A template-based approach for speech synthesis intonation generation using LSTMs

Srikanth Ronanki, Gustav Eje Henter, Zhizheng Wu, Simon King

The Centre for Speech Technology Research (CSTR), The University of Edinburgh, UK

srikanth.ronanki@ed.ac.uk, simon.king@ed.ac.uk

Abstract

The absence of convincing intonation makes current parametric speech synthesis systems sound dull and lifeless, even when trained on expressive speech data. Typically, these systems use regression techniques to predict the fundamental frequency (F0) frame-by-frame. This approach leads to overly-smooth pitch contours and fails to construct an appropriate prosodic structure across the full utterance. In order to capture and reproduce larger-scale pitch patterns, this paper proposes a template-based approach for automatic F0 generation, where per-syllable pitch-contour templates (from a small, automatically learned set) are predicted by a recurrent neural network (RNN). The use of syllable templates mitigates the over-smoothing problem and is able to reproduce pitch patterns observed in the data. The use of an RNN, paired with connectionist temporal classification (CTC), enables the prediction of structure in the pitch contour spanning the entire utterance. This novel F0 prediction system is used alongside separate LSTMs for predicting phone durations and the other acoustic features, to construct a complete text-to-speech system. We report the results of objective and subjective tests on an expressive speech corpus of children's audiobooks, and include comparisons to a conventional baseline that predicts F0 directly at the frame level.

Index Terms: speech synthesis, intonation modelling, F0 templates, LSTM, CTC

1. Introduction

Despite decades of research, synthetic speech is still not convincingly natural [1]. A particular shortcoming is prosody: durations, energy, and intonation tend to be dull, inappropriate, and unattractive. These issues prevent the use of synthesised speech in otherwise appealing applications, such as audiobooks.

In this paper we focus on intonation (pitch contour) generation in statistical parametric speech synthesis (SPSS). Although intonation is a supra-segmental property, conventional approaches, such as HTS [2], predict F0 frame-by-frame, based on limited linguistic contextual information (quinphone, part-of-speech and positional information). Predictions within an HMM state are assumed conditionally independent of *surrounding F0 predictions*, given the linguistic information. Surrounding F0 values only affect the local smoothing of the final contour (via MLPG [3]). Such a myopic approach may not be the best way to produce utterance-level supra-segmental structure.

We propose to generate more variable and naturalistic intonation through the use of syllable-level pitch-contour templates learned from data. The hope is that treating intonation prediction as a *classification* rather than a *regression problem* will mitigate the over-smoothing seen in conventional techniques. By using a recurrent neural network (RNN) to predict the pitch-template sequence, long-range information about linguistic context and surrounding predictor state can influence the output, enabling high-level prosodic structure to be expressed.

To the best of our knowledge, our proposal to use an RNN to predict syllable-level F0 templates is new. We also consider improvements to this basic idea, including learned smoothing of template joins and hierarchical prediction of the templates, and evaluate against a state-of-the-art SPSS baseline.

2. Background

2.1. Conventional intonation modelling

Conventional HMM+regression tree speech synthesis predicts frame-wise F0 from contextual linguistic features either using a multi-space probability distribution (MSD) [4] or by continuous F0 modelling [5]. Recently [6] the regression tree has been replaced by a deep neural network (DNN), significantly improving subjective naturalness [7].

2.2. Long memory in prosody modelling

Recurrent neural networks (RNNs), in principle, enable long-range dependencies to affect prediction output. They provide superior output compared to feedforward DNNs for acoustic modelling [8, 9], at least when natural speech durations are used for synthesis. Our investigation considers the more practical scenario of F0 prediction with RNNs in a full synthesis system, without 'oracle' cues from natural speech durations. [10] proposes predicting state-level F0 statistics and durations using so-called long short-term memory (LSTM) units. However, predictions were made at the state level (3-state HMM), while we predict syllable-level templates that do not assume within-state stationarity. We use simplified LSTMs [9], and parametrise our F0 contours using templates. [11] proposes two different model structures: cascade and parallel DNN, to embody hierarchical and additive properties of F0. Similarly, part of our approach considers using a hierarchy of DNNs for predicting F0.

2.3. Template-based F0 generation

While the vast majority of F0 prediction approaches are based on regression rather than classification, the idea of clustering and/or discretising F0 contours is not new (cf. [12, 13]). In [14], k-means clustering identified representative accent shapes over syllables, with phrase-level structure predicted by a regression tree. Modelling local and global components of F0 separately was shown to improve upon conventional F0 models.

