INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

On Smoothing and Enhancing Dynamics of Pitch Contours Represented by Discrete Orthogonal Polynomials for Prosody Generation

Chen-Yu CHIANG

Dept. of Communication Engineering, National Taipei University, New Taipei City, Taiwan
cychiang@mail.ntpu.edu.tw

Abstract

This paper presents a new pitch contour generation algorithm for statistical syllable-based logF0 generation models which represent logF0 contours of syllables by coefficients of discrete orthogonal polynomials, i.e. orthogonal expansion coefficients (OECs). The conventional statistical logF0 models can generate smooth pitch contour within a syllable because of the continuity property of polynomials. However, the models do not ensure to produce continuous and smooth logF0 contours in the proximity of syllable junctures. Besides, dynamic range of the generated logF0 contours is generally smaller than the one of real speech. The above two shortcomings would result in unnatural and monotonous prosody. To overcome these shortcomings, juncture-smooth and dynamics-enhancing OEC generation algorithms are hence proposed in this paper. Analysis on the generated logF0 contours by the proposed algorithm shows some improvements in logF0 smoothness at syllable junctures and enhanced logF0 dynamic range. In addition, a perceptual evaluation of the logF0 contour generated by the proposed algorithm shows an improvement in naturalness of the synthesized speech.

Index Terms: prosody, pitch contour, orthogonal expansion polynomial, text-to-speech system

1. Introduction

Discrete orthogonal polynomials are widely used to represent syllabic pitch contours of Mandarin [1-6] and Chinese dialects [7]. In Chen and Wang's study of vector quantization of pitch information for Mandarin [1], the pitch contour of each syllable is parameterized by a 3-rd order discrete orthogonal polynomial expansion expressed by

$$\hat{F}(i)=\sum_{j=0}^{3}a(j)\cdot\phi_j(i/M)\quad \text{for } i=0\sim M, \tag{1}$$

where $a(j)$ is called the orthogonal expansion coefficient (OEC) for j-th orthogonal polynomial basis, $\phi_j(i/M)$; $\hat{F}(i)$ represents a reconstructed pitch value of i-th voicing frame; M+1 is the length of the syllable pitch contour. The bases are normalized to [0,1] in length and are expressed by:

$$\begin{aligned}
&\phi_0(i/M)=1\\
&\phi_1(i/M)=[12\cdot M/(M+2)]^{1/2}\cdot[i/M-0.5]\\
&\phi_2(i/M)=[\frac{180\cdot M^3}{(M-1)(M+2)(M+3)}]^{1/2}\cdot[(\frac{i}{M})^2-\frac{i}{M}+\frac{M-1}{6\cdot M}]\\
&\phi_3(i/M)=[\frac{2800\cdot M^5}{(M-1)(M-2)(M+2)(M+3)(M+4)}]^{1/2}\\
&\cdot[(\frac{i}{M})^3-\frac{3}{2}(\frac{i}{M})^2+\frac{6M^2-3M+2}{10\cdot M^2}(\frac{i}{M})-\frac{(M-1)(M-2)}{20\cdot M^2}]
\end{aligned}\tag{2}$$

These bases are chosen since they can represent basic patterns of logF0 contours for Chinese tones. Their corresponding coefficients, i.e. OECs, can be found by

$$a(j)=\tfrac{1}{M+1}\sum_{i=0}^{M}F(i)\cdot\phi_j(i/M)\qquad j=0\sim 3 \tag{3}$$

where $F(i)$ is the observed i-th voicing frame of a syllable. Therefore, the pitch contour of n-th syllable in an utterance, $\mathbf{sp}_n$, can be represented by a four-dimensional vector:

$$\mathbf{sp}_n=[a_n(0)\quad a_n(1)\quad a_n(2)\quad a_n(3)]^T \tag{4}$$

where $a_n(0)$, $a_n(1)$, $a_n(2)$, and $a_n(3)$ represent respectively the mean, slope, acceleration and curvature of the logF0 contour. Generally, the error between the observed pitch contour, $F(i)$, and the reconstructed one, $\hat{F}(i)$, is very small. Therefore, the OECs of syllables can be directly taken as prediction targets in prosody generation tasks [1-7]. In the previous studies [2,6], logF0 generation models were formulated based on a maximum likelihood (ML) criterion:

$$\mathbf{sp}_n^*=\boldsymbol{\beta}(t_{n-1}^{n+1},p_n,B_{n-1}^n)=\text{argmax}_{\mathbf{sp}_n}N(\mathbf{sp}_n;\boldsymbol{\beta}(t_{n-1}^{n+1},p_n,B_{n-1}^n),\mathbf{V}_n) \tag{5}$$

where $\mathbf{sp}_n^*$ is the ML-generated syllable logF0 contour; $\boldsymbol{\beta}(\cdot)$ is the mean vector of a Gaussian distribution with covariance matrix $\mathbf{V}_n$, which is a function of the tone triplet, $t_{n-1}^{n+1}=(t_{n-1},t_n,t_{n+1})$, the pitch prosodic state, p_n, and the adjacent prosodic break types $B_{n-1}^n=(B_{n-1},B_n)$. It is noted that $\boldsymbol{\beta}(\cdot)$ can be obtained by superimposing several four-dimensional vectors (i.e. OECs) which represent effects of tone, prosodic state, or prosodic break [5,6], i.e.

