IDIAP RESEARCH REPORT

CONVOLUTIONAL PITCH TARGET APPROXIMATION MODEL FOR SPEECH SYNTHESIS

Xingyu Na[a] Philip N. Garner

Idiap-RR-05-2013

MARCH 2013

[a]Idiap Research Institute

Convolutional Pitch Target Approximation Model for Speech Synthesis

Xingyu Na, and Philip N. Garner

Abstract

In this paper, we investigate pitch contour modelling in speech synthesis based on segmental units. A convolutional pitch target approximation model is proposed. This model allows jointly stochastic modelling of framewise pitch and pitch contour of longer units, of which the intuitive relations are revealed by a convolutional target approximation filter. The pitch contour is stylized by a linear representation called pitch target. In synthesis stage, the likelihood of the framewise model and the pitch target model are jointly maximized using a Toeplitz matrix representing the discrete convolutional filter.

Index Terms

Pitch modelling, speech synthesis, pitch target approximation.

I. Introduction

We are interested in general in modelling the prosodic contour of synthetic speech for statistical text to speech synthesis (TTS). In particular, we seek a model that allows the prosody to be modified in order to convey a particular intention. For instance, the dialogue component of a conversational system may wish to choose intonation indicative of a question rather than a statement, or to emphasise a given phrase. This in turn implies control over pitch, intensity and duration of speech segments. In this letter we focus on pitch. Two issues are apparent:

1) Pitch contours associated with emotion are supra-segmental; they span phonemes.
2) To respond to high level semantic cues, the model should have physical meaning.

From a superpositional point of view, additive pitch models based on a hierarchical prosody structure decompose pitch contour into various levels and model them separately [1]. Other than using additive components, contour parameterization was an alternative approach, such as discrete cosine transform (DCT) [2]. By changing the analytical granularity, it can also be used on a hierarchical structure [3]. However, neither additive model nor contour parameterization reveals the intuitive relations between the framewise pitch and the segmental unit contour.

The Parallel ENcoding and Target Approximation (PENTA) model [4] was proposed to study the linguistic phenomena observed in an acoutic sense. Assuming that the pitch targets are sequentially assigned at syllable boundaries before being realized, they are defined as the straight line,

$$u(t) = mt + b. \tag{1}$$

The parameters, m and b, have explicit physical meanings, i.e. target slope and target height.

Prior to any implementation, we note that evaluation of the proposed model in a supra-segmental sense is difficult; it involves rather subjective tests. In order to validate the *model*, we make use of Mandarin. Being tonal, Mandarin has two advantages for this:

1) Pitch carries syntactic as well as semantic meaning; the former can be evaluated more objectively.
2) The tonality is segmental; this eases implementation in a HMM based TTS system.

In the following sections, we describe a stochastic pitch model based on a contour approximation simulated by a discrete convolutional filter. We then validate the model qualitatively by using it to generate (tonal) pitch contours for Mandarin speech. The method is found to generate pitch with the right syntactic meaning, albeit with a slight accent. We conclude that the model merits further investigation.

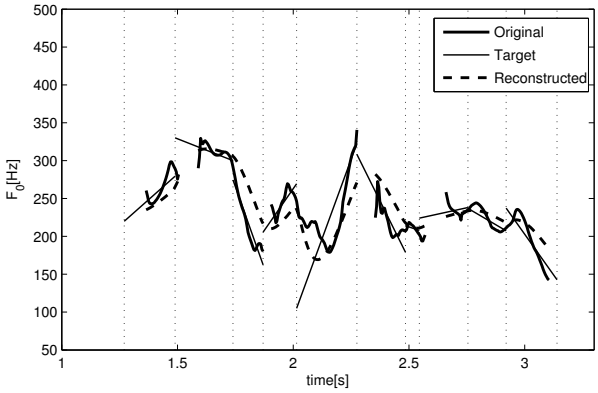

Fig. 1. Illustrative test of the discrete version of the target approximation filter on a Mandarin utterance. Dot lines represent the syllable boundaries. Target parameters are extracted by minimizing the reconstructive error.

