 4 is additional for tonal languages that these represent the valley shape, where $f_0(t)$ is the F0 value at a time t, A is a rising or falling amplitude, D is a rising or falling duration and A_{abs} is an absolute F0 value at the starting point of the rising or falling curve.

$$f_0(t) = A_{abs} + A - 2 \cdot A \cdot \left(\frac{t}{D}\right)^2 \qquad 0 < t < \frac{D}{2} \quad (1)$$

$$f_0(t) = A_{abs} + 2 \cdot A \cdot (1 - \frac{t}{D})^2 \qquad \frac{D}{2} < t < D \quad (2)$$

$$f_0(t) = A_{abs} + A - A \cdot (\frac{t}{D})^2 \qquad 0 < t < D \quad (3)$$

$$f_0(t) = A_{abs} + A \cdot (1 - \frac{t}{D})^2 \qquad 0 < t < D \quad (4)$$

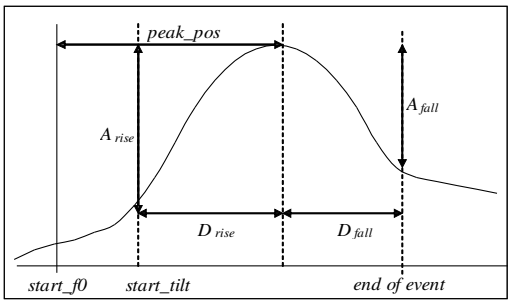

Figure 1: *Parameterization in the T-Tilt model.*

3. Optimization Methods

An overall procedure of the proposed optimized model is illustrated in Figure 2. This section will explain the model in four subsections including T-Tilt model fitting, F0 shape labeling and prediction, T-Tilt parameter clustering and prediction, and F0 contour generation.

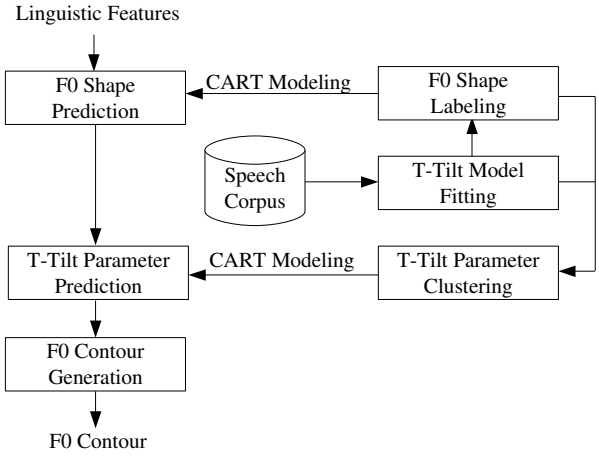

Figure 2: *The proposed optimized model.*

3.1. T-Tilt Model fitting

The model fitting is the first step of F0 parameterization. The F0 contour of each syllables are fitted to mathematic formulations containing a set of parameters. Traditional models required F0 transitions between adjacent tones. An interpolation process is often used to fill in F0 values around unvoiced regions of speech. A smoothing process is also common to remove sudden changes after interpolation, and eliminate micro prosody and measurement errors often found at voiced-unvoiced boundaries. Major drawbacks of these processes are the distortion from original shapes; introducing noise that leads to F0 shape ambiguity. Figure 3 shows examples of two syllabic F0 contours A and B (the upper contour) which should be similarly parameterized. However, after the interpolation and smoothing processes (the lower contour), they can be differently modeled.

There are three constraints in our proposed model fitting method. First, instead of having a Tilt starting at any position in the syllable, we restricted that the Tilt curve always expands over the whole syllable. Second, we removed the interpolation and smoothing process so that the fault F0 movements came from this process could be eliminated. As a result, the number of T-Tilt parameters was reduced from eight to five, including *start_f0, tTilt_amp, event_amp, peak_pos,* and *shape_type.* Finally, the five T-Tilt parameters are estimated by an automatic analysis-by-synthesis algorithm. The algorithm used the grid search method to determine the best fitted T-Tilt curve to starting, peak, and ending points of each syllable. Among all hypotheses, the root mean square error (RMSE) between modeled and exact F0 curves are used to determine the best T-Tilt shape.