$$\boldsymbol{\beta}(t_{n-1}^{n+1},p_n,B_{n-1}^n)=\boldsymbol{\beta}_{t_n}+\boldsymbol{\beta}_{p_n}+\boldsymbol{\beta}^f_{B_{n-1},tp_{n-1}}+\boldsymbol{\beta}^b_{B_n,tp_n}+\boldsymbol{\mu}_{sp} \tag{6}$$

The pitch modeling by OECs is advantageous in several aspects. First, logF0 contours in different lengths can be parameterized by vectors with identical dimensions, i.e. four OECs, so as to facilitate a syllable-based pitch modeling. Second, the OECs can well describe a smooth pitch contour of a syllable since the fitting error between the observed logF0 contour and the reconstructed one is negligibly small. Third, the logF0 contours can be easily manipulated for different lengths caused by different speaking rate (SR) [6] since the bases, $\phi_j(i/M)$, are normalized to [0,1] and the length can be simply specified by the parameter M.

However, the ML-based logF0 generation in Eq. (5) does not ensure to produce smooth consecutive logF0 contours on syllable junctures. By analyzing the logF0 generation results, this discontinuity usually takes place at junctures of a preceding voiced final and a following voiced initial and results in unnatural prosody. Besides, the ML-based generated OECs encounter an over-smoothing problem that the dynamic ranges of the ML-based generated OECs are generally narrower than the ones of the authentic OECs, resulting a monotonous prosody. It can be seen from a typical example of this discontinuity and the over-smoothing problem in Figure 1

http://dx.doi.org/10.21437/Interspeech.2016-409

that the original F0 contours are smoothly connected at some syllable junctures but the predicted F0 contours are not. The F0 dynamic range of the authentic speech is wider than the the synthesized one. Therefore, to tackle the above-mentioned two issues, this paper proposes a juncture-smooth OEC generation algorithm to ensure smoothness for pitch contours on voiced syllable junctures. Based on the proposed smooth OEC algorithm, a tonality-enhanced OEC algorithm and a dynamics-controlled OEC algorithm are hence developed to increase dynamic range of predicted OECs.

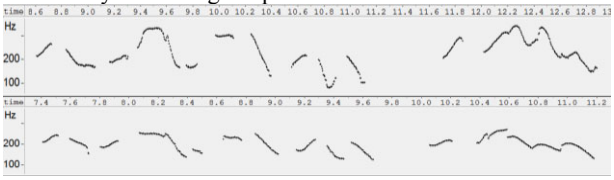

Figure 1: *Examples of F0 contours for authentic speech (upper) and the synthesized speech (lower)*

2. Juncture-Smooth OEC Generation

The total log-likelihood function for ML-based logF0 generation of an utterance is expressed by

$$L(\mathbf{x}) = -\tfrac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \mathbf{a}_n)^T \mathbf{V}_n^{-1}(\mathbf{x}_n - \mathbf{a}_n) \quad (7)$$

where N represents number of syllable in an utterance; $\mathbf{a}_n$ is a 4-by-1 vector composed of four OECs representing mean vector of the Gaussian distribution generated by a prosodic model given with some linguistic/prosodic context of n-th syllable; $\mathbf{V}_n$ is a 4-by-4 covariance matrix for the Gaussian; $\mathbf{x}_n$ represents the predicted OEC vector for n-th syllable. To ensure smoothness of logF0 contours at voiced-continuous syllable junctures, the following two conditions are required:

$$F_{\hat{n}_k}(I_{\hat{n}_k}+1)=F_{\hat{n}_k+1}(0) \text{ and } F_{\hat{n}_k}(I_{\hat{n}_k})=F_{\hat{n}_k+1}(-1) \text{ for } k=1\sim K \quad (8)$$

where $\hat{n}_k \in [1, N-1]$ is a syllable index of k-th smooth juncture, i.e. the pitch contours of $\hat{n}_k$-th and $(\hat{n}_k+1)$-th syllables are smoothly connected at the juncture between $\hat{n}_k$-th and $(\hat{n}_k+1)$-th syllables (referred as $\hat{n}_k$-th juncture); K is number of smooth junctures; $F_{\hat{n}_k}(i)$ is the OEC expanded logF0 value of i-th frame in $\hat{n}_k$-th syllable obtained by

$$F_{\hat{n}_k}(i) = \sum_{j=0}^{3} \mathbf{x}_{\hat{n}_k}(j)\cdot\phi_j(i/I_{\hat{n}_k}), \text{ for } \hat{n}_k = 1\sim I_{\hat{n}_k}; \quad (9)$$

