

Predicting Underlying Pitch Targets for Intonation Modeling

Xuejing Sun

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

Abstract

The present paper reports our preliminary attempt on modeling intonation using underlying pitch targets. The underlying pitch targets were derived using a nonlinear regression technique under the pitch target approximation model [17, 19]. We assume that the use of underlying pitch targets can capture the most important intonation patterns while maintaining critical predictive power. Another important aspect of our approach is that we do not rely on pitch accent as a component in the system. To predict the parameters of the underlying targets, we used a recurrent neural network combined with a time-delay window. Comparing the predicted and original pitch targets, the root mean square error (RMSE) is 7.90 Hz, and the correlation coefficient (r) is 0.78. The results are encouraging and suggesting that the use of underlying pitch targets is a promising approach to intonation modeling.

1. Introduction

Intonation modeling is an important component in speech synthesis. The complex nature of intonation patterns has made their prediction a difficult task. To describe intonation, 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 [15]). Prediction methods include rule-based and data-driven approaches with the latter becoming more popular.

In this study, we present a new approach to intonation modeling, where a time-delay recurrent neural network is trained to predict underlying pitch targets instead of surface F0 contours.

A key idea of the present approach is that we try to model the underlying form of F0 contours rather than the surface form. It is known that the variation of surface F0 contour is the consequence of linguistic constraints, contextual effects and segmental effects. The variation due to higher-level linguistic inputs is often obscured by contextual and segmental effects, which make the identification of the underlying forms difficult. In Xu and Wang [19], a pitch target approximation model was proposed aiming at teasing apart the underlying form from the surface F0 contour. This model was later quantified in Xu, *et al.*[17]. In this model, the surface F0 contour is viewed as the result of asymptotic approximation of the underlying pitch target. A pitch target is defined as the smallest unit that is articulatorily operable. The host unit of a pitch target is assumed to be the syllable (for Mandarin, at least). This model is primarily based on evidence from Mandarin Chinese, where four basic pitch targets (High, Low, Rise, and Fall) are associated with the tonal targets, and can be determined lexically. In English, however, the concept of underlying pitch targets is less

straightforward. In each segment, for instance a syllable, what kind of target a speaker intends to realize cannot be lexically determined. Nevertheless, for English, we can also assume that in a host unit, the speaker is trying to reach a simple target in the form of a straight line, and the parameters are continuous variables.

Another important aspect of our system is that we try to get around the accent prediction process. Many previous approaches are pitch accent based models (e.g. ToBI [10], the Tilt model [15]), or use accent as an important feature to predict intonation. However, we argue that, using accent may have limitations: (1) To train an accent prediction system, a pre-labeled database is necessary. Such labeling process can be labor intensive and time-consuming; (2) Labeling usually involves labelers' subjective interpretation of the sentences. As a result, some acoustic unit may be marked as accented due to its linguistic strength rather than a strong acoustical manifestation. For example, a less frequent word to the listener may sound more like accented than a more frequent word, even though it does not have a strong acoustic manifestation. These "accents" are mixed with the true ones that are indeed more prominent both acoustically and perceptually. In other words, human labeled accents usually reflect both linguistic effects and acoustic effects. This makes them difficult to predict for automatic machine learning methods. In the present study, we assume that underlying pitch targets reflect the intonation patterns that the speaker tries to convey and the listener hears [20]. They can be predicted from higher-level linguistic features. The accent information can be reflected by the height of the pitch target (intercept) and its velocity (slope). For example, for a given duration, a larger slope may more likely indicate an accent. The direction and the intercept of the target can further provide more specific information about the status of the host unit.

In the present study, our central task is therefore to predict the underlying pitch targets. The surface F0 contours can then be obtained by interpolating through these pitch targets. Although the resulting F0 contours may not be very accurate numerically, it should be fairly close perceptually as it presumably retains the most important intonation pattern.

To predict underlying pitch targets, we used a recurrent neural network combined with a time-delay window. Recurrent neural network has been used in prosodic modeling quite extensively (e.g. [2,16]). Its dynamical structure is able to capture the temporal information from input feature space, which is desirable for intonation modeling.

