# An RNN-based Quantized F0 Model with Multi-tier Feedback Links for Text-to-Speech Synthesis

**Citation for published version:**
Wang, X, Takaki, S & Yamagishi, J 2017, An RNN-based Quantized F0 Model with Multi-tier Feedback Links for Text-to-Speech Synthesis. in *Proceedings Interspeech 2017.* Interspeech, International Speech Communication Association, pp. 1059-1063, Interspeech 2017, Stockholm, Sweden, 20/08/17. https://doi.org/10.21437/Interspeech.2017-246

**Digital Object Identifier (DOI):**
10.21437/Interspeech.2017-246

**Link:**
Link to publication record in Edinburgh Research Explorer

**Document Version:**
Peer reviewed version

**Published In:**
Proceedings Interspeech 2017

# An RNN-based Quantized F0 Model with Multi-tier Feedback Links for Text-to-Speech Synthesis

Xin Wang$^{1,2}$, Shinji Takaki$^{1}$, Junichi Yamagishi$^{1,2,3}$

$^{1}$National Institute of Informatics, Japan
$^{2}$SOKENDAI University, Japan
$^{3}$University of Edinburgh, UK

wangxin@nii.ac.jp, takaki@nii.ac.jp, jyamagis@nii.ac.jp

## Abstract

A recurrent-neural-network-based F0 model for text-to-speech (TTS) synthesis that generates F0 contours given textual features is proposed. In contrast to related F0 models, the proposed one is designed to learn the temporal correlation of F0 contours at multiple levels. The frame-level correlation is covered by feeding back the F0 output of the previous frame as the additional input of the current frame; meanwhile, the correlation over long-time spans is similarly modeled but by using F0 features aggregated over the phoneme and syllable. Another difference is that the output of the proposed model is not the interpolated continuous-valued F0 contour but rather a sequence of discrete symbols, including quantized F0 levels and a symbol for the unvoiced condition. By using the discrete F0 symbols, the proposed model avoids the influence of artificially interpolated F0 curves. Experiments demonstrated that the proposed F0 model, which was trained using a dropout strategy, generated smooth F0 contours with relatively better perceived quality than those from baseline RNN models.

**Index Terms**: Text-to-speech, F0 model, recurrent neural network

## 1. Introduction

Fundamental frequency (F0) is an essential acoustic feature that realizes the speech prosody. Generating the F0 contour on the basis of textual features of the input text is necessary in many text-to-speech (TTS) synthesis systems. A TTS system may either transplant the generated F0 contour to a waveform [1] or convert the F0 contour and other spectral features into a speech waveform [2]. Even the WaveNet system, which directly generates waveform sampling points, gains from using externally generated F0 as additional input [3].

Many F0 modeling methods for TTS map the textual features to the commands that drive a hardwired models to generate the F0 contour [4, 5, 6, 7]. Some other methods use the unit-selection approach and generate F0 contours by concatenating F0 units [1, 8]. Alternatively, F0 modeling can use statistical models such as the hidden Markov model (HMM) [2, 9, 10, 11] and neural networks (NNs) [12, 13, 14, 15] to directly convert the textual features to the F0 contour.

This work proposes an F0 model that uses recurrent neural networks (RNNs) and further considers the temporal correlation of F0 contours by adding feedback links from the output to the input of the RNN. Motivated by the WaveNet [3] and a model for hand-writing synthesis [16], the proposed F0 model feeds the F0 of the previous frame as additional input for the current frame. Above the frame level, the proposed model feeds back the F0 features aggregated over the phoneme and syllable. By using the multi-tier feedback F0 features, the proposed model is expected to learn the movement of F0 contours and thus is different from existing models using multi-tier F0 features but not feedback links [17, 18]. Another difference is that the output of the proposed model is not the commonly used interpolated F0 contour but rather a sequence of discrete F0 symbols representing the quantized F0 levels and the unvoiced condition. As far as we know, the quantized F0 has been used as the input information for prosodic labeling [19] and conventional F0 modeling [20] but not as the output of the F0 model. By using the discrete F0 symbols, the proposed model avoids the influence of the interpolated F0 curves.

For practical considerations, this work uses a data dropout training method to reduce the *exposure bias* [21] that hampers the proposed model. This method may also be applicable to any model with feedback links and normal input. Additionally, this work introduces a hierarchical softmax layer to model the discrete F0 data. Experiments demonstrated that the proposed model generated smooth F0 contours with relatively better perceived quality than the baseline RNN model.

