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

Model-based Parametric Prosody Synthesis with Deep Neural Network

Hao Liu[1], *Heng Lu*[2], *Xu Shao*[2], *Yi Xu*[1]

[1]Department of Speech, Hearing and Phonetic Sciences, University College London, UK
[2]Nuance Communications, USA

{h.liu.12, yi.xu}@ucl.ac.uk, {heng.lu, xu.shao}@nuance.com

Abstract

Conventional statistical parametric speech synthesis (SPSS) captures only frame-wise acoustic observations and computes probability densities at HMM state level to obtain statistical acoustic models combined with decision trees, which is therefore a purely statistical data-driven approach without explicit integration of any articulatory mechanisms found in speech production research. The present study explores an alternative paradigm, namely, model-based parametric prosody synthesis (MPPS), which integrates dynamic mechanisms of human speech production as a core component of F0 generation. In this paradigm, contextual variations in prosody are processed in two separate yet integrated stages: linguistic to motor, and motor to acoustic. Here the motor model is target approximation (TA), which generates syllable-sized F0 contours with only three motor parameters that are associated to linguistic functions. In this study, we simulate this two-stage process by linking the TA model to a deep neural network (DNN), which learns the "linguistic-motor" mapping given the "motor-acoustic" mapping provided by TA-based syllable-wise F0 production. The proposed prosody modeling system outperforms the HMM-based baseline system in both objective and subjective evaluations.

Index Terms: F0 modeling, prosody, syllable, target approximation, speech synthesis, deep neural network

1. Introduction

Statistical parametric speech synthesis (SPSS) [1] has been dominating the field of text-to-speech (TTS) synthesis in the last decade. Its success mainly relies on the use of hidden Markov models (HMMs) and Gaussian mixture models (GMMs) [2]. In its conventional approach, spectral and F0 features are first extracted frame-wise from training data. Then linguistic context-dependent phone HMMs, which represent nonstationary acoustic feature distributions by a sequence of hidden states (usually five states per phone model), are trained via the maximum likelihood (ML) criterion. State-level single Gaussian or GMM conditional probability density functions (PDFs) are computed. A binary decision tree is then constructed to cluster and tie contextually-similar states together and set up a mapping from contextual linguistic features (obtained from text analysis via front end) to GMM-HMM states. At the synthesis stage, acoustic parameters are generated from decision-tree-selected HMM sequence based on the maximum likelihood parameter generation (MLPG) algorithm [3] with static and dynamic features [4] before being sent to a vocoder (e.g. STRAIGHT [5]) for synthesizing waveforms.

More recently, along with its successful application in automatic speech recognition (ASR), deep neural network (DNN) has shown its power to improve the accuracy of statistical acoustic modeling in speech synthesis [6–9]. In general, it overcomes some issues (e.g. complexity limit, training data fragmentation) faced by decision tree-based approaches by offering a highly complex and nonlinear yet efficient mapping between linguistic features and state-level acoustic features via a compact hierarchical structure. During synthesis, acoustic features are predicted by DNN and then set as means of Gaussian distributions. Some very recent studies demonstrate even better results by implicitly embedding the parameter generation process inside recurrent neural network (RNN) with long short-term memory (LSTM) architecture and directly predicting static acoustic feature sequence [10, 11]. The DNN/RNN-based approaches have become state-of-the-art in speech synthesis nowadays.

However, amongst the acoustic features, F0 exhibits strong segmental as well as supra-segmental characteristics which have been hard to model at the phone/state level [12]. All aforementioned approaches tried to resolve this issue by considering numerous contextual prosodic factors (e.g. phone position in phrase/sentence) in an attempt to better represent longer-term F0 patterns [13, 14]. Hierarchical constraint strategies have also been developed either by layer-wise modeling prosodic components at different phonetic levels [15, 16] or by relying on discrete cosine transform (DCT) to capture phrase level F0 patterns [17–19]. RNN-based approaches, on the other hand, offer a solution of sequence-to-sequence mapping so that the dynamic process of speech production is implicitly embedded. What is common in these methods, however, is that they treat articulatory mechanisms of F0 production only implicitly. Even in approaches that try to integrate articulatory features into speech synthesis [20–24], articulatory mechanisms are treated as unknown.