2. The pitch target approximation model

In the following, we first outline the pitch target approximation model based on [17], and then describe its implementation and make some practical modifications. For

the illustration purpose, the expression of the model is somewhat different from that in [17].

The essence of the model is the assumption that the surface F0 contour is the result of an asymptotic approximation to an underlying pitch target. Speaking mathematically, we have:

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

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

where $T(\cdot)$ represents the underlying 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. In other words, it represents the speed at which the underlying target is approached. Parameters a and b are the slope and intercept of the underlying pitch target.

The estimation of these parameters can be done through nonlinear regression. However, as pointed out by [9], the above model or alike does not consistently have good estimation properties. Thus, we need to replace some of the parameters with so-called *expected-value* parameters [9].

Let (t_0, y_0) denote the first point on the F0 contour. By plugging this point into Eq. (2), we can replace parameter β , and have:

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

Note that here we assume $t_0 = 0$.

Next, let (t_1, y_1) denote a point where the exponential component becomes zero, i.e., the underlying pitch target has been approached. Note that an exponential function can never be zero but approximates zero indefinitely. Here we force the exponential part to be zero in order to simplify the model and make the regression analysis more robust. Thus Eq. (3) becomes:

$$y_1 = at_1 + b \quad (4)$$

With Eq. (4) we can replace either a or b in Eq. (3). That is:

$$y(t) = (y_0 - b) \exp(-\lambda t) + \left(\frac{y_1 - b}{t_1}\right)t + b \quad (5)$$

or

$$y(t) = (y_0 - y_1 + at_1) \exp(-\lambda t) + at + y_1 - at_1 \quad (6)$$

After experimentation, we chose Eq. (6) as it showed better estimation properties. We can estimate λ and a using Eq. (6), and derive b using Eq. (4).

The nonlinear regression routine used for estimation is an implementation of the widely used Levenberg-Marquardt algorithm. 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 due to the F0 interpolation. For (t_1, y_1) , currently we use the point in the middle of a segment, which seems to work best. This is also an interesting result as it implies that in general the speaker in the present database presumably approaches the underlying targets near the middle of the host segments.

The above analysis procedures have been tested on both Mandarin Chinese [12] and the current English database, and the results were satisfactory.

To illustrate the concept of the model and the underlying pitch targets, two nonlinear regression results are shown in Figure 1.

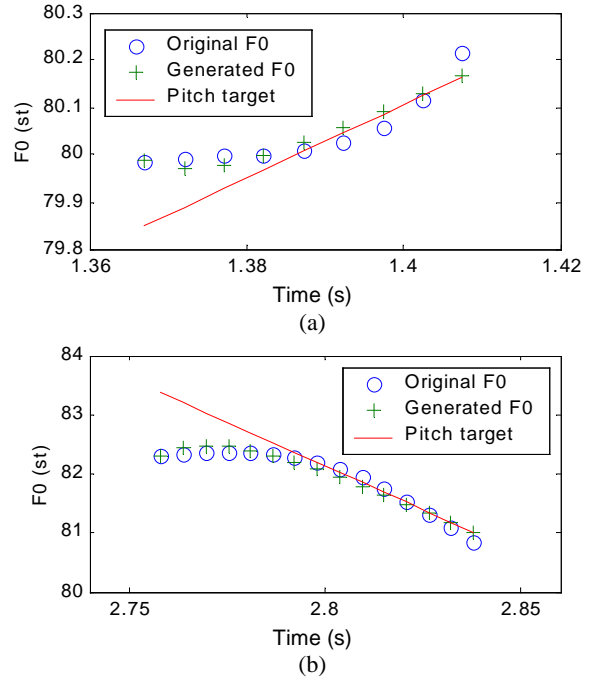


Figure 1. Examples of underlying pitch targets and its surface F0 contours of both original and generated by the model (a) target with positive slope; (b) target with negative slope

The parameters of underlying pitch targets are linguistically meaningful, and we could predict them either using rules or statistical based methods. In the present study, we view these parameters as continuous variables, which would be better handled by data-driven approaches.