$I_{\hat{n}_k}+1$ is the length of the voiced frame for $\hat{n}_k$-th syllable; $\mathbf{x}_{\hat{n}_k}(j)$ represents j-th element of the OEC vector; $F_{\hat{n}_k}(I_{\hat{n}_k}+1)$ and $F_{\hat{n}_k+1}(-1)$ means respectively a right extrapolated logF0 values from $\hat{n}_k$-th syllable and a left extrapolated logF0 value from $(\hat{n}_k+1)$-th syllable. If one of the conditions $F_{\hat{n}_k}(I_{\hat{n}_k}+1)=F_{\hat{n}_k+1}(0)$ or $F_{\hat{n}_k}(I_{\hat{n}_k})=F_{\hat{n}_k+1}(-1)$ is satisfied, the logF0 contour at $\hat{n}_k$-th juncture is continuous. Furthermore, if the two conditions are both satisfied, the logF0 contour is smoothly connected at the juncture.

The problem to generate juncture-smooth OECs can be found by the following objective function:

$$O(\mathbf{x},\boldsymbol{\lambda}^f,\boldsymbol{\lambda}^b) = L(\mathbf{x}) + \sum_{k=1}^{K}\lambda_k^f\left[F_{\hat{n}_k}(I_{\hat{n}_k}+1)-F_{\hat{n}_k+1}(0)\right] + \sum_{k=1}^{K}\lambda_k^b\left[F_{\hat{n}_k}(I_{\hat{n}_k})-F_{\hat{n}_k+1}(-1)\right] \quad (10)$$

where λ_k^f and λ_k^b are the Lagrange multipliers respectively for right and left extrapolated logF0 constraints. Since the juncture-smooth constraints stated in Eqs. (8) and (10) are linear transformations of the predicted logF0, $\mathbf{x}$, the objective function Eq. (10) can be re-written by a matrix form:

$$O(\mathbf{x},\boldsymbol{\lambda}) = O(\mathbf{x},\boldsymbol{\lambda}^f,\boldsymbol{\lambda}^b) = -\tfrac{1}{2}(\mathbf{x}-\mathbf{a})^T\mathbf{R}(\mathbf{x}-\mathbf{a}) + \boldsymbol{\lambda}^T\mathbf{Z}\mathbf{x} \quad (11)$$

where $\mathbf{x} = \begin{bmatrix}\mathbf{x}_1^T & \mathbf{x}_2^T & \dots & \mathbf{x}_N^T\end{bmatrix}^T$ and $\mathbf{a} = \begin{bmatrix}\mathbf{a}_1^T & \mathbf{a}_2^T & \dots & \mathbf{a}_N^T\end{bmatrix}^T$; $\mathbf{R}$ is a squared matrix formed by the inverse matrixes of the covariance matrixes, $\mathbf{V}_n$; $\boldsymbol{\lambda}$ is a vectorial Lagrange multiplier formed by λ_k^f and λ_k^b; $\mathbf{Z}$ is a linear transformation matrix formed by the smooth-juncture constraint in Eq. (8) in terms of the linear combinations of bases $\phi_j(\cdot)$. Then, the optimal pitch OECs can be derived by $\partial O(\mathbf{x},\boldsymbol{\lambda})/\partial\mathbf{x} = \mathbf{0}$, resulting in

$$\mathbf{x}^* = (\mathbf{R}^T)^{-1}\mathbf{Z}^T\boldsymbol{\lambda} + \mathbf{a} \quad (12)$$

By knowing the juncture-smooth constraints made by Eq. (8), i.e. $\mathbf{Z}\mathbf{x}^* = \mathbf{0}$, $\boldsymbol{\lambda}$ is obtained by

$$\boldsymbol{\lambda} = -\left[\mathbf{Z}(\mathbf{R}^T)^{-1}\mathbf{Z}^T\right]^{-1}\mathbf{Z}\mathbf{a} \quad (13)$$

3. Dynamics-Enhanced OEC Generation

To increase the dynamic ranges of the generated OECs, a new objective function is defined by

$$U(\mathbf{x},\boldsymbol{\lambda}) = O(\mathbf{x},\boldsymbol{\lambda}) + \tfrac{N}{2}\sum_{j=0}^{3} w_j v(j) \quad (14)$$

where $O(\mathbf{x},\boldsymbol{\lambda})$ is the objective function defined in Section 2; $v(j)$ is the variance of j-th dimension of the generated OEC vector for an utterance; w_j is a weight for variance of j-th dimension. The significance of w_j is to tune the dynamics of the OECs of j-th dimension. Generally, as w_j is larger, the dynamic range of the generated OECs of j-th dimension is larger. The variance, $v(j)$, is expressed by

$$v(j) = \tfrac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n(j))^2 - \left(\tfrac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n(j)\right)^2 \quad (15)$$