II. Convolutional target approximation filter

The generation of pitch contour can be viewed as successively approaching a sequence of assigned pitch targets. Linear representation of pitch targets has been proven to be adequate for realizing natural prosody [5]. To reconstruct the pitch contour from that, we parameterize the pitch realization process as

$$f_u(t) = u(t) * h(t), \tag{2}$$

in which the convolutional form allows fast digital implementation.

Zhang et al. [6] designed an FIR filter to reconstruct pitch contour from straight target, which is similar to a Gaussian filter.

$$h_G(t) = \frac{\sqrt{\pi}}{\alpha} \cdot \exp\left(-\frac{\pi^2}{\alpha^2}t^2\right) \tag{3}$$

The Fujisaki model, also known as the command-response (CR) model or the generation process model in the field of prosody research, assumes that the pitch contour is the superposition of phrase and accent components [7], which are modelled as the output of second-order critically-damped filters with exponential form [8].

$$h_E(t) = \alpha^2 t \exp(-\alpha t) \tag{4}$$

In a preliminary experiment, we tested two filters with artificial pitch targets. The output of the exponential filter asymptotically approaches the target till the end of the target, while that of the Gaussian filter approaches in the middle of it. Several studies [9, 10] suggest that the target approximation process of pitch should be sequential and end at the boundaries. Therefore, the exponential filter is more suitable for the target approximation model.

III. Discrete convolutional filter

To introduce the target approximation into stochastic models, we need to derive the discrete version of the convolutional filter. The Laplace transform of exponential filter is

$$\begin{aligned} \mathcal{F}_E(s) &= \mathcal{L}(h_E(t))(s) \\ &= \int_0^{+\infty} \alpha^2 t \exp(-\alpha t) \cdot \exp(-st)\, dt \\ &= \frac{\alpha^2}{(s+\alpha)^2} \end{aligned} \tag{5}$$

First, transform the s-plane filter response function to the z-plane using first-order Padé approximant

$$s \approx \frac{1 - z^{-1}}{t_0} \tag{6}$$

where t_0 is the sampling period of the discrete-time representation. The inverse system in the z-domain is then

$$\mathcal{H}_E^{-1}(z) = \psi^2 - 2\psi(\psi - 1)z^{-1} + \psi^2 z^{-2} \tag{7}$$

where $\psi = 1 + 1/(\alpha t_0)$. Hence, the backward difference form is

$$u[t] = a_2 f_u[t-2] + a_1 f_u[t-1] + a_0 f_u[t] \tag{8}$$

where

$$\begin{aligned} a_0 &= \psi^2, \\ a_1 &= -2\psi(\psi - 1), \\ a_2 &= \psi^2. \end{aligned} \tag{9}$$

Hereafter, we use t to denote the discrete time. Given Eq.8, the pitch $\mathbf{f}_u \triangleq (f_u[t], \cdots, f_u[T])^{\mathsf{T}}$ can be written in terms of $\mathbf{u}$, such that

$$\mathbf{u} = \mathbf{A}\mathbf{f}_u, \tag{10}$$

and to introduce the target as a constraint in the stochastic model, we rewrite the equation as

$$\mathbf{f}_u = \mathbf{A}^{-1}\mathbf{u}, \tag{11}$$

where

$$\mathbf{A} \triangleq \begin{pmatrix} a_0 & & & & O \\ a_1 & a_0 & & & \\ a_2 & a_1 & a_0 & & \\ & \ddots & \ddots & \ddots & \\ O & & a_2 & a_1 & a_0 \end{pmatrix} \tag{12}$$

is a Toeplitz matrix representing the inverse of the discrete target approximation filter. To illustrate the approximation performance of the discrete convolution, we extract the line parameters m and b using a steepest descent gradient algorithm to minimize the root mean square error (RMSE), defined as

$$\text{Err} = \left(\frac{1}{T} \sum_{t=1}^{T} (f[t] - u[t])^2 \right)^{1/2}. \tag{13}$$

The pitch contour was estimated by STRAIGHT-TEMPO [11]. One sample utterred by a female speaker is shown in Fig .1. As suggested by Kameoka et al. [8], the approximation strength parameter α was set to be a constant for a given speaker. In this case, based on the grid search in term of the average RMSE between the original and the reconstructed pitch of 100 testing utterances, we set $\alpha = 30.8$. The pitch targets generally identify the shape of the pitch contours.