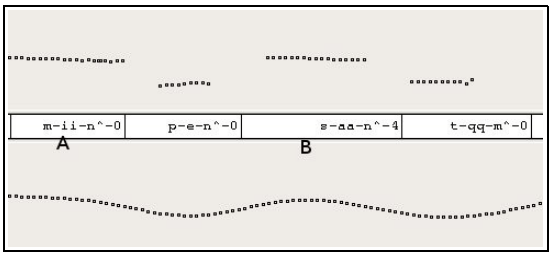

Figure 3: *The drawback of F0 contour interpolation and smoothing.*

3.2. Shape labeling and prediction

As mentioned in the Introduction, many F0 modeling approaches achieved a high performance in analysis but a rather low accuracy in synthesis. Major problems are the weak relationship between input linguistic features and targeted model parameters, and the large variations of F0 curves. An interesting way to alleviate this problem is to cluster F0 curves into several groups and model separately for each group. The accuracy of F0 prediction could not be improved when clustering by tones [10]. This is not surprising since, as described, the variation of F0 shapes within the tone class is extremely high. Instead of clustering by tones, we proposed to clustering by the F0 shape, based on its exact curvature in T-Tilt, and call this parameter *shape_type.* Eight shapes shown in Table 1 were defined.

Table 1. *The 'shape_type' parameter.*

Label	Shape
R	rising hill
R+	rising valley
F	falling hill
F+	falling valley
RF	rising hill followed by falling hill
RF+	rising valley followed by falling valley
FR	falling hill followed by rising hill
FR+	falling valley followed by rising valley

With an analysis on the relationship between the tone and the *shape_type* parameter, followings were observed.

- Each tone can indeed appear in all eight types of shape. An example given in Figure 4 is a distribution of the Thai falling tone in all eight shape types.
- In each tone appeared in polysyllabic words, the percentage of each shape type is not significantly different.
- However, when we divided syllables by their positions in a phrase (syllables at the beginning, at the end, and at other places), we found that the F0 shape of ending syllables were somewhat stable while beginning syllables were largely fluctuated. The Figure 4 also shows an example of the Thai falling tone distributed over the three syllable positions.

In our proposed model, a shape type of each syllable is first predicted using the CART model [11] given a set of linguistic features extracted from the input text, including

- tones of the current and surrounding syllables
- part-of-speeches (POSs) of the current and surrounding words
- durations of the current and surrounding syllables
- the quantity of syllable (open, closed)
- the length of vowel (short, long)
- the class of vowel (single, diphthong)
- the class of final consonant (single, clustered)

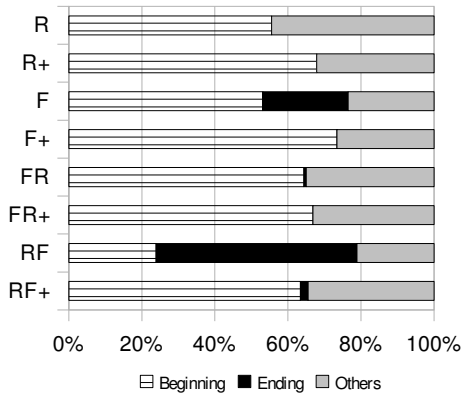

Figure 4: *The distribution of the Thai falling tone by shape types and syllable positions.*

3.3. T-Tilt parameter clustering and prediction

In [2], the exact value of each of eight T-Tilt parameters was directly modeled and predicted. In this paper, we apply k-mean clustering to reduce the variation of five T-Tilt parameters to a certain number of *k* parameter sets for each shape type. Each T-Tilt parameter was normalized using the z-score before clustering. According to our preliminary experiment, 32 clusters were optimal for our model. The RMSE between synthesized and natural F0 curves produced by this clustering model (29.8) was slightly lower than the one without parameter clustering (31.5). Finally, the same set of linguistic features used in shape type prediction and the CART model were applied for T-Tilt parameter prediction.

3.4. F0 contour generation

Finally, the synthesis process to convert the predicted T-Tilt description to F0 contours consists of two steps. First, T-Tilt representations are converted to RFC representations using Equations 5 to 8, derived from Taylor [7]. *syllable_duration* is the duration of the current syllable while the others have been described in the Section 2.

$$A_{rise} = \frac{event\_amp(1+tTilt\_amp)}{2} \quad (5)$$

$$A_{fall} = \frac{event\_amp(1-tTilt\_amp)}{2} \quad (6)$$

$$D_{rise} = peak\_pos \quad (7)$$

$$D_{fall} = syllable\_duration - peak\_pos \quad (8)$$

Then, the Equations 1 to 4 are used to produce the F0 contour of the syllable, the Equations 1 and 2 for the hill shape and the others for the valley shape.

4. Experiments

4.1. Experimental data

A large Thai phonetically-balanced speech corpus, namely TSynC-1 [12] was used in this work. It consists of about 14-hour female read speech tagged with word part-of-speech (POS), handed-labeled phrase breaks, and other useful information. The data set consists of 5,162 utterances, 4,645 utterances for training and the rest for testing. The training utterances contain 144,087 syllables, whereas the test utterances contain 15,856 syllables.

