Totally Data-driven Intonation Prediction Model Using a Novel F0 Contour Parametric Representation

Lifu Yi, Jian Li, Xiaoyan Lou, Jie Hao

Toshiba (China) Research and Development Center
{yilifu, lijian, louxiaoyan, haojie}@rdc.toshiba.com.cn

ABSTRACT

This paper proposes a novel parametric representation of mandarin intonation based on orthogonal polynomial approximation. The polynomial is a simplified representation of Parallel Encoding and Target Approximation (PENTA) intonation model that includes a target component and an approximation component. We also propose predicting the polynomial parameters from linguistic and phonetic attributes by generalized linear models (GLM). The optimal attributes are automatically selected by stepwise regression method. Thus both model structures and model coefficients are optimized in a totally data-driven manner. In addition, speaking rate is introduced as a new attribute for prediction. When the method is applied to intonation prediction of Mandarin speech, it achieves F0 RMSE of 30.21 Hz and correlation coefficients of 0.85 in open test. Informal perceptual experiments show that the predicted intonation is quite appropriate and natural.

Index Terms: intonation prediction, F0 contour parametric Representation, generalized linear models, speech synthesis

INTRODUCTION

Intonation modeling is a crucial issue affecting the intelligibility and naturalness of speech synthesis systems. In general, intonation modeling is divided into two steps. The first step is to represent F0 contour by a parametric or non-parametric model. The second step is to use data-driven methods to model the relationship between the representation parameters and linguistic/phonetic attributes. The representation model should represent F0 contour accurately and stably. The prediction of the representation parameters should be mathematically tractable.

In recent years, several parametric and non-parametric intonation models [1] [2] are proposed for F0 representation. Non-parametric model samples the original F0 contour directly. Fujisaki [3] [4], and PENTA [5] are two typical parametric models. Fujisaki model represents F0 contour by linear combination of long-term and short-term components, i.e., phrase and accent components. It is a superposition model and offers a distinct physiological interpretation that connects F0 contour with dynamics of the larynx. Several methods are proposed to automatically extract Fujisaki parameters. However it is still difficult because the decomposition of phrase and accent components is not unique [3] [5]. We also do not know the boundary between long-term and short-term components exactly because these components are not simply added together on a linear or nonlinear scale [5]. PENTA model is proposed by Xu [5] [6] [7], which is a typical linearly sequenced model and pays more attention on local events than big prosodic units in Fujisaki model. Xu and Sun proposed a simplified version of PENTA parametric representation in [7] [8]. However, unfortunately, parameter estimation of this representation is still complex and sometime unstable [8] [9].

Several machine-learning techniques are proposed to predict F0 parameters from linguistic/phonetic attributes, such as CART [1], Generalized Additive models [10] and neural networks [11]. These techniques are powerful, but their linguistic/phonetic attributes and attributes interactions are guided by existing knowledge [1] [2] [8], i.e., manually but not in a totally data-driven manner. Moreover, the techniques are too complex to have an intuitive interpretation.

In this paper, we propose a novel polynomial F0 representation model under pitch target approximation hypothesis [5]. The parameters are believed to be linguistically/phonetically meaningful for representation and efficient for prediction. Furthermore, we propose predicting the parameters by using generalized linear models [12]. Stepwise regression is used to automatically select the optimal attributes and attribute interactions. The whole process is performed in a totally data-driven manner.

This paper is organized as follows: in section 2, we describe the polynomial F0 representation model. We introduce basic concept of GLM and stepwise regression for parameter prediction in section 3. In section 4 we give the representation and prediction experimental results. Finally, we draw the main conclusions of this work.

2. F0 REPRESENTATION MODEL

The basic formulation of the F0 parametric representation is similar to the existing approaches [6] [7] [8], and we extend the basic assumption by using orthogonal polynomials.

Sun proposed a parametric representation of target approximation F0 model in [8]. This representation consists of an underlying target F0 component and an approximation F0 component, and defined as follows.

$$T(t) = at + b \tag{1}$$

$$y(t) = \beta \exp(-\lambda t) + at + b \tag{2}$$

Where $T(\cdot)$ represents the underlying F0 target, $\beta \exp(-\lambda t)$ is the approximation F0 component, y(·) represents the surface F0 contour. t is time, β and λ are model parameters. However, this representation still exhibits complex behaviors in parameter estimation by nonlinear regression techniques, just because the exponential approximation component is instable to estimate [8].

To overcome the instability of Sun's model, we assume that: a) In Mandarin syllable, F0 target component is approached symmetrically. The middle part of syllable is the most stable for target F0 representation. b) The target and approximation components are independent.