IV. Stochastic modelling

A meaningful stochastic model of pitch target assignment should satisfy the following requirements:

1) Targets are represented by a piecewise linear function.
2) Target slope and height are constants within a syllable.
3) The offset of one target is followed by the onset of the next target.

Assuming the pitch targets are the emissions of a set of HMMs which satisfy the above constraints, the topology is as shown in Fig.2. The segmental units are modelled by HMMs with one emitting state using single Gaussian distribution. g_k represents the stochastic distribution of the target model of the k^{th} segmental unit in the utterance, while $\mathbf{f}_k$ denotes the corresponding emitted pitch target.

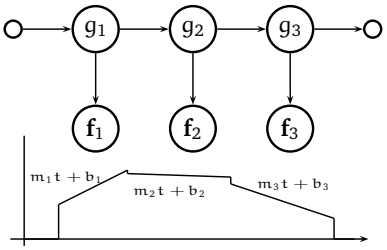

Fig. 2. Pitch target modelling using HMM considering constraints of target assignment.

A. Likelihood function

In this model, the output target of unit k at time t is observed given the hidden variable with additive white Gaussian noise.

$$\mathbf{o}_k[t] \triangleq u_k[t] = \mu_{g_k}[t] + \epsilon_{g_k}[t] \tag{14}$$

Let $\epsilon_{g_k}[t] \sim \mathcal{N}(0, \sigma^2_{g_k})$ then

$$\mathbf{o}_k[t] \sim \mathcal{N}(\mu_{g_k}[t], \sigma^2_{g_k}) \tag{15}$$

where $\mu_{g_k}[t] = m_{g_k} t + b_{g_k}$ and $\sigma^2_{g_k}$ are respectively the mean and variance of the target model g_k.

The free parameters in this model consist of the slope of the pitch target m_{g_k}, the height of the pitch target b_{g_k}, and the variance of the output distribution, $\sigma^2_{g_k}$. Here we use $\Theta_{\mathbf{g}}$ to denote the free parameter set of pitch target model as

$$\Theta_{\mathbf{g}} \triangleq \{\mathbf{M_g}, \mathbf{B_g}, \mathbf{\Sigma_g}\} \tag{16}$$

where $\mathbf{M_g} = (m_{g_1}, \ldots, m_{g_K})^\mathsf{T}$, $\mathbf{B} = (b_{g_1}, \ldots, b_{g_K})^\mathsf{T}$, and $\mathbf{\Sigma_g} = (\sigma^2_{g_1}\mathbf{I}, \ldots, \sigma^2_{g_K}\mathbf{I})$. $\mathbf{g} = (g_1, \ldots, g_K)$ represents the model sequence of the segmental units. From Eq. 15 and Eq. 16, we define

$$\mathbf{u} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{g}}, \mathbf{\Sigma_g}) \tag{17}$$

where $\mathbf{u} \triangleq (u[1], \ldots, u[T])^\mathsf{T}$ and $\boldsymbol{\mu}_{\mathbf{g}} \triangleq (\mu[1], \ldots, \mu[T])^\mathsf{T}$. Overall, the probability density function of target model parameters can be written as

$$\begin{aligned} p(\mathbf{u} \mid \Theta_{\mathbf{g}}) &= \mathcal{N}(\mathbf{u}; \boldsymbol{\mu}_{\mathbf{g}}, \mathbf{\Sigma_g}) \\ &= \frac{1}{(2\pi)^{T/2}|\mathbf{\Sigma_g}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{u} - \boldsymbol{\mu}_{\mathbf{g}})^\mathsf{T}\mathbf{\Sigma}_{\mathbf{g}}^{-1}(\mathbf{u} - \boldsymbol{\mu}_{\mathbf{g}})\right) \end{aligned} \tag{18}$$

where $\mathbf{u} = \mathbf{A}\mathbf{f}$ reveals the underlying relationship between variable $\mathbf{u}$ and $\mathbf{f}$ as defined by Eq. 10.