We combine key elements of a template-based approach similar to [14] with much stronger, long-range predictors (LSTMs). The use of templates is intended to capture and reproduce salient, syllable-level features of the F0 contour without

over-smoothing. The choice of template per syllable should *not* be made independently of surrounding choices, so we experimented with the connectionist temporal classification (CTC) framework [15] to predict *sequences* of templates.

2.4. Time-frequency decomposition for F0 modelling

In [16, 17], discrete cosine transform (DCT) coefficients were proposed for representing F0 patterns. The use of DCTs to represent F0 at, e.g., syllable and phrase levels, enables a compact representation of complex contours of varying durations [18]. The continuous wavelet transform (CWT) has also been found to improve F0 modelling in HMM SPSS [19]. [20] explored a multi-level representation of F0 by combining both DCT and CWT representation, at different wavelet scales representing F0-contour variations from phone up to utterance/phrase.

In [21], a regression tree was used to model DCT-parametrised phrase-level F0 trajectories, to generate smoothly varying contours. We similarly consider using the DCT to parameterise each syllable-level template, but with template class predicted, and joins smoothed, using deep learning techniques.

3. Proposed template-based approach

Our proposed method has three parts (Figure 1): 1) An inventory of syllable-level templates derived from training data; 2) A neural network classifier to predict template class from input text features; 3) F0 contour reconstruction from this sequence of templates.

3.1. Creating the inventory of syllable F0 templates

In a training database, the F0 contour (on the equivalent rectangular bandwidth (ERB) scale [22]) of each utterance is interpolated through unvoiced regions, including pauses then segmented into syllables using force-aligned phone labels. The database is described in Section 4.

We then (approximately) decompose each syllable contour into a sum of N weighted zero-phase cosine functions. If x is a contour of length N, then

$$c[k] = 2w[k] \sum_{n=0}^{N-1} x[n] \cos\left(\frac{\pi(2n+1)k}{2N}\right) \qquad (1)$$

for $0 \leq k < N$, where

$$w[k] = \begin{cases} \frac{1}{\sqrt{4N}} & \text{if } k = 0 \\ \frac{1}{\sqrt{2N}} & \text{if } 1 \leq k \leq N. \end{cases} \qquad (2)$$

This yields outputs: c_0, representing the mean F0 over the syllable, and $\boldsymbol{c} = [c_1, \ldots, c_{N-1}]^\intercal$, representing the shape of the contour. This representation is clustered hierarchically [23] to group similar intonation pattern vectors $\boldsymbol{c}$ together, as follows:

1. Assign each syllable F0 contour to a separate cluster. With M such contours there will be M clusters, each containing a single contour.
2. Find the pair of clusters with the smallest Euclidean distance between their mean contours and merge them into a single cluster.
3. Repeat previous step until stopping criterion is met.

We stop when the correlation between clusters calls below a threshold, or until we arrive at a specific number of clusters. Figure 2 illustrates a 6-template model. Figure 3 shows an example utterance represented as a sequence of these templates.

Following previous work on intonation modelling using DCT [17, 18], we used $N = 9$ coefficients for our model.