Instead of exclusively using statistical modeling to process all the variations, here we explore a two-stage paradigm: model-based parametric prosody synthesis (MPPS). In this paradigm, contextual linguistic features are associated to motor parameters of an articulatory F0 production model—Stage-I: linguistic to motor mapping, which dynamically generates F0 contours that can be mapped to those of natural speech—Stage-II: motor to acoustic mapping. The learning of linguistic-motor mapping can be achieved through either a decision tree (DT) or a DNN. During synthesis, the predicted motor parameters are then used to dynamically generate F0 contours for a TTS framework. In this way, the articulatory dynamics of F0 production becomes an integral part of both the learning and synthesis processes. Some previous research [25–28] also experimented on the same articulatory F0 production model as used here via various approaches. However, the model was used in either a hierarchical structure or a post-filtering way but seldom used on its own. While its efficacy has been demonstrated, the improvements introduced by the model were not satisfactory enough.

http://dx.doi.org/10.21437/Interspeech.2016-1325

2. Methods

2.1. Model-based Parametric Prosody Synthesis

In conventional SPSS approaches, all the acoustic features (e.g. MGC, F0 and duration) are jointly modeled at the phone/state level. For example, the production of an F0 contour is formed by a sequence of signal frames generated separately subject to learned state-level Gaussian distributions. In this way, articulatory mechanisms are largely ignored. While empirical successes haven been seen in SPSS, it is still widely recognized that acoustic modeling is a critical limit of speech synthesis [12]. Especially, positive results in F0 modeling have not been achieved [6, 29].

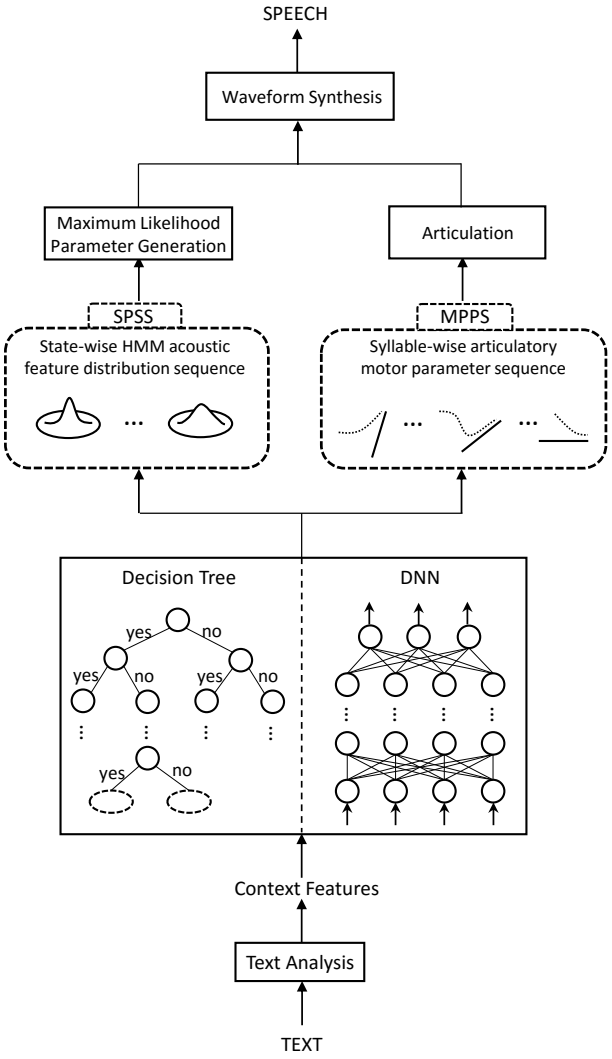

Figure 1: *Diagram of SPSS vs. MPPS for prosody modeling.*

Therefore, a new speech synthesis paradigm is proposed here, namely, model-based parametric prosody synthesis (MPPS). It addresses limitations in conventional SPSS paradigm by bringing in two major improvements:

- Instead of modeling acoustics with phone HMMs based on frame-wise observation and generation, MPPS aims at syllable-level acoustic modeling and segmental generation. Syllable is considered as a more plausible unit of speech production modeling than frames [30]. It also helps to resolve the temporal dependency problem which is currently coped with by tuning computationally expensive RNNs.
- In contrast to SPSS, MPPS does not heavily rely on sequences of Gaussian distributions for acoustic representation. Instead, it utilizes established articulatory models to represent phonetic segments from the perspective of motor control. By learning a small number of motor parameters, MPPS makes acoustic modeling more economical and effective.