In section 2 of this paper, the NN-based F0 models are discussed and then the proposed model is explained in detail. Section 3 shows the configuration and experiments on the proposed model. Section 4 further discusses the proposed model, and in section 5 we conclude with a brief summary.

## 2. Model Description

### 2.1. Baseline neural-network-based F0 models

This work focuses on F0 models that convert the sequence of textual features $\boldsymbol{x}_{1:T} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T]$ to an F0 contour $\boldsymbol{o}_{1:T} = [\boldsymbol{o}_1, \cdots, \boldsymbol{o}_T]$ in $T$ frames. A vector $\boldsymbol{o}_t \in \mathbb{R}^D$ is used to denote the F0 of one frame without loss of generality. Recent NN-based F0 models, or models that generate F0 and other acoustic features, have used RNNs with long-short-term memory units (LSTM) [15], mixture density networks (MDN) [22], and highway networks [23]. Despite the difference of NNs, most of these models can be written as

$$p(\boldsymbol{o}_{1:T}|\boldsymbol{x}_{1:T}) = \prod_{t=1}^{T} p(o_t|\text{NN}_t(\boldsymbol{x}_{1:T})). \quad (1)$$

Here, $\text{NN}_t(\boldsymbol{x}_{1:T})$ is used to denote the output of the NN at the frame $t$, although it may only depend on $\boldsymbol{x}_t$ or part of $\boldsymbol{x}_{1:T}$ for certain types of NN. The probability density function $p(o_t|\text{NN}_t(\boldsymbol{x}_{1:T}))$ can be a delta function $\delta(\boldsymbol{o}_t - \text{NN}_t(\boldsymbol{x}_{1:T}))$ or a Gaussian mixture model in the case of MDN. In either case, the weights of NN need to be trained. After the network training, an F0 contour can be generated as $\widehat{\boldsymbol{o}}_{1:T} = \arg\max_{\boldsymbol{o}_{1:T}} p(\boldsymbol{o}_{1:T}|\tilde{\boldsymbol{x}}_{1:T})$ for the new input $\tilde{\boldsymbol{x}}_{1:T}$.

### 2.2. Proposed F0 model

The proposed model is based on the RNN, but it does not assume the distribution of $\boldsymbol{o}_t$ to be independent across frames. Instead, it assumes that the distribution of $\boldsymbol{o}_t$ is conditioned by the F0 data of previous frames, which can be shown as

$$p(\boldsymbol{o}_{1:T}|\boldsymbol{x}_{1:T}) = \prod_{t=1}^{T} p(\boldsymbol{o}_t|\mathrm{NN}_t(\boldsymbol{x}_{1:T}, f(\boldsymbol{o}_{<t}))). \qquad (2)$$

Here, previous output data $\boldsymbol{o}_{<t}$ are summarized by a function $f(\cdot)$ and then fed back as the additional input to the network. With this data feedback, the model is expected to learn the movement of the F0 contour as a sequential model.

#### 2.2.1. *Multi-tier F0 feedback*

By setting $f(\boldsymbol{o}_{<t}) = \boldsymbol{o}_{t-1}$, the model uses the previous output for feedback and is expected to learn the temporal correlation of target data across frames [16, 24]. To capture the dependency of F0 across longer spans, the proposed model further feeds back F0 features summarized over the linguistic segments above the frame level. Suppose that the frame $t-1$ is in a phoneme that starts from the frame $t_{p_{(t-1)}}$. Then, the phoneme-level F0 feature can be computed as the moving average of the transformed F0 within that phoneme:

$$\boldsymbol{a}_{p_{(t-1)}} = \frac{1}{t - t_{p_{(t-1)}}} \sum_{i=t_{p_{(t-1)}}}^{t-1} \tanh(\boldsymbol{o}_i). \qquad (3)$$

Similarly, F0 features can be computed for syllables ($\boldsymbol{a}_{s_{(t-1)}}$) and other linguistic segments. By setting $f(\boldsymbol{o}_{<t}) = [\boldsymbol{o}_{t-1}^{\top}, \boldsymbol{a}_{p_{(t-1)}}^{\top}, \boldsymbol{a}_{s_{(t-1)}}^{\top}, \cdots]^{\top}$, the proposed model uses multi-tier F0 feedback features and is expected to learn the temporal dependency of F0 at multiple time scales.