We can rewrite Eq. (15) by a matrix equation:

$$v(j) = (\mathbf{C}_j\mathbf{x})^T(\mathbf{C}_j\mathbf{x}) - (\mathbf{e}^T\mathbf{C}_j\mathbf{x})^T(\mathbf{e}^T\mathbf{C}_j\mathbf{x}) = \mathbf{x}^T\mathbf{D}_j\mathbf{x} \quad (16)$$

where $\mathbf{C}_j$ is an $4N$-by-$4N$ squared matrix in which the element at r-th row and s-th column is defined by

$$\mathbf{C}_j(r,s) = \begin{cases} 1/\sqrt{N}, & \text{if } r=s \text{ and } (r \bmod 4) = j \\ 0, & \text{otherwise} \end{cases}; \quad (17)$$

$\mathbf{e}$ is a $4N$-by-1 vector expressed by

$$\mathbf{e}_{4N\times 1} = \begin{bmatrix}1/\sqrt{N} & \dots & 1/\sqrt{N}\end{bmatrix}^T; \quad (18)$$

and $\mathbf{D}_j = (\mathbf{C}_j)^T(\mathbf{I}-\mathbf{e}\mathbf{e}^T)\mathbf{C}_j$ is an $4N$-by-$4N$ symmetric matrix. The objective function in Eq. (14) can be re-written by

$$U(\mathbf{x},\boldsymbol{\lambda}) = -\tfrac{1}{2}(\mathbf{x}-\mathbf{a})^T\mathbf{R}(\mathbf{x}-\mathbf{a}) + \boldsymbol{\lambda}^T\mathbf{Z}\mathbf{x} + \tfrac{N}{2}\sum_{j=0}^{3} w_j\mathbf{x}^T\mathbf{D}_j\mathbf{x} \quad (19)$$

The optimal OECs, $\mathbf{x}^*$, given with the objective function $U(\mathbf{x},\boldsymbol{\lambda})$, can be derived by $\partial U(\mathbf{x},\boldsymbol{\lambda})/\partial\mathbf{x} = \mathbf{0}$ and the juncture-smooth constraint $\mathbf{Z}\mathbf{x}^* = \mathbf{0}$:

$$\mathbf{x}^* = \left[\mathbf{I} - \mathbf{H}^{-1}\mathbf{Z}^T(\mathbf{Z}\mathbf{H}^{-1}\mathbf{Z}^T)^{-1}\mathbf{Z}\right]\mathbf{H}^{-1}\mathbf{R}^T\mathbf{a} \quad (20)$$

where $\mathbf{H}$ is a symmetric squared matrix expressed by

$$\mathbf{H} = \mathbf{R}^T - N\sum_{j=0}^{3} w_j(\mathbf{D}_j)^T \quad (21)$$

4. Tonality-Enhanced OEC Generation

By an informal subjective test on the synthesized speeches by the dynamics-enhanced generation algorithm in Section 3, we found some tones of syllables would sound like different tones. These glitches may be resulted from disadvantages of the SR-HPM in which logF0 contour of each syllable is assumed to be additively combined by several affecting patterns. The combined logF0s do not ensure to preserve the tone of each syllable since tonality of syllable is affected not only by

absolute value of logF0 but also by the relative logF0 value w.r.t. the ones of adjacent syllables, i.e. dynamic logF0 values. In Xu's papers [8], several perceptual tests were conducted to show that the relative height and shape of pitch contour could affect perception of tone. We believe that incorporating the dynamic logF0 features in the logF0 generation model can improve and preserve tonality of the synthesized logF0. The idea is realized by introducing two dynamic logF0 features into the original logF0 log-likelihood function. The two dynamic logF0 features are the proceeding logF0 difference, $\mathbf{x}'_n$, and the post logF0 difference, $\mathbf{x}''_n$, defined respectively by

$$\mathbf{x}'_n = \mathbf{x}_n - \mathbf{x}_{n-1}, 2 \le n \le N \text{ and } \mathbf{x}''_n = \mathbf{x}_n - \mathbf{x}_{n+1}, 1 \le n \le N-1 \quad (22)$$

A new log-likelihood function is defined by

$$\begin{aligned} \hat{L}(\mathbf{x}) = g \cdot L(\mathbf{x}) &- \tfrac{g'}{2} \sum_{n=2}^{N} \left(\mathbf{x}'_n - \mathbf{a}'_n\right)^T \mathbf{V}'^{-1}_n \left(\mathbf{x}'_n - \mathbf{a}'_n\right) \\ &- \tfrac{g''}{2} \sum_{n=1}^{N-1} \left(\mathbf{x}''_n - \mathbf{a}''_n\right)^T \mathbf{V}''^{-1}_n \left(\mathbf{x}''_n - \mathbf{a}''_n\right) \end{aligned} \quad (23)$$