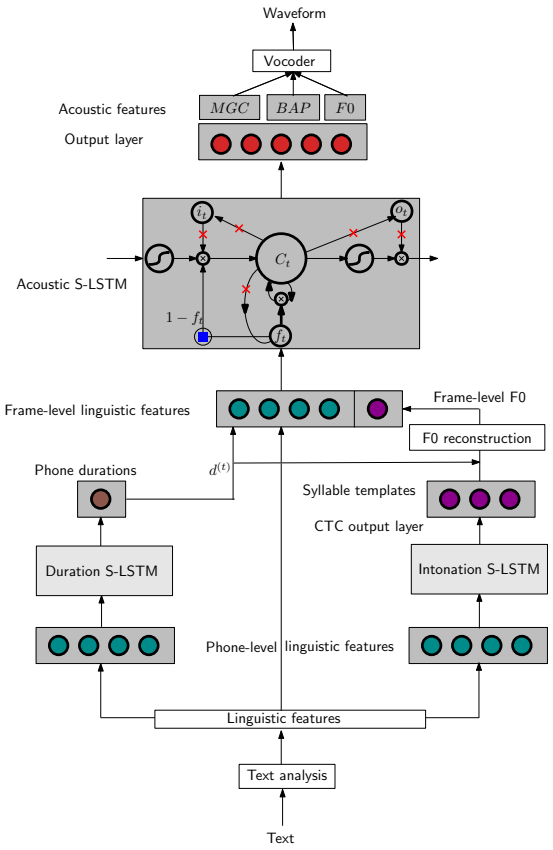

Figure 1: Schematic diagram of the proposed synthesis system CTC in Section 3.2.2 using syllable templates. A single LSTM unit is shown, although there is a full hidden layer of such units. Connections crossed out in red are present in standard LSTMs, but are omitted in the simplified LSTM units [9] used here.

3.2. Predicting templates using neural network classifiers

We now describe two different systems for predicting F0 templates using neural networks:

1. A hierarchical deep neural network classifier (HC).
2. A simplified LSTM with a connectionist temporal classification (CTC) output layer (SLSTM-CTC).

3.2.1. Hierarchical classifier

Table 1 reveals that the frequency distribution of templates is far from uniform. Template 4 (Figure 2), by far the flattest, accounts for 65% of the training data, presumably corresponding to unstressed syllables. Classifications made by a DNN trained to minimise either mean square error or cross-entropy on this data were even more biased towards predicting the single most frequent class (cf. "DNN" in Table 1), since it had such a large prior probability. This would produce flat and boring intonation.

We propose to use a hierarchy of DNN classifiers to diversify F0 template generation. A first classifier predicts whether or not to use the most frequent (and, here, flattest) template for the current syllable. If a less common template is called for, a second classifier chooses among the remaining templates. This gave a more diverse template distribution (cf. "HC" in Table 1), which should produce more interesting synthetic speech.

In our implementation, the first classifier is a simplified LSTM (S-LSTM) as in [9], while the second is a feed-forward

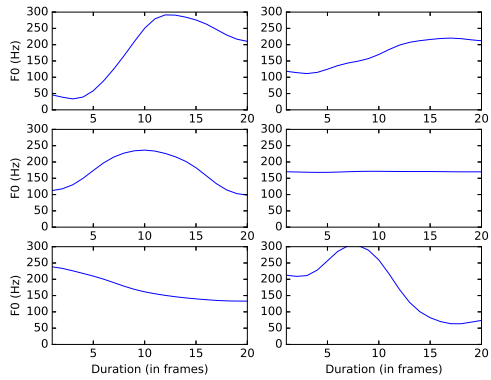

Figure 2: A set of six syllable F0 templates found by clustering of the data described in Section 4.1, plotted at the average F0 (180 Hz) and syllable duration (20 frames) of the speaker.

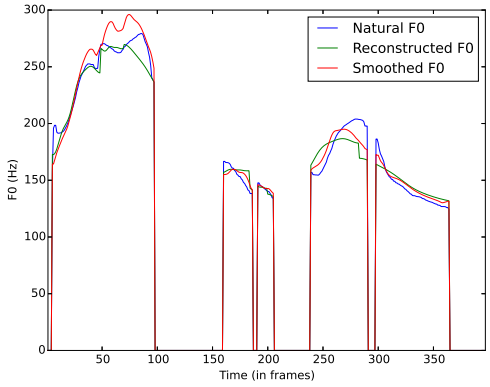

Figure 3: Natural F0 and reconstructed F0 contours for the sentence "Goldilocks and the three bears" at 5 ms frame rate.

DNN. (Using an RNN would not be straightforward, since the state has to be propagated through time instances where no second-level classification is needed.)

3.2.2. SLSTM-CTC

In another approach, we added a connectionist temporal classification (CTC) output layer [15] mapping from phone-level linguistic features to syllable-level templates, after the S-LSTM layer. The CTC output is a softmax output layer with $N + 1$ nodes, representing a probability distribution over the syllable templates plus an additional "blank" unit. The blank unit is the most common softmax output, particularly for phones in the beginning and middle of syllables; only when the network believes it is at the end of a syllable is the predicted template output node activated. Similar to HMMs, the CTC loss function also makes use of the forward-backward algorithm to make utterance-spanning decisions. This should enable a more appropriate distribution of predicted templates even for non-uniform data, without the need for two different predictors as in the proposed hierarchical DNN classifier.