The model with the 3-tier feedback (frame, phoneme, and syllable) is plotted in Figure 1. Note that the boundary of the linguistic segments is retrieved from the input textual features.

#### 2.2.2. *Clockwork RNN with a textual-feature-based clock table*

For better F0 modeling, input textural features can be processed with the long-term dependency taken into consideration. This is achieved by switching the normal RNN layer to the clockwork RNN layer [25, 26] with a clock table synchronized with the boundary of linguistic segments. A neuron in this clockwork layer only updates its output when the current frame is the start of a linguistic segment. Otherwise, it holds its previous output.

#### 2.2.3. *Quantized F0 modeling*

Baseline NN-based F0 models usually work on F0 contours where the unvoiced regions are interpolated by artificial F0 curves. However, the proposed model performs poorly on such F0 contours because it tends to fit the artificial F0 curves (see `BF` in table 3). One method is to remove the unvoiced frames, but it ignores the segmental [27] and super-segmental [28] correlation between the F0 contour and the unvoiced regions.

To directly model the un-interpolated contours, this work proposes to quantize the F0 value and then represent and model the F0 event of both voiced and unvoiced frames uniformly. Specifically, the raw F0 is first converted to the mel-scale, and then the mel-scale F0 is quantized into finite levels that cover the training data (see details in section 3.1). Suppose there were $N$ quantized F0 levels and one unvoiced symbol; accordingly, the target F0 event of one frame can be encoded as a one-hot vector $\boldsymbol{o}_t = [o_{t,0}, o_{t,1}, \cdots, o_{t,N}]$, where $o_{t,j} \in \{0, 1\}$ and $||\boldsymbol{o}_t||_1 = 1$. The probability for observing $\boldsymbol{o}_t$ can be calculated by using a hierarchical softmax layer [29] as

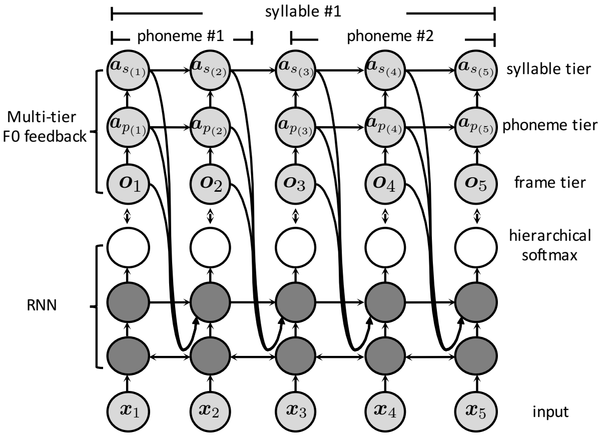

Figure 1: *The proposed F0 model with feedback from the frame, phoneme, and syllable-tier. Details are in section 3.1.*

$$P(\boldsymbol{o}_t|\boldsymbol{h}_t) = \begin{cases} \dfrac{e^{h_{t,0}}}{1+e^{h_{t,0}}}, & \text{if } o_{t,0} = 1 \\ \dfrac{1}{1+e^{h_{t,0}}} \dfrac{e^{h_{t,j}}}{\sum_{k=1}^{N} e^{h_{t,k}}}, & \text{if } o_{t,j} = 1, 1 \leq j \leq N \end{cases}, \qquad (4)$$

where $\boldsymbol{h_t} = \mathrm{NN}_t(\boldsymbol{x}_{1:T}, f(\boldsymbol{o}_{<t}))$ is defined to simplify the notation and $o_{t,0} = 1$ is used to denote the unvoiced event. The first level of the hierarchical softmax computes the probability for the unvoiced/voiced (U/V) condition, and the second level computes the probability for each quantized F0 level in an voiced frame. In the generation stage, $\widehat{o}_{t,0}$ is set to one if $\frac{e^{h_{t,0}}}{1+e^{h_{t,0}}} > 0.5$; otherwise, $\widehat{o}_{t,j^*}$ is set to one, where $j^* = \arg\max_j \frac{e^{h_{t,j}}}{\sum_{k=1}^{N} e^{h_{t,k}}}$ and $1 \leq j \leq N$.

F0 quantization is theoretically reasonable because humans cannot detect frequency change below the 'frequency difference limen' [30]. Another practical advantage is that F0 can be modeled without using either F0 interpolation or complicated statistical models [31, 32]. Note that this one-hot vector representation of F0 is also used for the multi-tier F0 feedback.