As the prosody modeling diagram shown in Figure 1, SPSS and MPPS share the same upstream from text analysis to DT/DNN mapping. The DT outputs of SPSS stream are acoustic feature distributions which directly relate to detailed acoustic realizations via MLPG algorithm plus smoothing. In the MPPS stream, in contrast, articulatory motor parameters are predicted from DNN, which is then executed via an articulatory model to generate segmental trajectories.

In other words, conventional SPSS paradigm jumps over the physical process of human speech production and sets up a frame-wise direct "linguistic-acoustic" mapping, while MPPS models the human speech production pipeline with a segmental "linguistic-motor-acoustic" mapping.

2.2. The articulatory F0 production model

Target approximation (TA) [31], shown in Figure 2, is the articulatory F0 production model that is used in the current MPPS paradigm. Its basic concept is that continuous surface F0 contours are the results of successive, non-overlapping articulatory movements, each approaching an underlying target associated with a host syllable.

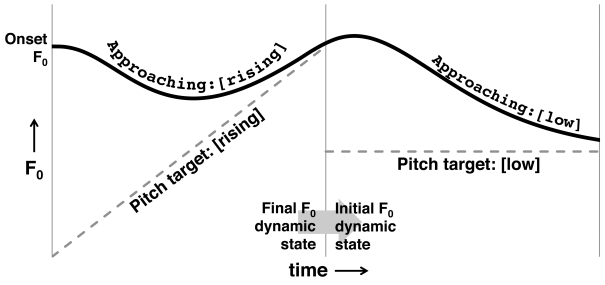

Figure 2: *Target approximation model.*

The concept of the TA model has been algorithmically implemented as the quantitative target approximation (qTA) model [32]. In this model, a target can be either static or dynamic, which can be represented by a simple linear equation

$$x(t) = mt + b, \tag{1}$$

where m and b represent the spatial properties of the target in terms of target height and slope, respectively, and t is time relative to the onset of the host syllable.

The realization of the target is through a third-order critically damped linear system defined by the following equation

$$f_0(t) = x(t) + (c_1 + c_2 t + c_3 t^2)e^{-\lambda t}, \tag{2}$$

where $f_0(t)$ is the complete form of the fundamental frequency in semitones, $x(t)$ is the forced response and the polynomial

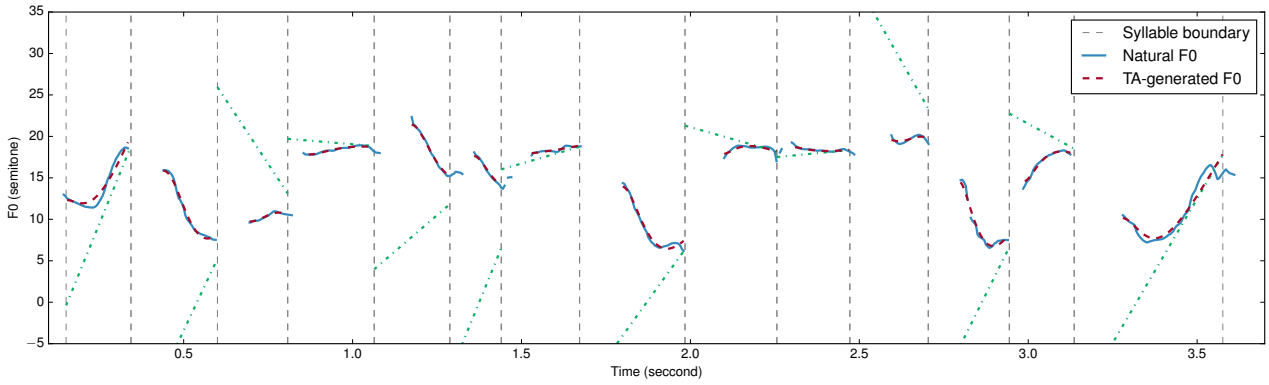

Figure 3: *Syllabified natural F0 contours and those generated by the TA model with local optimal motor parameters (a training set utterance).*

