Investigating Very Deep Highway Networks for Parametric Speech Synthesis

Xin Wang\textsuperscript{a,b,c}, Shinji Takaki\textsuperscript{b}, Junichi Yamagishi\textsuperscript{a,b,c}

\textsuperscript{a}National Institute of Informatics, 2-1-2, Hitotsubashi-cho, Chiyoda-ku, Tokyo, 101-8430, Japan.
\textsuperscript{b}SOKENDAI, 2-1-2, Hitotsubashi, Chiyoda, Tokyo, 101-8430, Japan.
\textsuperscript{c}The Centre for Speech Technology Research (CSTR), University of Edinburgh, Edinburgh, EH8 9LW, United Kingdom.

Abstract

Deep neural networks are powerful tools for classification and regression tasks. While a network with more than 100 hidden layers has been reported for image classification, how such a non-recurrent neural network with more than 10 hidden layers will perform for speech synthesis is as yet unknown. This work investigates the performance of deep networks on statistical parametric speech synthesis, particularly the question of whether different acoustic features can be better generated by a deeper network. To answer this question, this work examines a multi-stream highway network that separately generates spectral and F0 acoustic features based on the highway architecture. Experiments on the Blizzard Challenge 2011 corpus show that the accuracy of the generated spectral features consistently improves as the depth of the network increases from 2 to 40, but the F0 trajectory can be generated equally well by either a deep or a shallow network. Additional experiments on a single-stream highway and normal feedforward network, both of which generate spectral and F0 features from a single network, show that these networks must be deep enough to generate both kinds of acoustic features well. The difference in the performance of multi- and single-stream highway networks is further analyzed on the basis of the networks’ activation and sensitivity to input features. In general, the highway network with more than 10 hidden layers, either multi- or single-stream, performs better on the experimental corpus than does a shallow network.

Keywords: Text-to-Speech, Statistical parametric speech synthesis, Deep neural network, Highway neural network

1. Introduction

Speech synthesis aims at creating natural-sounding speech waveforms and is used in various types of application with speech waveforms as output. A widely used application is Text-to-Speech (TTS) synthesis [1], where the speech is synthesized to read aloud the input text. Its social value is obvious in human-machine and human-human communication, e.g., when a disabled human speaker cannot articulate sounds.

TTS is difficult because of the ambiguous association between text and speech. Despite the recent trend towards end-to-end TTS [2], most existing TTS systems consist of front-end back-ends. The front-end infers the phonemic and prosodic information from the text, and then the back-end synthesizes the speech waveform given the output of the front-end. The TTS back-end, or the acoustic model, can be implemented on the basis of statistical parametric speech synthesis (SPSS), a framework that uses statistical models such as a hidden Markov model (HMM) to generate speech acoustic features and then construct the waveform [3, 4]. Recently, various acoustic models based on neural networks (NNs) have been proposed to augment the HMM-based SPSS framework [5, 6, 7]. One reason is that well designed NNs can model various aspects of speech that have been ignored by HMM, including the complex correlation among linguistic features [6], cross-time [8, 9], and cross-dimension dependency [10] of speech acoustic features. Additionally, NNs can better describe the distribution of high-dimensional acoustic features [2, 11, 12] or waveform [13, 14], which may alleviate the artefact due to vocoding.

Despite the progress of NN-based SPSS, some questions remain unaddressed. One such question is how a NN’s depth affects its performance on acoustic feature modeling, particularly on the commonly used low-dimensional spectral and F0 features. We try to answer this question in the context of non-recurrent NNs. Although theories [15] and research works in image classification [16] have shown that very deep NNs can perform better than shallow networks, SPSS may need specific experiments because it is a regression task with heterogeneous target features. It would be useful to know whether both F0 and spectral features can be better modeled by a deeper NN and whether the classical feedforward NN should be used as it is.