where $\mathbf{a}'_n$ and $\mathbf{a}''_n$ are the two new-added 4-by-1 mean vectors of the Gaussian distributions for the logF0 differences between current syllable and proceeding syllable, and the one between current syllable and following syllable; $\mathbf{V}'_n$ and $\mathbf{V}''_n$ are the two covariances of the associated Gaussian distributions; g, g', and g'' are respectively weights for the total log-likelihoods of $\mathbf{x}_n$, $\mathbf{x}'_n$, and $\mathbf{x}''_n$. The parameters $\mathbf{a}'_n$, $\mathbf{a}''_n$, $\mathbf{V}'_n$ and $\mathbf{V}''_n$, can be obtained by a training of decision tree with a question set formed by features of contextual tones, contextual break types, and some linguistic features. Since $\mathbf{x}'_n$ and $\mathbf{x}''_n$ are linearly combined by $\mathbf{x}_n$, we can rewrite the new log-likelihood function as a matrix form by

$$\hat{L}(\mathbf{x}) = -\tfrac{1}{2}\left(\mathbf{B}\mathbf{x} - \boldsymbol{\mu}\right)^T \hat{\mathbf{R}} \left(\mathbf{B}\mathbf{x} - \boldsymbol{\mu}\right) \quad (24)$$

where $\boldsymbol{\mu}$ is a mean vector comprising $\mathbf{a}_n$, $\mathbf{a}'_n$, and $\mathbf{a}''_n$; $\hat{\mathbf{R}}$ is a square matrix composed of inverse covariance matrixes scaled by the inverses of the weights g, g', and g''; $\mathbf{B}$ is a transformation matrix for deriving a vector composed of $\mathbf{x}_n$, $\mathbf{x}'_n$, and $\mathbf{x}''_n$. The objective function for generating the tonality-enhanced logF0 contour is defined by

$$\hat{U}(\mathbf{x}, \boldsymbol{\lambda}) = \hat{L}(\mathbf{x}) + \boldsymbol{\lambda}^T \mathbf{Z}\mathbf{x} + \tfrac{N}{2} \sum_{j=0}^{3} w_j \mathbf{x}^T \mathbf{D}_j \mathbf{x} \quad (25)$$

The optimal OECs, $\hat{\mathbf{x}}$, can be derived by differentiation of $\hat{U}(\mathbf{x}, \boldsymbol{\lambda})$ w.r.t. $\mathbf{x}$:

$$\hat{\mathbf{x}} = \left[\mathbf{I} - \mathbf{G}^{-1}\mathbf{Z}^T \left(\mathbf{Z}\mathbf{G}^{-1}\mathbf{Z}^T\right)^{-1} \mathbf{Z}\right] \mathbf{G}^{-1} \mathbf{B}^T \hat{\mathbf{R}} \boldsymbol{\mu} \quad (26)$$

where

$$\mathbf{G} = \mathbf{B}^T \hat{\mathbf{R}} \mathbf{B} - \tfrac{N}{2} \sum_{j=0}^{3} w_j (\mathbf{D}_j)^T \quad (27)$$

It is noted that the modeling of the dynamic features may help keep the appropriate relative heights and shapes of the generated logF0 as the variance term $v(j)$ is incorporated.

5. Dynamics-Controlled OEC Generation

By an analysis on the prosody labeled speech corpus [3], we found that dynamic ranges of OECs could be affected by SR, length of utterance, distributions of tones and break types. Therefore, it is desired to precisely control dynamic ranges of generated OECs. To this end, for each utterance, the following equation must be satisfied for finding the weights w_j's for the desired variance of each dimension of OEC, i.e.

$$h_j(\hat{\mathbf{w}}) = \hat{\mathbf{x}}^T \mathbf{D}_j \hat{\mathbf{x}} - \hat{v}_j = 0, \text{ for } j = 0 \sim 3 \quad (28)$$

where $\hat{\mathbf{w}} = \{\hat{w}_0, \hat{w}_1, \hat{w}_2, \hat{w}_3\}$ is the optimal weight set; $\hat{\mathbf{x}}^T \mathbf{D}_j \hat{\mathbf{x}}$ and $\hat{v}_j$ represents represent respectively the variance of j-th dimension of the generated OECs and the desired variance of j-th dimension. As shown in Eqs. (14) and (25), the optimal OEC, $\hat{\mathbf{x}}$, is a function of the weight $\hat{\mathbf{w}}$. To find the optimal weights that satisfy Eq. (28), an iterative algorithm based on the Newton's method is applied to find the roots of Eq. (28), i.e. $h_j(\hat{w}_0, \hat{w}_1, \hat{w}_2, \hat{w}_3) = 0$, for j=0~3. The basic form for updating weights can be expressed by

$$\hat{w}_i^{(k+1)} = \hat{w}_i^{(k)} - h_j(\hat{\mathbf{w}}^{(k)}) \Big/ \nabla_{w_i} h_j(\hat{\mathbf{w}}^{(k)}), \text{ for all } i \text{ and } j \quad (29)$$