Two important caveats should be mentioned about our usage of the pitch target approximation model in the current study. Firstly, by predicting only the underlying pitch targets, we ignored the other two parameters in Eq. (2), namely β and λ . Although theoretically appealing, those two parameters seem to be difficult to predict in practice, at least for the current database. In contrast, the parameters for the underlying targets behave in a more predictable manner. The reconstructed F0 contours from the full parameter set and underlying pitch targets do not have much difference perceptually in most cases. Secondly, another important aspect of the pitch target approximation model is to always assume an appropriate host domain for the pitch targets. In [19], the syllable is assumed to be the host unit for Mandarin. In the present study, we found it more straightforward to work on the vowel domain only, although this might be theoretically sub-optimal. We therefore assume that there is only one pitch target for each vowel. It should be noted that, however, this assumption may not be very accurate as sometime there may be more than one pitch targets in a long vowel, e.g., rise-fall, or fall-rise. Nevertheless, as an initial attempt, we adopted the simplest form.

3. Prediction of underlying pitch targets

3.1. Structure of the neural network

We used a recurrent time-delay neural network in our training and testing tasks. The design of the network follows closely to the work of Ström [11]. NICO toolkit developed by Ström was thus used for training and testing the network, which can be obtained from [7]. The input layer has 33 neurons, which are derived from 9 input features. Instead of using the input features directly, we employed the 1-of-N coding scheme for discrete variables [1], and achieved slightly better results. The hidden layer has 100 “Tanhyp” type units. There are two neurons in the output layer, which are of “Linear” type. Each layer is forward connected to the next layer. There is also a short-cut connection between the input and output layer. To capture the temporal information from the input, the network contains a recurrent connection in the hidden layer, and a time-delay window in the connection from the input layer to hidden layer. The time-delay window includes one look-ahead and one look-back input pattern, respectively. We also utilized the sparse connection feature provided by the toolkit, that is, instead of building a fully connected network, we only randomly generated a fraction of connections. The benefit of this is allowing us to use more hidden units and reduce computational cost [11]. In our network, if we fully connect each layer, 70 neurons, or in general twice as many as the input neurons in the hidden layer seemed to be a reasonable choice. With sparse connections, we used 100 hidden units, and indeed achieved faster convergence and slightly better results. Both input and output values are normalized into the range of [-1, 1]. Training was done with back-propagation through time algorithm.

3.2. Input and output features of the neural network

The input feature set includes the followings, most of which have been commonly used in other systems:

- Vowel duration
- Glissando threshold
- Phone position in a syllable (initial, medial, final, and single)
- Syllable position in a word (initial, medial, final, and single)
- Number of phones in a syllable
- Syllable stress (1 or 0)
- Word position in a sentence (float number between 0 and 1, with 0 representing sentence initial position, and 1 representing sentence final position)
- Number of syllables in a word
- Part-of-speech of the current word (18 categories)

For output features, instead of using the slope and intercept of the target directly, we adopted the following features:

- Absolute ratio
- Direction of the underlying pitch targets (1 or -1)
- Intercept of the underlying pitch targets

Most of these input and output features are straightforward, except for glissando threshold and absolute ratio, which we will explain below.

As we know, some pitch targets are more perceptually significant, and more responsible for the natural intonation. On the other hand, in some cases, even though the slope is very steep, the host vowel does not sound like having a clear pitch movement, i.e., rise or fall, because of the short duration. Thus, we need to weigh the targets differently, and convey this information to the neural network. In attempting to achieve this goal, we computed the so-called *glissando* threshold, a perceptual measure, which is defined as [3,14]:

$$G_{tr} = \frac{0.16}{D^2} \quad (7)$$

where the unit of G_{tr} is semitone/s and D is the duration of the tone.

The glissando threshold means that a pitch movement must have a velocity greater than the threshold in order to be perceivable. From Eq. (7) we can see that a shorter duration of the tone results in a larger glissando threshold. This is reasonable intuitively as it would be more and more difficult to perceive a pitch movement when the duration of the tone becomes shorter. We then divided the slope of the underlying pitch targets with the corresponding glissando threshold. Such a ratio was used as an output feature instead of the slope itself. It is assumed that this nonlinear transformation could roughly make those perceptually more prominent targets have greater absolute ratio values. Experiment results indeed showed its effectiveness.