3.3. F0 contour reconstruction

A preliminary F0 contour is reconstructed from the sequence of predicted templates using the inverse discrete cosine transform (IDCT) per template

$$x[n] = \frac{c[0]}{\sqrt{N}} + \sqrt{\frac{2}{N}} \sum_{k=1}^{N-1} c[k] \cos\left(\frac{\pi(n+0.5)k}{N}\right). \quad (3)$$

Template	1	2	3	4	5	6
Train	852	5216	1853	26725	5784	1013
Dev.	26	139	50	653	147	28
Test	65	362	106	1553	377	40
DNN	0	127	4	2247	155	0
HC	15	298	27	1676	499	18
CTC	16	200	61	1958	287	11

Table 1: Template counts in the data (see Section 4.1) and corresponding test-set predictions by the methods of Section 4.3.

with the mean F0 (c_0) taken from the baseline system and c_1–c_{N-1} from the selected template. This contour may be discontinuous at template joins, which was found to degrade perceived naturalness. To solve this, the discontinuous F0 contour value at each frame was appended to the frame-level linguistic features used as inputs to the acoustic model (see Figure 1), allowing that model to predict the final F0 (as statics and dynamics) and the voiced/unvoiced flag needed for waveform generation. During training, a template decomposition of the natural F0 was used to train the S-LSTM acoustic model to reconstruct smooth and natural F0 contours from discontinuous input. Figure 3 shows the preliminary (discontinuous) reconstruction along with the output after smoothing. A pilot listening test indicated that this smoothing removed discontinuity artifacts in the F0 contours.

4. Experimental set-up

4.1. Data

We evaluated on prosodically interesting data from children's audiobooks, as provided by Usborne Publishing Ltd. for the 2016 Blizzard Challenge[1], containing text and speech (downsampled to 16 kHz) of 50 children's audiobooks read by a British female speaker. Three stories with a combined duration of approximately 13 minutes (6% of all data) were set aside as a test set, leaving approximately 3.5 hours of audio for training.

4.2. Feature extraction

The data was force-aligned at state level using monophone HMMs. Festvox's `ehmm` [24] was used to insert pauses into the phone-label sequences based on the acoustics. The resulting label sequences were coupled with text-derived linguistic features encoding a subset of the questions used by the decision-tree clustering in HTS [2]. This produced a vector of 592 binary input features, to which 9 numerical features were appended as in [25] and all features normalised to the range $[0.01, 0.99]$.

The proposed CTC system only utilised the phoneme-level binary input features to predict a distribution over the syllable templates from Sec. 3.1. For HC, per-syllable input feature vectors were created by removing phoneme-level features for position and articulation, and replacing quinphone identities by a syllable-identity vector concatenating 7 50-dimensional one-hot phone vectors, allowing seven phones per syllable.

Duration prediction, identical for all systems, used the unmodified binary features as input to a DNN to predict a six-dimensional vector, comprising five-state durations and total phone duration (the last to regularise training, and thus not used in synthesis).

STRAIGHT [26] was used to extract 60 MGC, 25 BAP and $\log F_0$, along with delta and delta-delta features every 5ms. Per-component mean and variance normalisation was performed.

[1] http://www.synsig.org/index.php/Blizzard_Challenge_2016

	Classification measures		F0 measures	
Model	Accuracy	F1 score	RMSE	Corr.
MSE	-	-	45.9	0.40
HC	61.1%	0.590	46.9	0.36
CTC	63.8%	0.593	46.1	0.40
Oracle	100%	1	40.8	0.58

Table 2: Classification correctness and F1 score of predicted syllable templates, along with RMSE and correlation of the predicted F0 contour, all measured w.r.t. held-out natural speech.

4.3. Reference systems

Alongside the two proposed systems (Section 3.2) we created several reference SPSS systems which, except where noted, also predicted all acoustic outputs from linguistic features: MSE was a framewise-regression baseline predicting F0 using S-LSTMS. BMK used forced-aligned natural durations and natural F0 contours from the test-set recordings. VOC was a top line of vocoded speech. BOT was a bottom line using piecewise-constant F0 per syllable (the mean natural F0). Oracle used templates derived from the natural F0 contour of the test utterance. However, all systems except VOC used the same acoustic S-LSTM to generate frame-level acoustic features (MGC+BAP).