4.2. Experimental measurement

Objective and subjective tests were conducted to measure the difference between synthesized and natural speech. For the objective test, the root mean squared error (RMSE) and correlation between predicted and original F0 contours were computed to examine how well the model works. The original means the F0 contour extracted from the natural speech by Praat [13]. The original contour was preprocessed by deleting isolated spurious F0 points caused by the extraction using thresholds. For the subjective test, the mean opinion score (MOS) was used to examine the perceptual quality of the F0 contour in the synthesized speech. The value 1 to 5 in the MOS scale ranks the naturalness of speech utterances from the worst to the best.

4.3. Experimental results

The first experiment aimed to test our new model fitting approach, which avoided the F0 interpolation and smoothing processes and constrained the T-Tilt shape to always expand over the whole syllable. Results are shown in Table 2. The first row shows results given by the previous T-Tilt model [2], called hereafter the "baseline" model. The second row shows results of our optimized model, called hereafter the "proposed" model. From these results, we can conclude that the proposed model-fitting approach obviously outperforms the baseline model. The major reason of improvement is that parameters produced in the baseline model were biased to interpolated and smoothed signals which were somewhat distorted from their original.

Table 2. *Comparative results of the baseline and proposed model-fitting approaches.*

Model	RMSE	Correlation
Baseline	14.98	0.95
Proposed	9.44	0.98

Table 3. *Comparative results of F0 prediction models with and without F0 shape prediction.*

Model	RMSE	Correlation
Without F0 shape prediction	39.09	0.79
With F0 shape prediction	31.48	0.83

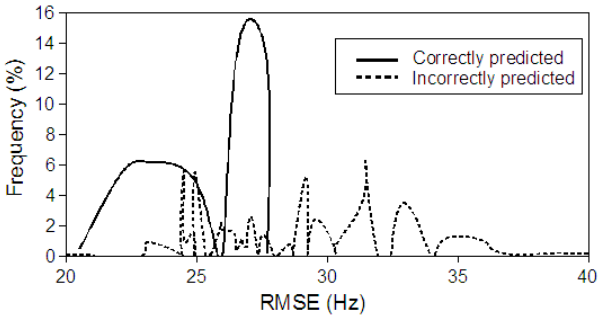

Figure 5: *RMSE distributions of syllable units with correctly (solid line) and incorrectly (dashed line) predicted F0 shapes.*

The second experiment regards the application of F0 shape prediction described in the Section 3.2. The CART-based shape predictor was applied to the optimized T-Tilt model. In evaluation, inputs were texts with word POSs, phrase breaks, and other useful information needed for T-Tilt

parameter prediction, and outputs were synthesized F0 contours. Table 3 presents RMSE and correlation results given by the models with and without F0 shape prediction. Although the shape predictor yielded only 43.1% accuracy, its effectiveness when applying to the T-Tilt model was clearly shown. Figure 5 plots RMSE distributions of all syllable units in the test set with F0 shapes correctly and incorrectly predicted. While the average RMSE of correctly predicted syllables was about 25.7 Hz, more than 28% of wrongly predicted syllables still achieved a lower RMSE. This became a major reason of such effectiveness. An intensive analysis showed that many times, an F0 shape was misclassified to ones with somewhat similar shapes or ones with the desired shape as an ingredient. For examples, R was often classified to be FR, R+ and F were often predicted as FR+, FR and FR+ were often interchanged. Therefore, with similar F0 shapes, the T-Tilt models could be shared.

The next experiment aimed to examine the effect of the syllable position in phrases to the proposed T-Tilt modeling method. We divided all test syllables into three groups; syllables at the beginning, at the end, and at other places in the phrase. As shown in Table 4, we found that the good result came from the syllables located at the end of phrases. This backups the finding in the Figure 4 that the syllable at the end of phrases is less varied and hence easier to model. Also, these results insist the strong relationship between the model performance and the shape variation.

Table 4. *Results separated by the position of syllables in test utterances.*

Syllable position	RMSE	Correlation
Beginning	36.97	0.78
Ending	20.10	0.89
Other places	31.06	0.82

In terms of the subjective test, 4 Thai native females and 7 Thai native males, aged between 21 to 40 years old, were asked to listen to three sets of speech utterances and submit their MOS ranks through a website. The three sets were as follows:

- S1: 15 utterances of natural speech,
- S2: 15 utterances with F0 given by the baseline model,
- S3: 15 utterances with F0 given by the proposed model without F0 shape prediction,
- S4: 15 utterances with F0 given by the proposed model with F0 shape prediction.