and the exponential are the natural response. λ is the rate of target approximation, i.e., how rapidly the target is approached, which controls the strength of target approximation movement. The transient coefficients c_1, c_2 and c_3 are jointly determined by the initial F0 dynamic state of the syllable, consisting of F0 level, velocity as well as acceleration transferred from the offset of the preceding syllable (as such they are not free parameters):

$$c_1 = f_0(0) - b, \tag{3}$$

$$c_2 = f_0'(0) + c_1\lambda - m, \tag{4}$$

$$c_3 = (f_0''(0) + 2c_2\lambda - c_1\lambda^2)/2. \tag{5}$$

At the end of the syllable, the final F0 dynamic state is transferred to the next syllable to become its initial state, which results in a smooth and continuous F0 trajectory across the syllable boundary (Figure 2).

In short, the process of F0 production is simulated by the TA model at the syllable level by controlling just three motor parameters (m, b and λ), and this process forms a deterministic "motor-acoustic" mapping. To achieve the "linguistic-motor" mapping, the motor parameters can be trained with input features via DNN learning, as is done in this study.

3. Experiments

3.1. Experimental setup

A Mandarin Chinese speech dataset was used in the experiment, which consisted of 6233 phonetically balanced utterances (around 5 hours) as the training set and 60 extra utterances as the test set. Of the 6233 utterances in the training set, 701 were questions and the rest were statements. The test set was evenly divided into statements and questions. The dataset was recorded from a female speaker in 22.5kHz/16bit format. Spectral analysis was performed with 25-ms hamming window shifted every 5 ms. Extracted acoustic features include logarithmic F0 (by the RAPT algorithm [33]), 31-order Mel-generalized cepstrum (MGC) coefficients as well as their delta and delta-delta. Phone durations were obtained through forced alignment, and the contextual linguistic features include tri-phone, phone position in word and in phrase, syllable and its position in word and in phrase, word/phrase length, sentence length, sentence type, phone/syllable stress, prominence, word part-of-speech (POS), etc. To test the proposed paradigm, we built two systems for comparison, one is HMM-based SPSS as the baseline with all the features and the other is DNN-based MPPS with only features above the syllable level.

The HMM-based SPSS baseline system is typical as used in other studies with five-state left-to-right-with-no-skip HMM contextual phone models, and each HMM state is modeled by a single Gaussian output distribution with diagonal covariance. In particular, the log F0s with voiced/unvoiced observations were modeled by multi-space probability distributions (MSD) [13]. A total number of 5268 questions were used for decision tree-based state clustering with the minimum description length (MDL) criterion factor α set to 1 [34].

For MPPS, because it is a two-stage "linguistic-motor-acoustic" paradigm, each stage needs to be optimized separately to achieve an optimal end-to-end mapping. As described above, the TA process is implemented by a dynamical system, the Levenberg-Marquardt nonlinear least-squares method [35] can be easily applied to find locally most fitted TA parameters of each syllable [36]. Although it is increasingly popular to do interpolation on unvoiced segments in order to obtain overall continuous F0 contours for universal modeling [19, 37], and it is also reasonable to do so with the hypothesis that articulatory movements are continuous even during unvoiced session [38], here in our local fitting task unstable TA parameters were found due to the undesirable errors introduced by the pitch tracking and interpolation in the current system. Therefore, heuristic strategies were developed to skip initial unvoiced parts in syllables with voiceless consonants, and optimal TA parameters were obtained based only on the voiced parts. Based on previous studies [32, 38], certain ranges were applied to limit the search range of TA parameters: $m \in [-100, 100]$, $b \in [-30, 30]$ and $\lambda \in [1, 80]$.

Figure 3 illustrates the performance of TA model when local optimal parameters were found (m and b are plotted as underlying pitch targets defined in TA, λ is not presented).