4.4. Training and synthesis

The SLSTM-CTC system used a six-layer network with the first four layers being feed-forward DNNs with `tanh` activation functions of 1024 nodes each, followed by an S-LSTM layer with 512 nodes and finally a softmax layer with seven output nodes. The hierarchical DNN used a similar architecture with softmax output layers of two and five output nodes.

Each network was initialised using small random weights and subsequently optimised using stochastic gradient descent (no pre-training) with a fixed learning rate, manually tuned to yield close-to-optimal results on the development set in 30 epochs or less. Early stopping was used to avoid over-fitting, selecting the model with best dev-set performance.

One acoustic-model S-LSTM and one duration S-LSTM were trained, similar in structure to the intonation-prediction S-LSTM but with a linear output layer. Meta-parameters such as batch size and regularisation criteria followed [25].

For synthesis, `ehmm` phone sequences derived from the test data were used as inputs to duration and intonation prediction. This simply amounts to using oracle pause insertion, with no other acoustics-derived information provided to the predictors.

After duration prediction, a sequence of frame-level linguistic features was generated using the predicted durations, and fed into the acoustic model to generate post-filtered MLPG [3] parameter trajectories as in [25]. Global variance from the dev. data was applied to the generated MGCs and trajectories further enhanced using conventional post-filtering. Finally, waveforms were synthesised using the STRAIGHT vocoder [26] and normalised according to ITU P.56 [27].

5. Results

5.1. Objective measures

To evaluate the different systems, we calculated the root-mean-square error (RMSE) and Pearson correlation between natural and predicted F0 contours. The template-based approaches were additionally evaluated by classification accuracy and F1 score (Table 2). The objective numbers for the Oracle systems (correlation 0.6; lowest RMSE) indicate that natural F0 contours can be reasonably reconstructed from just six syllable templates. Note, however, that Oracle, HC, and CTC all used c_0 values based on the mean F0 predicted by MSE. Instead using natural mean F0 for c_0 increased the Oracle correlation to 0.89.

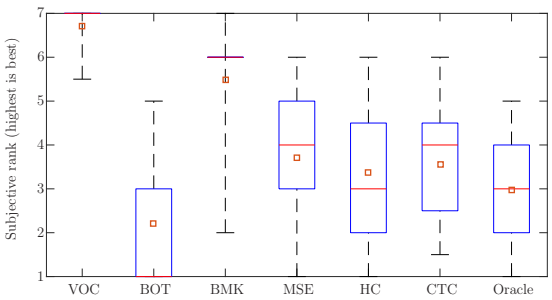

Figure 4: Box plot of aggregate ranks from listening test. Red lines are medians, orange squares means. Box edges (which may coincide) show quartiles. Whiskers cover 90% of the data.

5.2. Subjective evaluation

Next, the systems were evaluated subjectively using a hybrid between a MUSHRA [28] and a preference test. 20 native English speakers with no known hearing impairments, who were remunerated for their time, used GUI sliders to rank the relative naturalness of different systems producing the same sentence. Each listener scored 20 out of 32 held-out sentences in a randomised but balanced design; this was preceded by a two-sentence training phase and a tutorial screen. Tests took place in sound-insulated booths over high-quality headphones.

Figure 4 shows the results. Vocoded speech was clearly ranked the highest, followed by the natural-F0 BMK system; the bottom line was definitely the least preferred. A Mann-Whitney U test and a Wilcoxon signed-rank test both found all pairs of systems to be significantly different ($p < 0.05$) except (MSE, CTC) and (HC, Oracle). Holm-Bonferroni [29] correction was applied because of the multiple comparisons. The proposed CTC system performed as well as but unfortunately not (yet) better than the conventional baseline (MSE).

In stark contrast to the objective results (Table 2), the Oracle F0 contour was the least preferred among the main systems. This may point at an interesting interaction between duration and F0 that would affect many other results in the literature from systems that use oracle values for one, but synthesise the other. Perhaps a mismatch between the durations of the natural speech from which the templates were taken, and the durations of the synthetic speech onto which they were imposed, causes some inconsistency that listeners attend to. This explanation is consistent with [10], which suggests that joint prediction of duration and F0 may be important in DNN-based acoustic models.