#### 2.2.4. *Strategies to alleviate the exposure bias*

The proposed model can be trained with the natural F0 as the feedback data. During generation, it has to use the generated F0 because the natural F0 is unknown. However, this combination of training and generation methods leads to *exposure bias* [21], and the model may accumulate generation errors. Furthermore, because the natural $\boldsymbol{o}_{t-1}$ for feedback and the target $\boldsymbol{o}_t$ have similar F0 values, the trained model may be addicted to the feedback data but insensitive to the input textual features.

A better training method might be *scheduled sampling* [24], which randomly uses either the ground truth $\boldsymbol{o}_{t-1}$ or the generated $\widehat{\boldsymbol{o}}_{t-1}$ for feedback in the training stage. However, this method is theoretically flawed [33], and our trials on it were unsuccessful (see results in section 3.2). Alternatively, this work proposes a *data dropout* strategy that enables the model to randomly set the feedback data to zero. If the feedback data are zero, the model is forced to learn and predict F0 events only on the basis of the textual features. With the data dropout, the

Table 1: *Results of pilot tests. Metrics (RMSE, CORR, U/V-ER, f-GV and $\Delta f$-O) are defined in section 3.1. Note that $P_s$ is the probability for data dropout or schedule sampling. For reference, GV of the natural F0 is 8.21.*

| | | Conventional softmax layer | | | Hierarchical softmax layer | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Model | Training strategy | RMSE | CORR | U/V-ER | RMSE | CORR | U/V-ER | $\Delta f$-O | $f$-GV |
| $\mathtt{QN}$ | - | 43.5 | 0.745 | 9.59% | 42.1 | 0.760 | 4.88% | 5.68% | 8.10 |
| $\mathtt{QF}_{\mathtt{FT}}$ | data dropout ($P_s = 0.75$) | 43.7 | 0.740 | 6.78% | 44.0 | 0.739 | 4.92% | 1.41% | 8.12 |
| $\mathtt{QF}_{\mathtt{FT}}$ | data dropout ($P_s = 0.50$) | 46.3 | 0.716 | 6.10% | 45.5 | 0.727 | 4.98% | 1.18% | 8.21 |
| $\mathtt{QF}_{\mathtt{FT}}$ | data dropout ($P_s = 0.25$) | 49.0 | 0.687 | 5.90% | 49.2 | 0.695 | 5.06% | 1.18% | 8.14 |
| $\mathtt{QF}_{\mathtt{FT}}$ | data dropout ($P_s = 0.00$) | 53.9 | 0.633 | 5.98% | 54.4 | 0.634 | 5.24% | 1.22% | 8.22 |
| $\mathtt{QF}_{\mathtt{FT}}$ | schedule sampling ($P_s = 0.50$) | - | - | - | 49.4 | 0.693 | 8.45% | - | - |

Table 2: *Experimental models*

| ID | Quantized F0 | Feedback frame tier | Feedback pho. & syl. tiers | Clock LSTM |
|---|---|---|---|---|
| $\mathtt{BN}$ | - | - | - | - |
| $\mathtt{BF}$ | - | + | - | - |
| $\mathtt{QN}$ | + | - | - | - |
| $\mathtt{QF}_{\mathtt{FT}}$ | + | + | - | - |
| $\mathtt{QF}_{\mathtt{AT}}$ | + | + | + | - |
| $\mathtt{QF}_{\mathtt{CL}}$ | + | + | + | + |

model is expected to be less addicted to the feedback data and more robust to the generation errors of previous frames.

In the generation stage, a better strategy is to use the probability vector calculated by the softmax layer in (4) rather than the one-hot vector $\widehat{\boldsymbol{o}}_t$ as the feedback data. Using $\widehat{\boldsymbol{o}}_t$ failed to generate acceptable F0 contours in our experiments.

## 3. Experiments

### 3.1. Data and model configuration

The Blizzard Challenge 2011 corpus of 12,072 English utterances [34] was used for the experiments. Both the test and validation set contained 500 randomly selected utterances. Text analysis on the entire corpus was conducted using the Flite toolkit [35]. The outputs of Flite were converted into vectors of order 382 as the input textual features. The CURRENNT library [36] was modified to implement the proposed model [1].