B. Maximum likelihood pitch generation

In HMM-based speech synthesis, pitch is modelled using framewise multi-space distribution HMMs. To capture the temporal dependency, static features are generated considering the constraint imposed by dynamic window coefficients [12]. The probability density function of pitch model is [13]

$$\begin{aligned} p(\mathbf{W}\mathbf{f} \mid \Theta_{\mathbf{q}}) = \\ \frac{1}{(2\pi)^{3T/2}|\mathbf{\Sigma_q}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{W}\mathbf{f} - \boldsymbol{\mu}_{\mathbf{q}})^\mathsf{T}\mathbf{\Sigma}_{\mathbf{q}}^{-1}(\mathbf{W}\mathbf{f} - \boldsymbol{\mu}_{\mathbf{q}})\right) \end{aligned} \tag{19}$$

where $\Theta_{\mathbf{q}}$ represents the framewise model parameter set of state sequence $\mathbf{q}$, and $\mathbf{W}$ is the window coefficient matrix given by

$$\begin{aligned} \mathbf{W} &= [\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_T]^\mathsf{T}, \\ \mathbf{W}_t &= \left[\mathbf{w}_t^{(0)}, \mathbf{w}_t^{(1)}, \mathbf{w}_t^{(2)}\right]. \end{aligned} \tag{20}$$

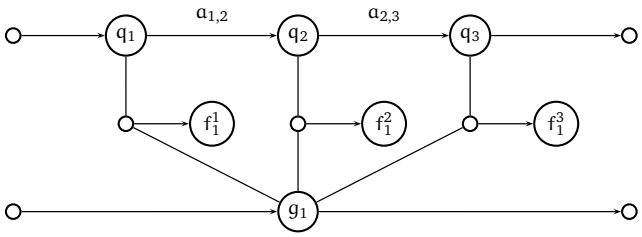

Fig. 3. Joint pitch model of a segmental unit, assuming the framewise model has three emitting states.

$\mathbf{w}_t^{(d)}$ is the window coefficient vector for calculating the d^{th} order dynamic feature of frame t, which only has non-zero values on the t^{th} and adjacent elements, depending on the window length.

To combine the framewise model with the segmental model, the state sequence of the framewise model and that of the pitch target model are aligned in the synthesis stage. $\mathbf{f}$ is determined by maximizing the joint likelihood considering dynamic features and target model as

$$\hat{\mathbf{f}} = \arg\max_{\mathbf{f}} p(\mathbf{Wf} \mid \Theta_{\mathbf{q}}) p(\mathbf{u} \mid \Theta_{\mathbf{g}}) \tag{21}$$

used for imposing $p(\mathbf{u} \mid \Theta_{\mathbf{g}})$ as a constraint term in parameter optimization. The topology is shown in Fig.3. A new set of joint distributions are formed by imposing the aligned target model. This generative model of pitch contour describes not only the windowed pitch distribution at each frame, but also the target distribution of the current segmental unit.

Ignoring the parts independent of $\mathbf{f}$, the objective function to be maximized is defined as the log likelihood function

$$\begin{aligned} L_{\Theta_{\mathbf{q}},\Theta_{\mathbf{g}}} &= \log p(\mathbf{Wf} \mid \Theta_{\mathbf{q}}) + \log p(\mathbf{u} \mid \Theta_{\mathbf{g}}) \\ &\propto -\frac{1}{2}\left(\mathbf{f}^\top \mathbf{W}^\top \boldsymbol{\Sigma}_{\mathbf{q}}^{-1} \mathbf{Wf} - 2\boldsymbol{\mu}_{\mathbf{q}}^\top \boldsymbol{\Sigma}_{\mathbf{q}}^{-1} \mathbf{Wf}\right) \\ &\quad -\frac{1}{2}\left(\mathbf{f}^\top \mathbf{A}^\top \boldsymbol{\Sigma}_{\mathbf{g}}^{-1} \mathbf{Af} - 2\boldsymbol{\mu}_{\mathbf{g}}^\top \boldsymbol{\Sigma}_{\mathbf{g}}^{-1} \mathbf{Af}\right) \end{aligned} \tag{22}$$