The input dimensions of the DNN-based MPPS were 287 formed by 35 binary features with one-hot encoding and 23 numeric features with zero-mean unit-variance normalization. The output dimensions were 6 including the three TA parameters as well as dynamic onset state of the syllable (F0 level, velocity and acceleration), which were normalized to $[0.01, 0.99]$ based on their minimum and maximum values in the dataset. The best DNN structure achieved to date for this system is 3 hidden layers $\times$ 1024 nodes, more layers with fewer nodes on each layer achieved similar results. The activation functions used were hy-

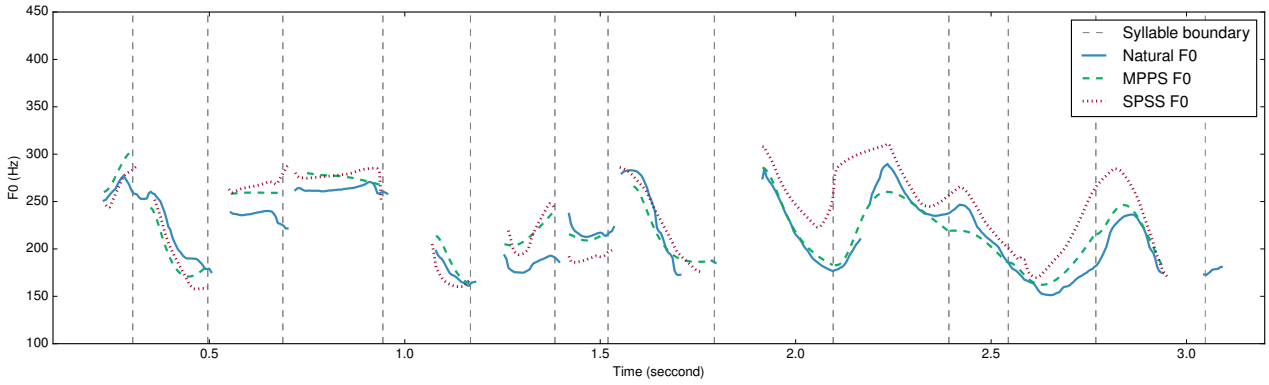

Figure 4: *Syllabified natural F0 contours together with those generated by TA model with predicted motor parameters in MPPS and those generated via MLPG with PDFs in SPSS (a test set utterance).*

perbolic tangent for the hidden layers and linear for the output layer. The DNN was trained with the backpropagation algorithm using mini-batch stochastic gradient descent (SGD) as the optimizer.

3.2. Evaluations

As mentioned earlier, the test set consisted of 30 statements and 30 questions. The evaluations were therefore run separately for them to show differences in performance. Both objective and subjective evaluations were conducted. During synthesis, durations obtained through state-level forced alignment on the test set were used for both systems. The MPPS system generally followed the voiced/unvoiced decisions predicted by the SPSS system.

Table 1: *Objective scores of each system on different sentence types.*

System	Statement		Question	
	RMSE	Corr.	RMSE	Corr.
SPSS	22.32	0.91	33.67	0.85
MPPS	21.10	0.91	33.20	0.86

For objective test, F0 discrepancies between natural and synthetic speech were measured with root mean square errors (RMSE) in Hz as well as correlation scores reported for each system in each task. As shown in Table 1, while the systems achieved similar correlation scores, MPPS outperformed SPSS system in RMSE tests for both statement (-1.22) and question (-0.47) tasks.

Subjective test was focused on comparing naturalness of sentence prosody only. Similar to the objective test, statements and questions were tested separately. Subjects were asked to do A/B preference test based on the synthetic sentence pairs that they heard. Fifteen sentence pairs for each sentence type were randomly selected from the test set. Twenty native speakers participated in the test. The preference scores are shown in Table 2 with p-values from two-tailed t-test. It can be seen that the MPPS system achieved significantly better performance for both sentence types. More importantly, the MPPS system doubled its score in questions from that of statements.

A comparison between F0 contours generated by the two systems is shown in Figure 4. It can be seen that the F0 contours generated by the MPPS system show much greater resemblance to the natural ones than those generated by the SPSS baseline system. Note that the MPPS system only needs 3 parameters (m, b and λ) with syllable onset state (transferred from the offset of its preceding syllable if no voiceless interruption, otherwise the predicted values are used) to generate any syllable, whereas the SPSS system needs at least 10 (5 states $\times$ 2 parameters per state including mean and variance if MSD is not considered) for a single-phone syllable, and most syllables are multi-phone in spontaneous speech. Therefore, the MPPS system is also more economical in acoustic parameter generation.