To explore the influence of network depth, this work conducts experiments on the multi-stream highway network because it facilitates the training of very deep networks [17] and reduces the interaction between sub-networks for spectral and F0 modeling [18]. Experiments comparing the network with different depths show that spectral features can be better generated by a network with up to 40 hidden layers, while F0 can be well generated by a network with just 4 hidden layers. Then, this work conduct experiments on single-stream highway and conventional feedforward networks, where all acoustic features are generated from a single network. It was...
found that the single-stream networks must be sufficiently deep to generate both spectral and F0 features well. These findings are further supported by a histogram analysis of each network’s activation and the neurons’ sensitivity to the input textual features. These results indicate that the commonly used single-stream feedforward network should be sufficiently deep to well model both spectral and F0 features. Comparing the single- and multi-stream networks in a subjective test also shows that the performance of all the experimental networks is better when the depth is increased from 4 to 40.

This work is closely related to the pioneering work using feedforward NN for SPSS, where networks with up to 7 hidden layers were reported [6, 19]. However, this work explores deeper networks. More importantly, based on the experiments on single- and multi-stream networks, this work provides results and analysis on the different influence of network depth on spectral and F0 modeling. This result may not be revealed by only using the normal single-stream networks. We think it may be informative for the widely used NN-based SPSS because neither the method or result of this work was reported before.

Section 2 of this paper introduces the multi- and single-stream highway networks. Section 3 then introduces the tools to analyze the neural network. Section 4 explains the details of the experiments, including the performance of highway networks with different depths when modeling spectral and F0 features. The trained networks are further analyzed in section 5 based on the tools introduced in section 3. Section 6 discusses the remaining issues, and section 7 concludes.

2. Acoustic model based on highway networks

2.1. Neural-network-based acoustic model for TTS

This work focuses on the neural-network-based acoustic model for TTS systems with a pipeline structure [1]. This type of TTS system uses a front-end to derive the information on the pronunciation and prosody of the input text. Based on the linguistic features, the SPSS-based back-end generates the waveform. For the common configuration in English TTS, the prosodic features may include the pitch accent of syllables and boundary tone of phrases, and the acoustic features may consist of F0 and compact spectral features of each speech frame [20]. The acoustic model can be implemented on the basis of feedforward neural networks, where the input linguistic features are transformed into acoustic features frame by frame [6].

An advantage of a deep NN is its ability to extract structural features from data [15, 21, 22]. For example, a non-recurrent network with more than 100 hidden layers performed best in a recent image classification task [16]. For acoustic modeling, existing work has explored feedforward neural networks with up to 7 hidden layers [6, 19]. Although this showed that a deeper network improved the accuracy of generated acoustic features, the performance of a network with more than 10 hidden layers remains unknown. This work thus aims to investigate the performance of deeper non-recurrent networks by analyzing their performance and behavior.

2.2. Highway-network-based acoustic model

A deep neural network cannot be easily trained using the back-propagation algorithm and the random initialization strategy. Various methods have been proposed to facilitate the training of deep neural networks, including using better initialization methods based on unsupervised pre-training [23, 24] and unsaturated activation functions [25, 26]. Another thread of research focuses on the network architecture that alleviates the difficulty of network training. One new architecture is the highway network [17], which combats the gradient-vanishing problem. It has been shown that a deep highway network with more than 10 hidden layers can be well trained for speech recognition without requiring complicated engineering tunings [27]. Therefore, this work investigates deep highway networks for parametric speech synthesis because of their simplicity and effectiveness.

2.2.1. Highway block

The highway network consists of one or multiple highway blocks. In one highway block, the input linguistic feature vector \( x \) is transformed by a conventional feedforward layer as

\[
\mathcal{H}(x) = f(W_H x + b_H).
\]  

(1)

Here, \( f(\cdot) \) is the non-linear activation function, \( b_H \) is the bias vector and \( W_H \) is the transformation matrix. Furthermore, the highway block uses a highway gate to compute a control vector

\[
\mathcal{T}(x) = \sigma(W_T x + b_T),
\]  

(2)

and then merges the transformed feature vector \( \mathcal{H}(x) \) with the input \( x \) as the output of this highway block:

\[
y = \mathcal{T}(x) \odot \mathcal{H}(x) + [1 - \mathcal{T}(x)] \odot x.
\]  

(3)

Here, \( \odot \) denotes element-wise multiplication, and the sigmoid function \( \sigma(x) = \frac{1}{1 + e^{-x}} \) is used in the highway gate. The above computational flow in one highway block is plotted in Figure 1.