The F0 data were extracted using STRAIGHT [37] and converted into a mel-scale using $m = 1127 \log(1 + \text{F0}/700)$ [38]. The mel-scale F0 data were quantized into 127 levels between 133 and 571, which were the minimum and maximum values of the mel-scale F0 in the corpus. Then, the quantized F0 data and the unvoiced condition were encoded as the one-hot vectors $\boldsymbol{o}_t \in \{0, 1\}^{128}$. An analysis-synthesis test showed that using 127 levels was sufficient to avoid the perceptible 'quantization noise' on the corpus. For baseline models, F0 contours were interpolated by using an exponential function.

Experimental models are listed in Table 2. The baseline $\mathtt{BN}$ had two feed-forward layers and two bi-directional RNN LSTM layers with the layer size as $(512, 512, 256, 128)$. Its output layer was a linear transformation layer. Other quantized F0 models adopted a similar network structure but with a softmax output layer. Furthermore, models $\mathtt{QF}_{\mathtt{FT}}$ , $\mathtt{QF}_{\mathtt{AT}}$, and $\mathtt{QF}_{\mathtt{CL}}$ switched the second LSTM layer to a uni-directional one and took the feedback data as the input to that layer. This configuration was based on our founding that the feedback data should be processed by at least one LSTM layer. Model $\mathtt{QF}_{\mathtt{CL}}$ further replaced the first RNN LSTM layer with a clock-LSTM layer, where 16 and another 16 neurons were updated according to the phoneme and syllable boundary, respectively. For reference, a model $\mathtt{BF}$ with feedback links was trained to model the continuous-valued interpolated F0 data. Note that only $\mathtt{BN}$ and $\mathtt{BF}$ used the F0 delta and delta-delta features as well as the MLPG generation algorithm [39]. Natural alignment was used for generation, and the waveforms were synthesized given the output F0 from experimental models and spectral features from another RNN model.

For objective evaluation, the root mean square error (RMSE), correlation coefficients (CORR) and unvoiced/voiced error rate (U/V-ER) were calculated against the continuous-valued natural F0 data. The global variance of F0 in Hz domain ($f$-GV) was also computed. Finally, delta F0 outliers ($\Delta f$-O) was counted to measure the amount of unnatural jump of the F0 contours. The delta F0 ($\Delta f$) was the difference between the F0 values of two adjacent voiced frames. Given the mean ($m_{\Delta f}$) and standard deviation ($\sigma_{\Delta f}$) of the natural $\Delta f$, the outlier of the generated $\Delta f$ was identified if it lay outside $m_{\Delta f} \pm 3\sigma_{\Delta f}$. The percentage of outlier in the test set was defined as $\Delta f$-O.

### 3.2. Pilot tests on the proposed model

Pilot tests were conducted to compare the effect of different softmax layer types and training strategies, and the results are listed in Table 1. Each row of Table 1 compares the hierarchical softmax against and the conventional plain softmax. Although the hierarchical one did not consistently improve the F0 RMSE and CORR, it decreased U/V-ER in all cases. One reason may be that the hierarchical softmax layer is more suitable to model the unbalanced distribution of the F0 events, where the number of the unvoiced event is much larger than any other F0 event.

As for training strategies, comparison across the rows of Table 1 reveals that the model performed better if the data dropout was more frequently used. As section 2.2.4 argues, the reason may be that more data dropout makes the model less addicted to the ground-truth feedback data and thus more robust to the 'errors' from generated feedback data. However, the model with more frequent dropout generated F0 contours with more unnatural transitions (see $\mathtt{QN}$ in Figure 2) and thus acquired larger $\Delta f$-O. On the contrary, $\mathtt{QF}_{\mathtt{FT}}$ without data dropout generated smooth F0 contours, but it tends to ignore the textual features and generated incorrect F0 in the linguistic sense, e.g., pitch accents on unstressed syllables. Our informal subjective test suggested that the data dropout with $P_s = 0.5$ was appropriate to strike the balance. Note that the scheduled sampling failed to decrease the U/V error rate, even though it improved the F0 CORR and RMSE.

### 3.3. Evaluation on the baseline and proposed models

On the basis of the pilot tests, $\mathtt{QF}_{\mathtt{AT}}$, $\mathtt{QF}_{\mathtt{CL}}$ and $\mathtt{BF}$ were trained with data dropout ($P_s = 0.5$). Objective and subjective results are listed in Table 3 and 4. The subjective preference test was participated by 29 native English speakers from universities, where each participant evaluated 14 pairs of samples.