where k represents iteration index; $\nabla_{w_i} h_j(\hat{\mathbf{w}}^{(k)}) = \partial\left(\hat{\mathbf{x}}^T \mathbf{D}_j \hat{\mathbf{x}} - \hat{v}_j\right) / \partial w_i \Big|_{\mathbf{w} = \hat{\mathbf{w}}^{(k)}}$ is the gradient of $h_j(\mathbf{w})$ w.r.t. w_i at $\mathbf{w} = \hat{\mathbf{w}}^{(k)}$ expressed by

$$\nabla_{w_i} h_j(\hat{\mathbf{w}}^{(k)}) = 2N \cdot \left(\hat{\mathbf{x}}^{(k)}\right)^T \mathbf{D}_j \mathbf{U}^{(k)} \left(\mathbf{G}^{(k)}\right)^{-1} (\mathbf{D}_i)^T \hat{\mathbf{x}}^{(k)} \quad (30)$$

where

$$\begin{aligned} \mathbf{G}^{(k)} &= \mathbf{B}^T \hat{\mathbf{R}} \mathbf{B} - \tfrac{N}{2} \sum_{j=0}^{3} w_j^{(k)} (\mathbf{D}_j)^T \\ \mathbf{U}^{(k)} &= \mathbf{I} - \left(\mathbf{G}^{(k)}\right)^{-1} \mathbf{Z}^T \left(\mathbf{Z} \left(\mathbf{G}^{(k)}\right)^{-1} \mathbf{Z}^T\right)^{-1} \mathbf{Z} \\ \hat{\mathbf{x}}^{(k)} &= \mathbf{U}^{(k)} \left(\mathbf{G}^{(k)}\right)^{-1} \mathbf{B}^T \hat{\mathbf{R}} \boldsymbol{\mu} \end{aligned} \quad (31)$$

Since the four equations (for j=0~3) in Eq. (31) are the functions of w_j's, a special designed iterative algorithm is stated in the following to solve the roots for Eq. (28):

Step 1: Set initial weights to be $\hat{\mathbf{w}}^{(0)} = \{\hat{w}_0^{(0)}, \hat{w}_1^{(0)}, \hat{w}_2^{(0)}, \hat{w}_3^{(0)}\}$.

Step 2: Update weights by the following loops

for j=0 to 3

while k < a predefined maximum iteration number

- Update w_j by the Newton's method:
  $\hat{w}_j^{(k+1)} = \hat{w}_j^{(k)} - h_j(\hat{\mathbf{w}}^{(k)}) \Big/ \nabla_{w_j} h_j(\hat{\mathbf{w}}^{(k)})$
- If $h_j(\mathbf{w}) < \varepsilon_j$, exit the while loop, or k=k+1
  (ε_j : a predefined threshold for each j-th dimension)

end_while

end_for

Step 3: If $h_j(\mathbf{w}) < \varepsilon_j$ for all j, go to Step 4, or go to Step 2

Step 4: Exit and return the optimal generated OECs, $\hat{\mathbf{x}}$.

6. Experimental Results

To examine the effectiveness of the proposed methods, the parameters of prosodic acoustic features are generated by the Mandarin speaking-rate dependent hierarchical prosodic model (SR-HPM) [6]. The SR-HPM is trained by a female Mandarin speech database with four parallel speech corpora of slow, medium, normal and fast SRs. There are in total 1,478 utterances with 183,795 syllables, in which 176 utterances are taken as test set. To generate prosodic-acoustic features by the SR-HPM, the prosodic break type, B_n, is first predicted given with linguistic feature ($\mathbf{L}_n$) and a specified SR (s) by

$$\bar{B}_n = \arg\max_{B_n} P(B_n \mid \mathbf{L}_n, s) \quad (32)$$

The mean vector, $\mathbf{a}_n$, is obtained by an MMSE predictor [2]:

$$\mathbf{a}_n = E[\mathbf{sp}_n \mid \bar{\mathbf{B}}, \mathbf{L}] = \sum_{p_n} \boldsymbol{\beta}(t_{n-1}^{n+1}, p_n, \bar{B}_{n-1}^{n}) P(p_n \mid \bar{\mathbf{B}}, \mathbf{L}) \quad (33)$$

where $P(p_n \mid \mathbf{B}, \mathbf{L})$ is the *a posteriori* probability for the pitch prosodic state which is calculated by a forward/backward recursion [2] given with a synthesized probabilistic model:

$$P(p_n \mid p_{n-1}, B_n, \mathbf{L}) \approx P(p_n \mid p_{n-1}, B_n) + P(p_n \mid \mathbf{L}) \quad (34)$$

where $P(p_n \mid p_{n-1}, B_n)$ and $P(p_n \mid \mathbf{L})$ are respectively the prosodic state model and the prosodic state-syntax model [6]. The covariance matrix, $\mathbf{V}_n$, is obtained by

$$\mathbf{V}_n = \sum_{p_n} \left(\mathbf{V} + \boldsymbol{\beta}_{p_n} (\boldsymbol{\beta}_{p_n})^t\right) P(p_n \mid \bar{\mathbf{B}}, \mathbf{L}) - \mathbf{a}_n (\mathbf{a}_n)^t \quad (35)$$