Experimental evaluation revealed that further splitting the ratio into two features (the absolute value of the ratio, and its sign, i.e., the direction of the target) seemed to give better results. When the slope is zero, we regard it as a positive direction. Thus, the direction of pitch target is a discrete variable with the values 1 and -1. In total there are three output features: absolute ratio, pitch target direction, and intercept of the pitch target.

4. Experimental evaluation

4.1. The speech corpus

The speech corpus used in this study contains 448 TIMIT phonetically balanced sentences read by an American English speaker¹. It was collected at University of Edinburgh’s Center for Speech Technology Research, and can be obtained from [6]. The database is provided with hand-labeled phone-level boundaries, pitchmarks derived from EGG recordings, and Festival utterance structures, such as syllable boundary, lexical stress, pitch accent, etc. We split the database into training and test sets at a ratio of 9:1.

Although pitchmarks derived from EGG is fairly accurate, it retains all the irregularities of vocal folds vibrations, which is not desirable for intonation modeling. Thus, we processed the pitchmarks as follows:

- Convert pitchmarks to F0 values; Interpolate the F0 values using cubic-spline method and smooth the contours using seven-point median and linear filters

¹ The description of the database indicates there are 452 utterances, however we found kdt_411.*, kdt.427.*, kdt_428.pm, and kdt_452.Syllable are missing in our downloaded version.

- Convert F0 values from linear Hz scale to logarithmic semitone (st) scale
- Compute delta F0 values and remove those points that have delta values greater than certain threshold from F0 contours. The choice of threshold value is based on [13,18], in which a maximum speed of pitch change was derived at the value of 120 st/s for the 12 st condition. Since it has been found that the speed is slower with smaller pitch interval [13, 18], and considering that the current database contains read speech with relatively flat F0 contours, 120 st/s was assumed to be a sufficient threshold for the current analysis.

The original database does not provide part-of-speech information. We used TreeTagger program [5] to tag part-of-speech automatically. No manual correction was performed afterwards. Due to the scale of the current database, the original output part-of-speech tag set (36 tags) was reduced to 18 categories.

4.2. Results

The outputs of neural network on the test set were converted back to parameters for underlying targets, i.e. slope and intercept. For objective evaluation of our results, we computed root mean square error (RMSE) and correlation coefficient (r), which have been commonly used in the literature (e.g. Dusterhoff, *et al.* [4]). Since our system is to predict the underlying pitch targets rather than the surface F0 contours, we compared the predicted results from neural network to both the surface F0 contours and the original pitch targets derived by nonlinear regression analysis. Also, to show the eligibility of underlying pitch targets, a comparison between the surface F0 contours and the original underlying pitch targets was conducted. The results are presented in Table 1. Note that all the comparisons were performed on the vowel portion only.

Table 1: Comparison between surface F0 contours, original and predicted underlying pitch targets

	RMSE	r
Surface F0 – Original targets	2.45	0.97
Surface F0 - Predicted targets	7.68	0.77
Original Targets - Predicted targets	7.90	0.78

Table 1 shows that the original underlying targets are very close to the surface F0 contours (RMSE = 2.45; $r = 0.97$), which confirms the capability of underlying targets to represent the surface F0 contours. The predicted targets are reasonably close to both the original targets and surface F0 contours. When examining the details, we found that the predicted targets tend to have small magnitude, which results in more neutral F0 contours.

For the same database, Dusterhoff, *et al.* [4] obtained results with RMSE = 9.1, and $r = 0.74$. Although our results seem slightly better, these two studies may not be directly comparable. In [4], the results are based on the comparison of the original and predicted Tilt parameters for entire F0 contours, while ours were based on comparisons on the vowel portions only. Nevertheless, our results are encouraging and warrant further explorations.

To evaluate the results subjectively, we interpolated both the original and predicted pitch targets, and re-synthesized

sentences using TD-PSOLA [8] technique. The original duration was used in the synthesis. Informal listening tests showed that in most cases the intonation pattern generated from the original pitch targets were almost indistinguishable from those synthesized using the original F0 values. This suggests that the most important intonation patterns could potentially be captured using underlying pitch targets. Sentences synthesized with predicted targets were acceptable, but tended to sound more neutral.