Taking the first derivative gives

$$\begin{aligned} \frac{\partial L_{\Theta_{\mathbf{q}},\Theta_{\mathbf{g}}}}{\partial \mathbf{f}} = & \\ & -\frac{1}{2}\left(\frac{\partial}{\partial \mathbf{f}}\left(\mathbf{f}^\top \mathbf{W}^\top \boldsymbol{\Sigma}_{\mathbf{q}}^{-1} \mathbf{Wf}\right) - 2\frac{\partial}{\partial \mathbf{f}}\left(\boldsymbol{\mu}_{\mathbf{q}}^\top \boldsymbol{\Sigma}_{\mathbf{q}}^{-1} \mathbf{Wf}\right)\right) \\ & -\frac{1}{2}\left(\frac{\partial}{\partial \mathbf{f}}\left(\mathbf{f}^\top \mathbf{A}^\top \boldsymbol{\Sigma}_{\mathbf{g}}^{-1} \mathbf{Af}\right) - 2\frac{\partial}{\partial \mathbf{f}}\left(\boldsymbol{\mu}_{\mathbf{g}}^\top \boldsymbol{\Sigma}_{\mathbf{g}}^{-1} \mathbf{Af}\right)\right) \\ & = -\mathbf{Rf} + \mathbf{r} \end{aligned} \tag{23}$$

where

$$\begin{aligned} \mathbf{R} &= \mathbf{W}^\top \boldsymbol{\Sigma}_{\mathbf{q}}^{-1} \mathbf{W} + \mathbf{A}^\top \boldsymbol{\Sigma}_{\mathbf{g}}^{-1} \mathbf{A} \\ \mathbf{r} &= \mathbf{W}^\top \boldsymbol{\Sigma}_{\mathbf{q}}^{-1} \boldsymbol{\mu}_{\mathbf{q}} + \mathbf{A}^\top \boldsymbol{\Sigma}_{\mathbf{g}}^{-1} \boldsymbol{\mu}_{\mathbf{g}} \end{aligned} \tag{24}$$

Letting $\partial \mathbf{L}_{\theta_q,\theta_u} / \partial \mathbf{f} = 0$, we find

$$\mathbf{R}\hat{\mathbf{f}} = \mathbf{r}. \tag{25}$$

By solving Eq. 25, the maximum likelihood pitch contour is generated. This equation can be solved by decomposing $\mathbf{R}$ into triangular matrices and then using the forward-backward Gaussian substitution method [13].

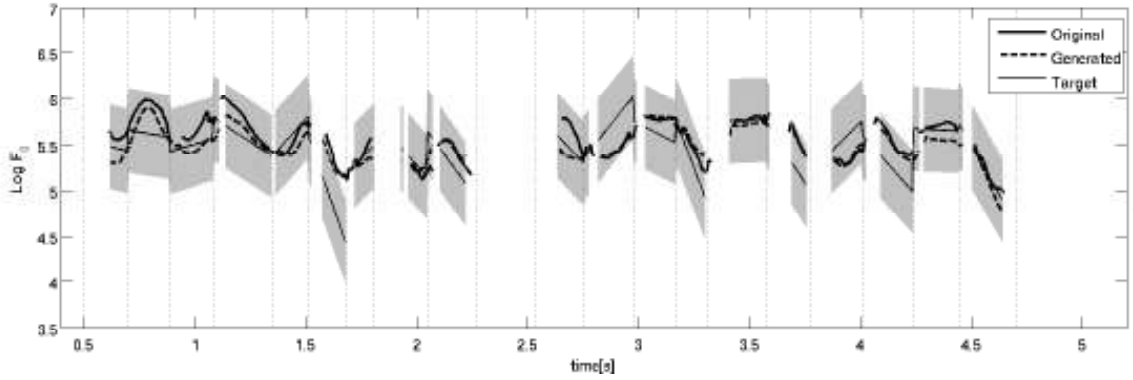

Fig. 4. A sample of pitch generation. The thick solid line is the original pitch of the testing utterance. The dash line is the pitch contour generated jointly by the framewise pitch model and the pitch target model. The thin solid line and shadow show the distributions of the targets predicted by the decision tree. The dot lines show the syllable boundaries.

TABLE I
WAVEFORMS DEMONSTRATING SYNTHESISER PERFORMANCE

	Female speaker
Original	`BIT_ASCCD_f001_012_06_05.ori.wav`
Synthsized	`BIT_ASCCD_f001_012_06_05.tgt.wav`

C. Illustrative Sample

To demonstrate the convolutional pitch model, both a framewise acoustic model and a pitch target model are trained using a subset of the Mandarin corpus ASCCD [14]. A decision tree is used to cluster the target model using the same lexical features as framewise models. Fig. 4 shows a testing utterance (`BIT_ASCCD_f001_012_06_05`). The generated pitch approaches the natural contour while following the targets predicted by decision tree. Most of the pitch dynamic ranges are covered by the target distributions.