Furthermore, we assume the components are orthogonal to each other. In experiments, we find that the second-order Legendre polynomials are suitable for the representation, and can be considered as approximations of Taylor's expansion of Eq.(2). The F0 representation is defined as follows:

$$T(t_N) = \beta_0 p_0(t_N) + \beta_1 p_1(t_N) = \beta_0 + \beta_1 t_N \quad (3)$$

$$y(t_N) = T(t_N) + \beta_2 p_2(t_N) = \beta_0 + \beta_1 t_N + 0.5\beta_2(3t_N^2 - 1) \quad (4)$$

Where t_N is the normalized segmental time, $t_N \in [-1,1]$. β_0, β_1 and β_2 are Legendre coefficients. $\{p_n(t_N)\}$ are the classes of Legendre polynomials that obey an orthogonality constraint in Eq.(5). δ_{mn} in Eq(5) is the Kronecker delta and $c_n = 2/(2n+1)$. The first three Legendre polynomials are shown in Eq. (6-8).

$$\int_{-1}^{1} P_m(t_N) P_n(t_N) dt_N = \delta_{mn} c_n \quad (5)$$

$$p_0(t_N) = 1 \quad (6)$$

$$p_1(t_N) = t \quad (7)$$

$$p_2(t_N) = \frac{1}{2}(3t_N^2 - 1) \quad (8)$$

There are two main differences between our F0 representation and Sun's: firstly, we use an orthogonal quadratic approximation instead of the exponential approximation of Sun's; secondly, we normalize the segmental duration range to [-1, 1]. By the orthogonal parametric form, the target and approximation components are totally separated. This makes parameter estimation easier as shown in section 4.

3. PARAMETER PREDICTION APPOACH

This paper models the F0 representation parameters by GLM. The optimization criteria are F-test and BIC. The optimal attributes and attributes interactions are selected by stepwise regression in a totally data-driven manner.

3.1 GLM

GLM is a generalization of multivariate linear regression model [12]. The GLM model predicts each F0 parameters $\hat{\beta}$ from attribute vector $\boldsymbol{A}$ of speech unit s by:

$$\hat{\beta} = h^{-1}(\gamma_0 + \sum_{i=1}^{p} \gamma_i a_i) \quad (9)$$

Where a_i is an attribute or attribute interaction, γ_i is a regression coefficient, p is the dimension of the GLM. And h is a link function. By using different link function h, we can get different statistical distribution of F0 representation parameters β. Here we assume Gaussian distribution for β, so h equals identity function accordingly. In GLM, a_i can be either an attribute or an attribute interaction, e.g., a_i can be defined as $(a'_{in} \times a'_{im})$, where a'_{in} and a'_{im} are attributes from attribute vector $\boldsymbol{A}$.

To measure performances of GLMs, we use Bayes Information Criterion (BIC), base on which we can choose a model that optimizes the trade-off between model complexity and goodness-of-fit.

3.2 Stepwise regression

We keep only the linear attributes and the second order attribute interactions in Eq.(9) for initialization. Stepwise regression can automatically selects the most important attributes and attribute interactions by an iterative training process as shown in Fig.1.

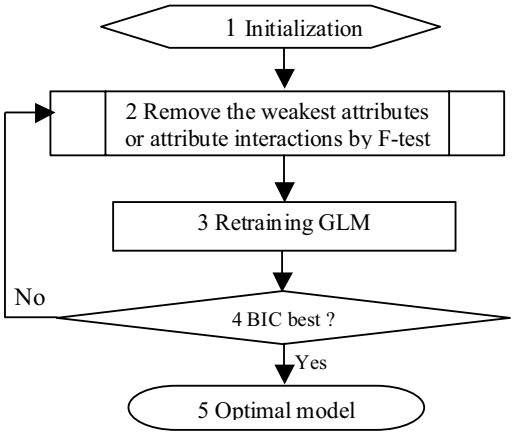

Fig.1: Stepwise regression of F0 representation parameters

For example, supposing that the F0 parameter β_0 is affected only by attributes "phone" (a_1) and "tone" (a_2), we will have the following model:

$$\hat{\beta}_0 = \gamma_0 + \gamma_1 a_1 + \gamma_2 a_2 + \gamma_{12}(a_1 \times a_2) \quad (10)$$

Where $a_1 \times a_2$ means the interaction (combination) of phone and tone. Eq.(10) is the initial model. Then we calculate F-test scores of each item. Maybe $a_1 \times a_2$ is the least important item, if so we just remove it and retrain the model without $a_1 \times a_2$ item. Then we calculate the BIC. If the BIC is minimized, we can stop here and get the optimal model in Eq.(11).

$$\hat{\beta}_0 = \gamma_0 + \gamma_1 a_1 + \gamma_2 a_2 \quad (11)$$

This attributes/attribute interactions selection process is off-line and totally data-driven. Similar optimal processes will be performed for β_0, β_1 and β_2. Finally we get the optimal attribute/attribute interaction sets of βs.