The proposed OEC generation algorithms are examined on the test set subjectively. The experiment setups are designed as follow: ***(1) BSL***: OECs are generated by the baseline ML criterion stated in Eq. (5), ***(2) SMT***: incorporating the juncture-smooth criterion expressed in Eq. (12), ***(3) TNL***: incorporating the likelihood terms in Section 4 that model the dynamic features to enhance the tonality, ***(4) DYC***: incorporating the dynamics-controlled algorithm stated in Section 5, and ***(6) GNM***: using the Gaussian normalization method to scale the OECs to the desired variance of each dimension. It is noted that, in the generation method by ***DYC*** and ***GNM***, the desired variance of *j*-th dimension for an utterance, i.e. $\hat{v}_j$ for $j = 0 \sim 3$ are predicted by four independent CART decision trees (each one for each *j*-th dimension) trained by samples of utterance-wise variances given with question sets formed by the features about SR, length of utterance, distribution of tones/break types in an utterance. The split criterion for the CART decision trees is likelihood gain. Besides, two independent decision trees are trained by the training set for the parameters of pre- and post- logF0 differences for the ***TNL*** method. The weight for ***TNL***, i.e. g, g', and g'' are empirically tuned by an informal listening test on the synthesized utterances with some texts of the training set. The SR is set to be a normal SR of 0.20 sec/syllable for the SR-HPM to generate prosody.

Mean opinion score (MOS) test and preference test were performed simultaneously by 11 subjects given with 10 synthesized long utterances with lengths from 68 to 128 syllables (101 syllables in average) for each prosody generation method. The texts of the 10 utterances were chosen from the texts in the testing set. There are four experimental settings for comparison: (1) Set1: +***SMT*** vs ***BSL***, (2) Set2: +***SMT***+***TNL*** vs. ***BSL***, (3) Set3: +***SMT***+***TNL***+***DYC*** vs. ***BSL***, and (4) Set4: +***SMT***+***TNL***+***DYC*** vs. ***BSL***+***GNM***. Table 1 shows statistics of variances of OECs for the authentic utterance, ***BSL***, and +***SMT***+***TNL***+***DYC*** and ***BSL***+***GNM***. It is found that the logF0 dynamics of ***BSL*** is much smaller than the one of the authentic speech while the dynamics of +***SMT***+***TNL***+***DYC*** or ***BSL***+***GNM*** which is predicted by the decision trees can be closer to the true variances.

Before listening to the synthesized utterances by ***BSL*** and the proposed methods, subjects were asked to listen to the authentic utterances of normal SR in the test speech corpus corresponding to the synthesized speeches for reference. The order of the synthesized utterances in the preference test is randomly set. Table 2 displays the results of the preferences and the MOSs for the subjective tests. The result in Set1 indicates the proposed juncture-smooth generation method could reduce logF0 discontinuity between voiced syllable junctures so as to make synthesized speech more natural and fluent. As shown in Set2, the proposed tonality-enhanced method (***TNL***) with the juncture-smooth criterion still performed better than ***BSL*** but the improvement was relatively smaller than Set1. This degradation may be due to over-enhancing the OEC differences between adjacent syllables. The significant improvement against ***BSL*** was shown in the result of Set3, proving that increasing dynamics of the generated logF0 and keeping juncture smoothness really improves the naturalness of the synthesized speech. In the Set4, the proposed method (+***SMT***+***TNL***+***DYC***) reached the highest MOS among the all subjective the testing settings and still performed slightly better than the baseline system with enhanced dynamics by ***GNM*** method. By an analysis on the generated logF0 contours, we found that though logF0 contours made by ***BSL***+***GNM*** have the same dynamic ranges as the ones made by the proposed +***SMT***+***TNL***+***DYC***, the logF0 contours by ***BSL***+***GNM*** are sometimes not smooth-connected on the voiced syllable junctures. Generally, enhancing dynamics of logF0 contours could significantly improve the naturalness the synthesized speech while ensuring logF0 smoothness on voiced syllable junctures by the juncture-smooth criterion could further improve the naturalness. Figure 2 displays an example illustrating logF0 contours by ***BSL***(upper), ***BSL***+***GNM***(middle), and the proposed +***SMT***+***TNL***+***DYC***(lower). It is obvious to see that the proposed +***SMT***+***TNL***+***DYC*** method is advantageous in keeping logF0 smoothness and enhancing dynamics of logF0 at the same time – the main merit of the proposed method.