It is noted that all proposed models in this experiment were already applied with the k-mean technique for T-Tilt parameter clustering. This technique has proved to be efficient in reducing parameter variations as described in the Section 3.3.

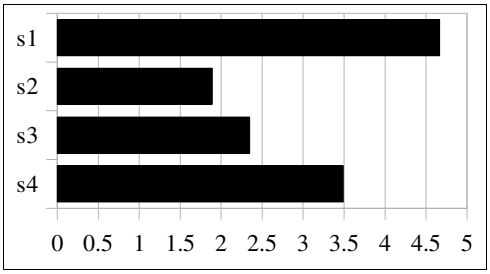

Figure 6: *MOS results in the naturalness test.*

According to results shown in Figure 6, the S1 set was rated unsurprisingly highest. This supports the objective test result shown in the Table 2 that the proposed model stated in the Table 2 (S3) obviously outperforms the baseline model. Since the S4 set was rated significantly higher than that of the S3 set, we can conclude that the F0 shape prediction module clearly helps to improve the model performance. Still some listeners found some syllables in the S4 set unacceptably distorted from their original. The problem was from incorrect F0 shape prediction and improper T-Tilt clustering.

5. Conclusions

In this paper, we have proposed several ways to improve the T-Tilt modeling, a modified Tilt model for F0 analysis and synthesis in tonal languages. We have investigated on a new model fitting approach for T-Tilt parameter estimation. The proposed model drastically reduced the number of T-Tilt parameters from eight to only five. We have found that classifying syllabic F0 curves into eight types and modeling separately for each type could effectively improve the modeling performance. Lastly, we have applied k-mean clustering to reduce the variation of T-Tilt parameters and hence prevent the problem of training data scarcity.

The listening test showed that the naturalness of the synthesized speech was significantly improved from the model without the proposed optimization approach. For tonal languages, the listeners were observed to be very sensitive to every syllable in utterances. If we need the good MOS, F0 curves at every syllable cannot be distorted. In the near future, we will find a solution to eliminate such errors.

6. References

[1] Taylor, P. A., "Analysis and synthesis of intonation using the Tilt model", Journal of the Acoustical Society of America, vol. 107, no.3, pp. 1697-1714, 2000.

[2] Thangthai, A., Thatphitakkul, N., Wutiwiwatchai, C., Rugchatjaroen, A., and Saychum, S., "T-Tilt: a modified Tilt model for F0 analysis and synthesis in tonal laguages" Proc. of Interspeech 2008, pp. 2270-2273, 2008.

[3] Fujisaki, H. and Hirose, K., "Analysis of voice fundamental frequency contours for declarative sentences of Japanese", Journal of the Acoustical Society of Japan (E), 5(4): pp.233-241, 1984.

[4] Pablo, D. A., Antonio, B., Lu, Y., and Juan, C.T, "Intonation modeling of Mandarin Chinese using a superpositional approach", Proc. of Interspeech 2008, pp. 2134-2137, 2008.

[5] Mohler, G. and Conkie, A., "Parametric modeling of intonation using vector quantization", Proc. of the 3rd ESCA Workshop on Speech Synthesis, Jenolan Caves, Australia, 1998.

[6] Xu, Y., "Contextual tonal variation in Mandarin", Ph.D. Dissertation, The University of Connecticut, 1993.

[7] Taylor, P. A., "The Rise/Fall/Connection model of intonation", Speech Communication, vol. 15, pp. 169-186, 1995.

[8] Xu, Y., and Wang, Q. E., "Pitch targets and their realization: evidence from Mandarin", Speech Communication, Vol. 33, pp. 319-337, 2001.

[9] Prom-on, S., "Pitch target analysis of Thai tones using quantitative target approximation model and unsupervised clustering", Proc. of Interspeech 2008, pp. 1116-1119, 2008.

[10] Chompan, S., and Kobayashi, T., "Design of tree-based context clustering for an HMM-based Thai speech synthesis system", Proc. of the 6th ISCA Workshop on Speech Synthesis (SSW6-2007), pp. 160-165, 2007.

[11] Taylor, P. A, Calery, R., and Black, A. W., "The Edinburgh speech tools library", The Centre for Speech Technology Research, University of Edinburgh, Edition 1.2.0, 1999, Available on http://festvox.org/docs/speech_tools-1.2.0/.

[12] Hansakunbuntheung, C., Tesprasit, V., and Sornlertlamvanich, V., "Thai tagged speech corpus for speech synthesis", Proc. of Oriental COCOSDA 2003, pp. 97-104, 2003.

[13] Boersma, P., Weenink, D., "Praat, doing phonetics by computer", April 2007, Available on http://www.fon.hum.uva.nl/praat/.