[1] Tthe toolkit and samples are available at http://tonywangx.github.io

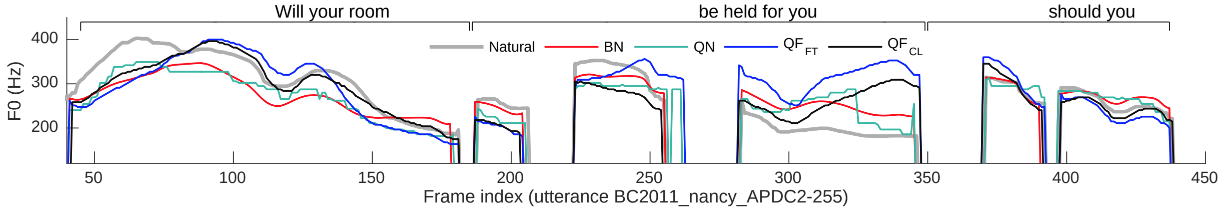

Figure 2: *Generated F0 for utterance '**Will your room be held for you should you** require a hospital or nursing home stay, ...'. Note that $QF_{FT}$ and $QF_{CL}$ were trained with data dropout ($P_s = 0.5$).*

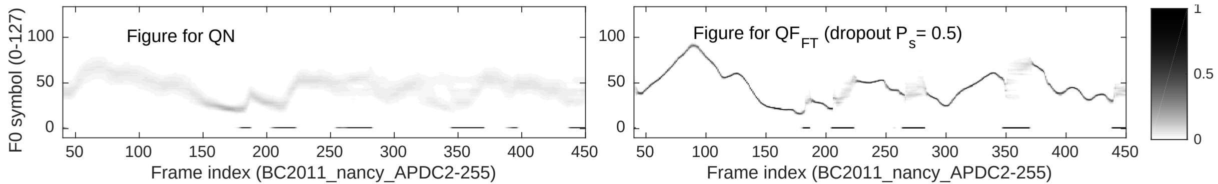

Figure 3: *Probability $P(o_{t,j}|\boldsymbol{h}_t)$ of each F0 symbol for each frame in a test utterance. The 0th event is for 'being unvoiced'.*

Table 3: *Results of the objective evaluation for section 3.3. Quantized F0 (Quan. F0) is shown for reference.*

| | RMSE | CORR | U/V-ER | $\Delta f$-O | $f$-GV |
|---|---|---|---|---|---|
| BN | 39.4 | 0.775 | 5.01% | 0.04% | 7.77 |
| QN | 42.1 | 0.760 | 4.88% | 5.68% | 8.10 |
| $\texttt{QF}_{\texttt{FT}}$ | 45.5 | 0.727 | 4.98% | 1.18% | 8.14 |
| $\texttt{QF}_{\texttt{AT}}$ | 43.7 | 0.744 | 4.91% | 1.11% | 8.17 |
| $\texttt{QF}_{\texttt{CL}}$ | 43.3 | 0.744 | 4.82% | 1.18% | 8.15 |
| BF | 79.3 | 0.211 | 9.38% | 0.19% | 8.32 |
| Quan. F0 | 1.1 | 0.999 | 0.00% | 0.06% | 8.21 |

First, BF performed the worst objectively. It was found that the F0 contours generated by BF repeated patterns similar to the interpolated F0 curves. As the proportion of unvoiced frames in the corpus is around 30%, BF may be biased to the interpolated curves. This result indicates that the F0 interpolation should be avoided for the model using feedback links.

The comparison among $\texttt{QF}_{\texttt{FT}}$ , $\texttt{QF}_{\texttt{AT}}$ , and $\texttt{QF}_{\texttt{CL}}$ indicates the usefulness of the multi-tier F0 feedback and clock LSTM in the objective test. However, these three models acquired higher RMSE and lower CORR than QN and BN. As previously argued, this gap is due to propagation of the 'errors' in the generation stage through the feedback link. A simple experiment on $\texttt{QF}_{\texttt{FT}}$ showed that, if the F0 of the test set was used for feedback during generation, the generated F0 was almost identical to the natural one. In other words, if the F0 data generated previously were close to the natural data, the F0 output at the current frame could also be accurate. On the other hand, any discrepancy from ground truth will affect the generation in the following frames.