Table 1: *Averages of variances ($\times 10^4$ logHz^2) of OECs for authentic utterances, the synthesized utterances by* ***BSL****, and by +****SMT***+***TNL***+***DYC*** *or* ***BSL***+***GNM*** *(DYN) of the testing set.*

OEC dimension	0	1	2	3
authentic speech	567.05	95.80	18.83	4.85
BSL	286.86	49.15	5.79	0.97
DYN	511.46	80.30	15.20	3.87

Table 2: *Preferences (%) and MOSs (numbers in brackets ± standard deviation) for the two subjective tests.*

Set 1		***Set 3***	
+***SMT***	38% (3.47 ± .49)	+***SMT***+***TNL***+***DYC***	41% (3.45 ± .62)
BSL	25% (3.34 ± .48)	***BSL***	25% (3.22 ± .56)
No prefer.	37%	No prefer.	34%
Set 2		***Set 4***	
+***SMT***+***TNL***	32% (3.40 ± .51)	+***SMT***+***TNL***+***DYC***	37% (3.50 ± .55)
BSL	26% (3.30 ± .52)	***BSL***+***GNM***	27% (3.43 ± .55)
No prefer.	41%	No prefer.	36%

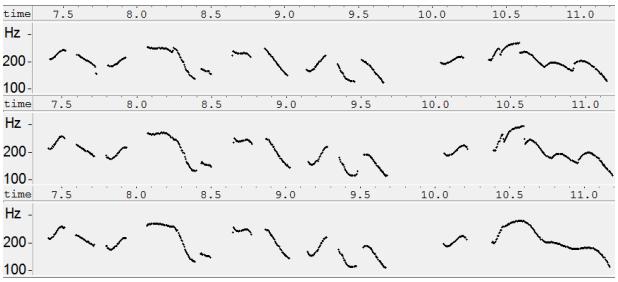

Figure 2: *Exemplar logF0 contours by* ***BSL*** *(upper),* ***BSL***+***GNM*** *(middle), and the proposed +****SMT***+***TNL***+***DYC*** *(lower).*

7. Conclusions and Future Works

This paper proposed a new logF0 contour generation method that combines ideas of the juncture smoothing, the tonality enhancing, and the dynamics enhancing. The proposed dynamics-controlled OEC generation algorithm can precisely generate synthesized speech with designated variances of OECs and smooth logF0 contours across voiced syllable junctures. The effectiveness of the proposed algorithm was proved by several subjective tests. In the future, it would be interesting to apply the proposed logF0 generation algorithm for emotional speech synthesis since dynamics of logF0 is highly correlated with types of emotions.

8. Acknowledgements

This work was supported the MOST of Taiwan under Contract No. NSC-102-2221-E-305-005-MY3.

9. References

[1] S. H. Chen and Y. R. Wang, "Vector quantization of pitch information in Mandarin speech," IEEE Trans. Commun., vol. 38, no. 9, pp. 1317-1320

[2] Chen-Yu Chiang, Sin-Horng Chen and Yih-Ru Wang, "Advanced Unsupervised Joint Prosody Labeling and Modeling for Mandarin Speech and Its Application to Prosody Generation for TTS," in Proc. Interspeech 2009, Brighton, UK, Sept. 2009, pp. 504-507.

[3] C. C. Hsia, C. H. Wu, and J. Y. Wu, "Exploiting Prosody Hierarchy and Dynamic Features for Pitch Modeling and Generation in HMM-Based Speech Synthesis," IEEE Trans. Audio, Speech, and Language Processing, vol. 18, no. 8, pp.1994-2003, August 2010.

[4] Sin-Horng Chen, Shaw-Hwa Hwang, and Yih-Ru Wang, "An RNN-Based Prosodic Information Synthesizer for Mandarin Text-to-Speech," IEEE Trans. Acoust., Speech, Signal Processing, vol. 6, pp. 226-239, May 1998

[5] C.-Y. Chiang, S.-H. Chen, H.-M. Yu, and Y.-R. Wang, "Unsupervised Joint Prosody Labeling and Modeling for Mandarin Speech," J. Acoust. Soc. Am. 125, No. 2, pp. 1164-1183 (2009).

[6] Sin-Horng Chen, Chiao-Hua Hsieh, Chen-Yu Chiang, Hsi-Chun Hsiao, Yih-Ru Wang, Yuan-Fu Liao, and Hsiu-Min Yu, "Modeling of Speaking Rate Influences on Mandarin Speech Prosody and Its Application to Speaking Rate-controlled TTS," in *Audio, Speech, and Language Processing, IEEE/ACM Transactions on* , vol.22, no.7, pp.1158-1171, July 2014

[7] W. C. Kuo, X. R. Zhong, Y. R. Wang and S. H. Chen, "A High-Performance Min-Nan/Taiwanese TTS System", in *Proc. ICASSP'03*, Hong Kong, April 2003.

[8] Y. Xu, "Contextual tonal variations in Mandarin," Journal of Phonetics **25**, 61-83. (1997)