Table 2: *Subjective preference scores (%) of each system on different sentence types. Systems achieved significantly better preference ($p < 0.01$) are in bold font. N/P stands for no preference.*

Statement		Question			
SPSS	MPPS	SPSS	MPPS	N/P	p-value
15.8	**32.0**	—	—	52.2	1.22×10^{-5}
—	—	9.5	**67.0**	23.5	6.74×10^{-14}

4. Conclusions

This study tested a model-based parametric prosody synthesis (MPPS) paradigm, which integrates an articulatory model of F0 production into the existing speech synthesis paradigm with DNN. The model is the target approximation (TA) model, which serves as the link between linguistic functions and surface acoustics. With TA, syllable is the basic prosody modeling unit instead of frames, which greatly increases processing economy. The results of our testing show that the MPPS system outperforms the SPSS baseline system in both objective and subjective evaluations. Thus MPPS is not only economical, but also may improve synthesis quality. Future work can try to combine the TA model with an RNN so that utterance-level dynamics can be captured by the RNN while syllable-level dynamics are effectively and economically simulated by the TA model. This may significantly reduce computational cost of existing RNN methods without degrading perceptual quality.

5. References

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

[2] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden Markov models," *Proc. IEEE*, vol. 101, no. 5, pp. 1234–1252, 2013.

[3] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP*, vol. 3, 2000, pp. 1315–1318.

[4] H. Zen, K. Tokuda, and T. Kitamura, "Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences," *Comput. Speech Lang.*, vol. 21, no. 1, pp. 153–173, 2007.

[5] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Commun.*, vol. 27, no. 3, pp. 187–207, 1999.

[6] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. ICASSP*, 2013, pp. 7962–7966.

[7] H. Zen and A. Senior, "Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis," in *Proc. ICASSP*, 2014, pp. 3844–3848.

[8] Z.-H. Ling, L. Deng, and D. Yu, "Modeling spectral envelopes using restricted Boltzmann machines and deep belief networks for statistical parametric speech synthesis," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 21, no. 10, pp. 2129–2139, 2013.

[9] Y. Qian, Y. Fan, W. Hu, and F. K. Soong, "On the training aspects of Deep Neural Network (DNN) for parametric TTS synthesis," in *Proc. ICASSP*, May 2014, pp. 3829–3833.

[10] Y. Fan, Y. Qian, F.-L. Xie, and F. K. Soong, "TTS synthesis with bidirectional LSTM based recurrent neural networks." in *Proc. Interspeech*, 2014, pp. 1964–1968.

[11] H. Zen and H. Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis," in *Proc. ICASSP*, 2015, pp. 4470–4474.

[12] Z.-H. Ling, S.-Y. Kang, H. Zen, A. Senior, M. Schuster, X.-J. Qian, H. M. Meng, and L. Deng, "Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends," *IEEE Signal Process. Mag.*, vol. 32, no. 3, pp. 35–52, 2015.

[13] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling," in *Proc. ICASSP*, vol. 1, 1999, pp. 229–232.

[14] H. Zen and N. Braunschweiler, "Context-dependent additive log F0 model for HMM-based speech synthesis." in *Proc. Interspeech*, 2009, pp. 2091–2094.

[15] M. Lei, Y.-J. Wu, F. K. Soong, Z.-H. Ling, and L.-R. Dai, "A hierarchical F0 modeling method for HMM-based speech synthesis." in *Proc. Interspeech*, 2010, pp. 2170–2173.

[16] Y. J. Wu and F. Soong, "Modeling pitch trajectory by hierarchical HMM with minimum generation error training," in *Proc. ICASSP*, 2012, pp. 4017–4020.

[17] J. Teutenberg, C. Watson, and P. Riddle, "Modelling and synthesising F0 contours with the discrete cosine transform," in *Proc. ICASSP*, 2008, pp. 3973–3976.

[18] Z. Wu, Y. Qian, F. K. Soong, and B. Zhang, "Modeling and generating tone contour with phrase intonation for Mandarin Chinese speech," in *Proc. ISCSLP*, 2008, pp. 1–4.