Despite the worse result on F0 RMSE and CORR, QF was preferred to QN in the subjective evaluation. The non-sequential model QN just counts the F0 event conditioned by the textual features. Because of the ambiguous association between the text and F0, QN only learned vague distribution of F0 events for unknown text, which can be seen in figure 3. As the result, QN generated F0 contours with more unnatural transitions as the contour in figure 2 and the large $\Delta f$-O in table 3 indicate. QF as well as $\texttt{QF}_{\texttt{AT}}$ and $\texttt{QF}_{\texttt{CL}}$ considered the temporal dependency of F0 contours. Therefore, the probability mass for F0 event from these models was sharp and the peak moved gradually across frames, which made the generated F0 contours smooth.

As for the comparison with BN, $\texttt{QF}_{\texttt{CL}}$ was preferred in the subjective test. Similar to QN, the non-sequential model BN learns the statistics of F0 data conditioned by the textual features. It achieved better F0 RMSE and CORR because its output was close to the mean F0. However, this modeling and generation framework leads to over-smoothed F0, which can be shown by the lower $f$-GV of BN. The second reason is that the ground truth F0 to be used for calculating RMSE and CORR is just one possible realization of F0 for the test text, and an F0 contour different from the ground truth may also be suitable. The proposed model may generate such F0 contours. One example can be seen in Figure 2, where $\texttt{QF}_{\texttt{CL}}$ generated a rising F0 curve for the words 'for you'. Because of the two possible reasons above, the F0 contour given by the $\texttt{QF}_{\texttt{CL}}$ may be preferred to the output from the baseline RNN model.

Table 4: *Results of subjective evaluation for section 3.3. The p-value of the two-tailed sign test is shown in the last column.*

| BN | QN | $\texttt{QF}_{\texttt{FT}}$ | $\texttt{QF}_{\texttt{AT}}$ | $\texttt{QF}_{\texttt{CL}}$ | $p$ |
|---|---|---|---|---|---|
| | 33.7% | 66.3% | | | 0.001 |
| | | 60.0% | 40.0% | | 0.119 |
| | | | 47.1% | 52.9% | 0.720 |
| 36.7% | | | | 63.3% | 0.011 |

## 4. Discussion

This work did not compare all possible configurations of the network structure, multi-tier F0 strategies, and different tiers of linguistic segments, and some of the tests are still underway. As this work focuses on the sequential modeling of the F0 contour, future work will further investigate one training method that alleviates the exposure bias of sequential modeling [40] and a differential softmax layer [41] to handle the unbalanced distribution of the quantized F0 events.

## 5. Conclusion

This work introduced an RNN-based sequential F0 model for the TTS task. This model differs from the baseline model as it feeds the previous F0 data as additional input of the current frame. These feedback F0 data not only include the F0 from the previous frame but also F0 features aggregated over multiple linguistic tiers. Other ideas put forth in this work include the use of discrete F0 data as the target of F0 modeling and a data dropout training strategy. As the experiments indicated, using the discrete F0 symbols is effective for sequential F0 modeling as it avoids the F0 interpolation. On the basis of discrete F0 symbols and the dropout training strategy, the proposed F0 model can generate F0 contours with somehow natural shapes and relatively better perceived quality.

# 6. References

[1] A. Raux and A. W. Black, "A unit selection approach to F0 modeling and its application to emphasis," in *Proc. ASRU*, 2003, pp. 700–705.

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

[3] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[4] H. Fujisaki and K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," *Journal of the Acoustical Society of Japan (E)*, vol. 5, no. 4, pp. 233–242, 1984.

[5] K. E. Dusterhoff, A. W. Black, and P. A. Taylor, "Using decision trees within the Tilt intonation model to predict F0 contours." *Proc. Eurospeech*, pp. 1627–1630, 1999.

[6] 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, 2009.

[7] X. Sun, "F0 generation for speech synthesis using a multi-tier approach," in *Proc. ICSLP*, 2002, pp. 2077–2080.

[8] J. Meron, "Prosodic unit selection using an imitation speech database," in *Proc. ITRW*, 2001, pp. 113–116.

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

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

[11] C. Wang, Z. Ling, B. Zhang, and L. Dai, "Multi-layer F0 modeling for HMM-based speech synthesis," in *Proc. ISCSLP*. IEEE, 2008, pp. 1–4.

[12] Y. Sagisaka, "On the prediction of global F0 shape for Japanese text-to-speech," in *Proc. ICASSP*, 1990, pp. 325–328.