6. Conclusion

We have described a classification-based approach to intonation prediction with syllable F0 templates replacing frame-level regression. The listening test suggests two things: that the proposed CTC approach matches the performance of the conventional approach, and has potential to exceed it once the issues with the Oracle template system are overcome; and, that there may be an important but under-explored interaction between duration and F0 that needs careful attention in the future.

This work was partially supported by EPSRC Programme Grant EP/I031022/1 Natural Speech Technology (NST). The NST research data collection can be accessed at http://hdl.handle.net/10283/786.

7. References

[1] S. King, "Measuring a decade of progress in text-to-speech," *Loquens*, vol. 1, no. 1, 2014.

[2] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," in *Proc. SSW*, vol. 6, 2007, pp. 294–299.

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

[4] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proc. Eurospeech*, 1999, pp. 2347–2350.

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

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

[7] O. Watts, G. E. Henter, T. Merritt, Z. Wu, and S. King, "From HMMs to DNNs: where do the improvements come from?" in *Proc. ICASSP*, 2016, pp. 5505–5509.

[8] 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.

[9] Z. Wu and S. King, "Investigating gated recurrent networks for speech synthesis," in *Proc. ICASSP*, 2016, pp. 5140–5144.

[10] R. Fernandez, A. Rendel, B. Ramabhadran, and R. Hoory, "Prosody contour prediction with long short-term memory, bi-directional, deep recurrent neural networks," in *Proc. Interspeech*, 2014, pp. 2268–2272.

[11] X. Yin, M. Lei, Y. Qian, F. K. Soong, L. He, Z.-H. Ling, and L.-R. Dai, "Modeling F0 trajectories in hierarchically structured deep neural networks," *Speech Commun.*, vol. 76, pp. 82–92, 2016.

[12] N. Obin, J. Beliao, C. Veaux, and A. Lacheret, "SLAM: Automatic stylization and labelling of speech melody," in *Proc. Speech Prosody*, 2014, pp. 246–250.

[13] R. Dall and X. Gonzalvo, "JNDSLAM: A SLAM extension for speech synthesis," in *Proc. Speech Prosody*, 2016, pp. 1024–1028.

[14] G. K. Anumanchipalli, L. C. Oliveira, and A. W. Black, "A statistical phrase/accent model for intonation modeling," in *Proc. Interspeech*, 2011, pp. 1813–1816.

[15] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proc. ICML*, 2006, pp. 369–376.

[16] J. Latorre and M. Akamine, "Multilevel parametric-base F0 model for speech synthesis," in *Proc. Interspeech*, 2008, pp. 2274–2277.

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

[19] A. Suni, D. Aalto, T. Raitio, P. Alku, and M. Vainio, "Wavelets for intonation modeling in HMM speech synthesis," in *Proc. SSW*, vol. 8, 2013, pp. 285–290.

[20] M. S. Ribeiro and R. A. J. Clark, "A multi-level representation of $f0$ using the continuous wavelet transform and the discrete cosine transform," in *Proc. ICASSP*, 2015, pp. 4909–4913.

[21] 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.

[22] J. O. Smith III and J. S. Abel, "Bark and ERB bilinear transforms," *IEEE T. Speech Audi. P.*, vol. 7, no. 6, pp. 697–708, 1999.

[23] S. C. Johnson, "Hierarchical clustering schemes," *Psychometrika*, vol. 32, no. 3, pp. 241–254, 1967.

[24] K. Prahallad, A. W. Black, and R. Mosur, "Sub-phonetic modeling for capturing pronunciation variations for conversational speech synthesis," in *Proc. ICASSP*, 2006, pp. I–853–I–856.

[25] Z. Wu, C. Valentini-Botinhao, O. Watts, and S. King, "Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis," in *Proc. ICASSP*, 2015, pp. 4460–4464.

[26] H. Kawahara, "STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds," *Acoust. Sci. Technol.*, vol. 27, no. 6, pp. 349–353, 2006.

[27] *Objective measurement of active speech level*, ITU Recommendation ITU-T P.56, International Telecommunication Union, Telecommunication Standardization Sector, Geneva, Switzerland, March 2011.

[28] *Method for the subjective assessment of intermediate quality levels of coding systems*, ITU Recommendation ITU-R BS.1534-3, International Telecommunication Union Radiocommunication Sector, Geneva, Switzerland, October 2015.

[29] S. Holm, "A simple sequentially rejective multiple test procedure," *Scand. J. Stat.*, vol. 6, no. 2, pp. 65–70, 1979.