4. EXPERIMENTAL EVALUATION

4.1. Experimental conditions

The models described above were trained and tested using our mandarin corpus. The corpus is narrated by a professional female broadcaster and contains 2,150 utterances sampled at 22.05 kHz. The pitch contours were extracted from the corpus by an autocorrelation algorithm with a 1ms resolution, then manually corrected and finally smoothed by a low-pass filter.

Theoretically, all linguistic and phonetic attributes are likely to influence F0. We use generally considered attributes, such as tone, POS (part-of-speech) and other contextual information. This attribute set is similar with that of other mandarin F0 research literatures [8][9]. The difference is that we introduce speaking rate as a new attribute, and it may interact with other attributes. Including speaking rate into the attribute set can improve the model precision as shown in section 4.2.

4.2 Experimental results

For a preliminary objective evaluation, we measure the goodness of fit in term of root mean square error (RMSE) and correlation coefficient (Corr) in this paper, which are often used in evaluation of F0 modeling [8],[10]. RMSE and Corr in this paper are calculated in the linear frequency domain (in Hertz).

4.2.1 Representation results

We compare the original F0 contour with the parametric ones by Eq.(2) of Sun and Eq.(4). Due to too many parameter estimation failures of Sun [8], we choose only the first 200 syllables in our corpus that were successfully parameterized by both Eq.(2) and Eq.(4). F0 values at voiced portion of every syllable are taken into account and the RMSE and Corr are calculated respectively for the 200 syllables. Table 1 shows the RMSE and Corr averaged on all the syllables. Our method is significantly better than Sun's.

Table 1: Comparison between parametric F0 representations

Parametric representation	RMSE	Corr
Proposed orthogonal representation	3.84	0.9555
Exponential presentation by Sun[8]	5.13	0.9344

Fig.2 compares the examples for the origin F0 contour and the parameterized ones by Sun [8] and the proposed orthogonal polynomial representation. It shows that the orthogonal polynomial model fits the real F0 contour better than Sun [8]. The estimation of the coefficients will not fail only if the number of sampled points of origin F0 contour is above three. We also calculated a RMSE and a Corr between the original F0 and proposed parameterized F0 for all syllables in the corpus, they are 11.15 Hz and 0.9812 respectively.

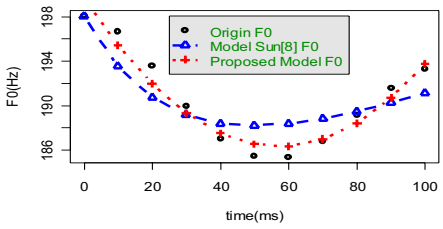

Fig.2*:* Examples for original surface F0 contour and parameterized ones.

The experiments show that the second-order orthogonal polynomial is sufficient for fitting F0 contours of mandarin syllables.

4.2.2 Prediction results

In prediction evaluation, the corpus was divided into a train set (75%) and a test set (25%). Since most previous works adopted CART as prediction method, we also implemented CART prediction for our parametric representation for comparison.

Table 2: Comparison between original F0, parameterized F0, and predicted F0 from proposed GLM method and CART method

	RMSE	Corr
Parameterized F0 – Predicted F0 (CART)	30.55	0.8488
Parameterized F0-Predicted F0 (GLM, no SpRate)	30.19	0.8502
Parameterized F0 - Predicted F0 (GLM)	**28.99**	**0.8661**
Original F0 – Predicted F0 (CART)	32.04	0.8355
Original F0–Predicted F0(GLM, no SpRate)	31.71	0.8369
Original F0 – Predicted F0 (GLM)	**30.54**	**0.8528**

Table 2 summaries comparison results between original F0, predicted F0 and Parameterized F0 in open tests. It can be seen from table 2, in terms of both RMSE and Corr, the proposed GLM method is consistently better than the CART method. We also compare the GLM model without speaking rate attribute that is significantly worse than the GLM with speaking rate. This tells us that including speaking rate as a new attribute does improve the prediction performances.