[13] C. Traber, "F0 generation with a data base of natural F0 patterns and with a neural network," in *Proc. ESCA Workshop on Speech Synthesis*, 1991, pp. 141–144.

[14] S. H. Chen, S. H. Hwang, and Y. R. Wang, "An RNN-based prosodic information synthesizer for Mandarin text-to-speech," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 3, pp. 226–239, 1998.

[15] 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.

[16] A. Graves, "Generating sequences with recurrent neural networks," *arXiv preprint arXiv:1308.0850*, 2013.

[17] 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.

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

[19] A. Rosenberg, "Classification of prosodic events using quantized contour modeling," in *Proc. ACL-HLT*, 2010, pp. 721–724.

[20] T. Nose, K. Ooki, and T. Kobayashi, "HMM-based speech synthesis with unsupervised labeling of accentual context based on F0 quantization and average voice model," in *Proc. ICASSP*, 2010, pp. 4622–4625.

[21] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba, "Sequence level training with recurrent neural networks," *Proc. ICLR*, 2016. [Online]. Available: https://arxiv.org/pdf/1511.06732.pdf

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

[23] X. Wang, S. Takaki, and J. Yamagishi, "Investigating very deep highway networks for parametric speech synthesis," in *Proc. SSW9*, 2016, pp. 181–186.

[24] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer, "Scheduled sampling for sequence prediction with recurrent neural networks," in *Proc. NIPS*, 2015, pp. 1171–1179.

[25] J. Koutnik, K. Greff, F. Gomez, and J. Schmidhuber, "A Clockwork RNN," in *Proc. ICML*, 2014, pp. 1863–1871.

[26] P. Liu, X. Qiu, X. Chen, S. Wu, and X. Huang, "Multi-timescale long short-term memory neural network for modeling sentences and documents," in *Proc. EMNLP*, 2015, pp. 2326–2335.

[27] K. E. A. Silverman, "The Structure and Processing of Fundamental Frequency Contours," Ph.D. dissertation, University of Cambridge, 1987.

[28] M. E. Beckman and G. Ayers, "Guidelines for ToBI labelling," *The OSU Research Foundation*, vol. 3, 1997. [Online]. Available: http://www.ling.ohio-state.edu//research/phonetics/E_ToBI/singer_tobi.html

[29] F. Morin and Y. Bengio, "Hierarchical probabilistic neural network language model," in *Proc. AISTATS*, vol. 5, 2005, pp. 246–252.

[30] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 6th ed. Brill, 2012.

[31] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Multi-space probability distribution HMM," *IEICE Trans. on Information and Systems*, vol. 85, no. 3, pp. 455–464, 2002.

[32] J. Latorre, M. J. F. Gales, K. Knill, and M. Akamine, "Training a supra-segmental parametric F0 model without interpolating F0," in *Proc. ICASSP*, 2013, pp. 6880–6884.

[33] F. Huszár, "How (not) to train your generative model: Scheduled sampling, likelihood, adversary?" *arXiv preprint arXiv:1511.05101*, 2015.

[34] S. King and V. Karaiskos, "The Blizzard Challenge 2011," in *Proc. Blizzard Challenge Workshop*, 2011, pp. 1–10.

[35] HTS Working Group, "The English TTS system Flite+HTS_engine," 2014. [Online]. Available: http://hts-engine.sourceforge.net/

[36] F. Weninger, J. Bergmann, and B. Schuller, "Introducing CURRENT: The Munich open-source CUDA recurrent neural network toolkit," *The Journal of Machine Learning Research*, vol. 16, no. 1, pp. 547–551, 2015.

[37] H. Kawahara, I. Masuda-Katsuse, and A. d. 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 Communication*, vol. 27, pp. 187–207, 1999.

[38] D. O'Shaughnessy, *Speech Communications: Human and Machine*. Institute of Electrical and Electronics Engineers, 2000.

[39] T. Keiichi, , Y. Takayoshi, M. Takashi, K. Takao, and K. Tadashi, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP*, 2000, pp. 936–939.

[40] S. Ross, G. J. Gordon, and D. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning." in *Proc. AISTATS*, vol. 1, no. 2, 2011, pp. 627–635.

[41] W. Chen, D. Grangier, and M. Auli, "Strategies for training large vocabulary neural language models," in *Proc. ACL*, 2016, pp. 1975–1985.