Besides tone, syllabic identity is also important. The target model allows unified acoustic modelling by joining these two factors. The performance is evidence from the waveforms included with in this submission, generated using the testing data in ASCCD, summarised in Table I. Judged by 5 native Mandarin speakers, although with a sligth accent, the speeches sound syntactically right. Hence, we believe it will work suprasegmentally.

V. CONCLUSION

In this letter, it has been shown that the pitch target model and associated discrete convolutional filter allow a unified framework to generate pitch contours based on knowledge of segmental units. Both discrete approximation filter and the resulting pitch model has been validated intuitively by real testing data. The tests on Mandarin speech synthesis provided right syntactic meaning. Therefore, this model can be used as segmental unit models in pitch modification scenarios, such as emotion adaptation.

The model as described, can be developed into multiple levels of segmental units as the DCT model by Qian et al. [3]. Incorporated with suitable syntactic and semantic cues, this model allows explicit prosody control. However, current implementation uses discontinuous pitch, which is not suitable for prosody modelling. Simple pitch interpolations generate quite different targets when there is a long unvoiced segment within the unit. To address this issue, incorporating the continuous pitch estimator by Garner et al. [15] is worth trying.

ACKNOWLEDGMENT

The first author is supported by the China Scholarship Council.

REFERENCES

[1] 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 Transactions on Audio, Speech and Language Processing*, vol. 18, no. 8, pp. 1994–2003, November 2010.

[2] J. Teutenberg, C. Watson, and P. Riddle, "Modelling and synthesising F0 contours with the discrete cosine transform," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, 2008, pp. 3973–3976.

[3] Y. Qian, Z. Wu, B. Gao, and F. K. Soong, "Improved prosody generation by maximizing joint probability of state and longer units," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 6, pp. 1702–1710, August 2011.

[4] S. Prom-on, Y. Xu, and B. Thipakorn, "Modeling tone and intonation in Mandarin and English as a process of target approximation," *Journal of the Acoustical Society of America*, vol. 125, no. 1, pp. 405–424, January 2009.

[5] S. Ravuri and D. P. W. Ellis, "Stylization of pitch with syllable-based linear segments," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, 2008, pp. 3985–3988.

[6] Z. Zhang, X. Wang, Y. Yu, and X. Wu, "Hierachical pitch target model for Mandarin speech," in *Proceedings of International Symposium on Chinese Spoken Language Processing*, Taiwan, 2010, pp. 378–382.

[7] H. Fujisaki and S. Nagashima, "A model for the synthesis of pitch contours of connected speech," Engineering Research Institute, University of Tokyo, Tech. Rep., 1969.

[8] H. Kameoka, J. L. Roux, and Y. Ohishi, "A statistical model of speech F_0 contours," in *ISCA Tutorial and Research Workshop on Statistical And Perceptual Audition*, September 2010, pp. 43–48.

[9] Y. Xu and Q. E. Wang, "Pitch targets and their realization: Evidence from Mandarin Chinese," *Speech Communication*, vol. 33, no. 4, pp. 319–337, March 2001.

[10] Y. Xu and A. Wallace, "Multiple effects of consonant manner of articulation and intonation type on F_0 in English," *Journal of the Acoustical Society of America*, vol. 115, no. 5, p. 2397, 2004.

[11] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch–adaptive time–frequency smoothing and an instantaneous–frequency–based F_0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, April 1999.

[12] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, November 2009.

[13] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, June 2000, pp. 1315–1318.

[14] Phonetics Lab, "ASCCD: Read discourse corpus with prosodic, segmental and syntactic annotation," http://ling.cass.cn/yuyin/english/resc6.htm, Institute of Linguistics, CASS.

[15] P. N. Garner, M. Cernak, and P. Motlicek, "A simple continuous pitch estimation algorithm," *IEEE Signal Processing Letters*, vol. 20, no. 1, pp. 102–105, January 2013.