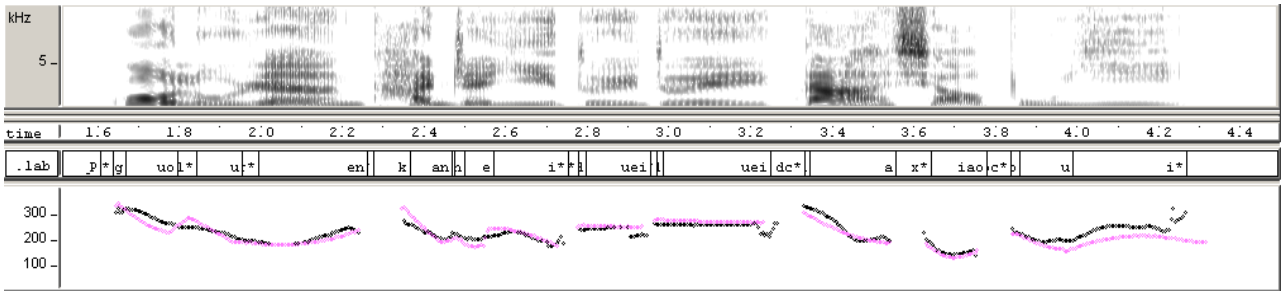

Fig.3: Examples for GLM predicted F0 contour (pink) and the original F0 contour (dark)

Fig.3 illustrates an example of the GLM predicted F0 contour and the original F0 contour for a sentence in the test set. The GLM predicted F0 contour is quite smooth and close to the original F0 contour.

Table 3: Comparison with other approaches

Method	RMSE	Corr
Proposed method (Mandarin)	30.5	0.85
Chen [9] (Mandarin)	22.5	0.72
Sun[8] (English)	33.1	0.72
Sakai [10] (English)	37.6	0.63
Agüero [2] (English)	14.4	0.71

Table 3 presents comparison results with other approaches. Due to the differences among languages, corpus, it is difficult to compare RMSEs directly. On the other hand, the proposed method achieves higher correlation.

5. CONCLUSIONS

In this paper, we propose a novel parametric F0 representation model from pitch target approximation hypothesis. It's simple, stable and efficient for F0 representation and prediction. Furthermore, we bring up generalized linear model to predict the parameters in this representation model. The optimal attributes and attribute interactions are automatically selected by stepwise regression based on F-test and BIC criteria. Therefore, the prediction method proposed in this paper is totally data-driven.

The F0 presentation and prediction methods can be used as a part of corpus-based or parametric speech synthesizer. It may also be used as in emotional prosody and speech conversion research. We also hope to apply this framework of F0 modeling/prediction to other languages, such as Japanese and English.

6. REFERENCES

[1] K. Dusterhoff, A. Black, and P. Taylor, "Using decision trees within the tilt intonation model to predict F0 contours," in *Proc. EUROSPEECH'99*, pp.1627–1630, 1999.

[2] T. Kagoshima, M. Morita, S. Seto and M. Akamine, "An F0 contour control model for totally speaker driven text to speech system," in *Proc. ICSLP98*, pp.1975-1978, 1998.

[3] Pablo Daniel Agüero, Klaus Wimmer and Antonio Bonafonte., "Joint Extraction and Prediction of Fujisaki's Intonation Model Parameters," *ICSLP 2004*, Jeju Island, Korea. 2004.

[4] Mixdorff, H., "A Novel Approach to the Fully Automatic Extraction of Fujisaki Model Parameters," in *Proc. ICASSP 2000*, pp.1281-1284, Turkey, 2000.

[5] Xu, Y., "The PENTA model of speech melody: Transmitting multiple communicative functions in parallel," in *Proceedings of From Sound to Sense: 50+ years of discoveries in speech communication*, Cambridge, MA, C-91-96, 2004.

[6] Xu, Y. and Wang, Q. E. "Pitch targets and their realization: Evidence from Mandarin Chinese," *Speech Communication* 33(4), pp.319-337, 2001.

[7] Xu, C. X., Xu, Y. and Luo, L-S., "A pitch target approximation model for F0 contours in Mandarin," in *Proceedings of The 14th International Congress of Phonetic Sciences*, San Francisco, pp.2359-2362, 1999.

[8] X. Sun, "F0 generation for speech synthesis using a multi-tier approach," in *Proc. ICSLP'02*, pp.2077-2080.

[9] Gaopeng Chen, Yu Hu and Renhua Wang, "Pitch Target Model's Realization Considering Speech Speed and Environment," *Journal of Chinese Information Processing*, Vol.18, No.3, pp. 81-85, 2004.

[10] S. Sakai, "Addictive Modeling of English F0 Contour for Speech Synthesis," in *Proc. ICASSP*05, pp.277-280, Philadelphia, 2005.

[11] Holm, B.Bailly and Gérard, "Learning the hidden structure of intonation n: implementing various functions of prosody," *Speech Prosody 2002*, pp. 399-402, France, 2002.

[12] McCullagh P and Nelder JA, *Generalized Linear Models*, Chapman & Hall, 1989.