[19] X. Yin, M. Lei, Y. Qian, F. K. Soong, L. He, Z.-H. Ling, and L.-R. Dai, "Modeling DCT parameterized F0 trajectory at intonation phrase level with DNN or decision tree." in *Proc. Interspeech*, 2014, pp. 2273–2277.

[20] T. Toda, A. W. Black, and K. Tokuda, "Mapping from articulatory movements to vocal tract spectrum with Gaussian mixture model for articulatory speech synthesis," in *Proc. SSW-5*, 2004, pp. 31–36.

[21] Z.-H. Ling, K. Richmond, J. Yamagishi, and R.-H. Wang, "Articulatory control of HMM-based parametric speech synthesis driven by phonetic knowledge," in *Proc. Interspeech*, 2008, pp. 573–576.

[22] Z.-H. Ling, K. Richmond, J. Yamagishi, and R.-H. Wang, "Integrating articulatory features into HMM-based parametric speech synthesis," *IEEE Trans. Audio Speech Lang. Process.*, vol. 17, no. 6, pp. 1171–1185, 2009.

[23] Z.-H. Ling, K. Richmond, and J. Yamagishi, "Vowel creation by articulatory control in HMM-based parametric speech synthesis," in *Proc. Interspeech*, 2012, pp. 991–994.

[24] Z.-H. Ling, K. Richmond, and J. Yamagishi, "Articulatory control of HMM-based parametric speech synthesis using feature-space-switched multiple regression," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 21, no. 1, pp. 207–219, 2013.

[25] Z. Zhang, X. Wang, Y. Yu, and X. Wu, "Hierarchical pitch target model for Mandarin speech," in *Proc. ISCSLP*, 2010, pp. 378–382.

[26] H. Pang, Z. Wu, and L. Cai, "Modeling pitch contour of Chinese Mandarin sentences with the PENTA model," *Tsinghua Sci. Technol.*, vol. 17, no. 2, pp. 218–224, 2012.

[27] X. Na and P. N. Garner, "Convolutional pitch target approximation model for speech synthesis," Idiap, Tech. Rep. No. EPFL-REPORT-192548, 2013.

[28] L. Gao, Z.-H. Ling, L.-H. Chen, and L.-R. Dai, "Improving F0 prediction using bidirectional associative memories and syllable-level F0 features for HMM-based Mandarin speech synthesis," in *Proc. ISCSLP*, 2014, pp. 275–279.

[29] S. Kang, X. Qian, and H. Meng, "Multi-distribution deep belief network for speech synthesis," in *Proc. ICASSP*, 2013, pp. 8012–8016.

[30] Y. Xu, "Speech melody as articulatorily implemented communicative functions," *Speech Commun.*, vol. 46, no. 34, pp. 220–251, 2005.

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

[32] S. Prom-on, Y. Xu, and B. Thipakorn, "Modeling tone and intonation in Mandarin and English as a process of target approximation," *J. Acoust. Soc. Am.*, vol. 125, no. 1, pp. 405–424, 2009.

[33] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech Coding and Synthesis*, pp. 495–518, 1995.

[34] K. Shinoda and T. Watanabe, "Acoustic modeling based on the MDL criterion for speech recognition," in *Proc. Eurospeech*, 1997, pp. 99–102.

[35] J. J. Moré, "The Levenberg-Marquardt algorithm: implementation and theory," in *Numerical Analysis*. Springer, 1978, pp. 105–116.

[36] M. Newville, T. Stensitzki, D. B. Allen, and A. Ingargiola, "LMFIT: Non-Linear Least-Square Minimization and Curve-Fitting for Python," 2014. [Online]. Available: dx.doi.org/10.5281/zenodo.11813

[37] K. Yu and S. Young, "Continuous F0 modeling for HMM based statistical parametric speech synthesis," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 5, pp. 1071–1079, 2011.

[38] Y. Xu and S. Prom-on, "Toward invariant functional representations of variable surface fundamental frequency contours: Synthesizing speech melody via model-based stochastic learning," *Speech Commun.*, vol. 57, pp. 181–208, 2014.