The preliminary results of the present study have shown that modeling intonation via underlying pitch target is promising. With a simple feature set, bypassing the accent layer, where labeling or predicting accents is a nontrivial task, we achieved results comparable to previous studies. This encourages us to explore further along these directions. Future studies should include experiments with a larger database, better feature set, and more sophisticated learning algorithms.

Note that current approach is somewhat similar to the automatic stylization approach in D'Alessandro and Mertens [3]. However, a major difference is that [3] is purely perception based, in which they ignored the small pitch movements that were perceptually insignificant as determined by using the glissando threshold. In our system, the key concept is the underlying pitch targets, which is based on an assumption about the pitch production process [19]. We ignored some parts of the surface F0 contour that are not belong to it underlying targets.

5. Conclusions

In the present study, we tried to model intonation with underlying pitch targets, which was derived from nonlinear regression analysis within the pitch target approximation model framework. The derived pitch targets are meant to reflect the underlying form of F0 contours, which is presumably more predictable than the surface F0 contours. A recurrent neural network combined with a time-delay window was employed to predict underlying targets. Both objective and perceptual evaluations gave promising results. It indicates that the current approach could potentially be used in both theoretical research and practical implementation of intonation models, and could also be used in recognition of prosodic labels, such as in ToBI.

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] Bishop, C. M. *Neural networks for pattern recognition*. Oxford University Press, UK, 1995.
- [2] Chen, S. H., Hwang, S. H., and Wang, Y. R. "An RNN-based prosodic information synthesizer for Mandarin text-to-speech," *IEEE Trans. Speech and Audio Proc.*, 6(3):226-239, 1998.
- [3] D'Alessandro, C., Mertens, P., "Automatic pitch contour stylisation using a model of tonal perception", *Comput. Speech Language* 9, 257-288, 1995.
- [4] 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 99*, 1999.

- [5] Schmid, H. "Improvements in Part-of-Speech Tagging with an Application to German". *Proceedings of the ACL SIGDAT-Workshop*. March 1995.
- [6] <http://www.festvox.org>
- [7] <http://www.speech.kth.se/NICO/index.html>
- [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] Ratkowsky, D. A. *Handbook of nonlinear regression models*, Marcel Dekker, Inc. New York, New York, 1990.
- [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 92*. Banff, Alberta, pp. 867-870, 1992.
- [11] Ström, N. "Phoneme probability estimation with dynamic sparsely connected artificial neural networks," *The Free Speech Journal* (www.cse.ogi.edu/CSLU/fsj/), 5, 1997.
- [12] Sun, X., Xu, C. X., and Xu, Y., "Experiment on pitch target approximation model for generating Mandarin F0 contour", *141st ASA meeting*, Chicago, 2001
- [13] Sundberg, J. "Maximum speed of pitch changes in singers and untrained subjects", *J. Phon.* 7, 71-79, 1979.
- [14] 't Hart, J., Collier, R., and Cohen, A. *A perceptual study of intonation*, Cambridge University Press, UK, 1990.
- [15] Taylor, P.A., "Analysis and synthesis of intonation using the Tilt model", *J. Acoust. Soc. Am.* 107, 1697-1714, 2000.
- [16] 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.
- [17] Xu, C. X., Xu, Y. and Luo, L-S. "A pitch target approximation model for F0 contours in Mandarin", *Proc. of The 14th International Congress of Phonetic Sciences*, San Francisco. pp. 2359-2362, 1999.
- [18] Xu, Y. and Sun, X., "How fast can we really change pitch? Maximum speed of pitch change revisited", *Proc. of The 6th International Conference on Spoken Language Processing*, Beijing, pp. 666-669, 2000.
- [19] Xu, Y. and Wang, E., "Pitch targets and their realization: Evidence from Mandarin Chinese", *Speech Communication* 33 (4), 319-337, 2001.
- [20] Xu, Y. and Wang, E., "Phonetic targets as the link between speech production and speech perception", *J. Acoust. Soc. Am.* 108, Pt. 2, 2531-2532, 2000.