Parameters \( W_T \) and \( b_T \) in the highway gate are also trainable. When the output of the gate \( \mathcal{T}(x) \) is approximately zero, the input \( x \) can be directly propagated forwards; i.e., \( y = x \). In this case, the gradient can also be propagated backwards without being attenuated by the feedforward transformation layer in the highway block. Thus, a very deep network based on highway blocks can be trained by using the standard gradient-descent back-propagation algorithm. Note that \( \mathcal{H}(x) \) can be a transformation conducted by multiple feedforward layers. In other words, one highway block can contain more than one feedforward transformation layer.

2.2.2. Single-stream and multi-stream highway networks

A highway network based on one or more highway blocks can be directly used as the acoustic model for TTS, which is shown on the right side of Figure 1. Because all the acoustic features are generated by a single network, we call it a single-stream highway network. The word ‘stream’ is borrowed from the HMM-based parametric framework [28]. Besides the single-stream architecture, we propose a multi-stream highway
network as shown in the middle of Figure 1. On the input side, a linear projection layer transforms the input vector into a shared hidden vector. The multi-stream highway network then uses several sub-networks to separately transform the shared vector into different acoustic features.

There are two reasons for us to investigate multi-stream networks. First, the hidden feature vector in a single-stream network encodes the information for generating both spectral and F0 features, so the effect of the network’s depth on the spectral and F0 features cannot be separately examined. In addition, the hidden vector may be biased toward high-dimensional spectral features and affect the performance of F0 generation. Thus, a multi-stream architecture is investigated to clearly show the result for each acoustic feature type.

Second, we want to compare the change of performance when the depth of single- and multi-stream networks is increased. Despite the claim that a single-stream network can model the correlation between spectral and F0 features [29], other studies have shown that the correlation between F0 and spectral features is somewhat weak, at least for reading speech in a neutral style [30, 31]. Recent work also shows that the F0 generation performance becomes worse even when spectral features are better generated on the basis of multi-task learning [32]. In addition, spectral and F0 features may rely on different input textual features [33], thus on different hidden representations in the network. Considering the above argument, we wonder that multi- and single-stream highway networks may perform differently on F0 and spectral feature generation as the depth is increased.

3. Tools to analyze the highway network

3.1. Histogram of the output of the highway gate

Each network’s performance can be evaluated by calculating the accuracy of the generated acoustic features. In addition, we introduce two tools to analyze the highway network. The first tool is the histogram of the output of the highway gate; i.e., $\mathbf{T}(\mathbf{x})$ in Equation 2. As discussed in section 2.2.1, a highway block tends to avoid using the transformed $\mathbf{H}(\mathbf{x})$ as its output when the highway gate is almost closed ($\mathbf{T}(\mathbf{x}) \approx 0$). Thus, the histogram on $\mathbf{T}(\mathbf{x})$ can reflect the behavior of the highway block and thus the whole network. In practice, we activate the network with the input data for one phoneme and collect $\mathbf{T}(\mathbf{x})$ to plot the histogram. Note that the order of the feature dimension is ignored.

3.2. Sensitivity of a neuron to input textual features

The second tool evaluates the sensitivity of a neuron to the input textual feature. Suppose the input feature vector at time $t$ takes the value $s$ for a feature class $S$. For example, when $S = \{/a/, /f/, \cdots\}$ refers to the phoneme identity, $s = /a/$ means this frame realizes the phoneme /a/. Then, all input vectors that take the same feature value $s$ can be fed to the neural network, and the output of the $k$-th neuron can be summarized as

\[ a_k(s) = \frac{\exp(\tilde{a}_k(s))}{\sum_{s \in S} \exp(\tilde{a}_k(s))}, \]

where

\[ \tilde{a}_k(s) = \sum_t \gamma_t(s) h_k(t). \]

Here, $\gamma_t(s)$ is an indicator whose value is 1 if the input feature at time $t$ takes the feature value $s$. On the basis of $a_k(s)$, the $k$-th neuron’s sensitivity to the feature class $S$ is defined as a normalized entropy over all possible feature values as

\[ E_{S,k} = -\frac{\sum_{s \in S} a_k(s) \log a_k(s)}{-\sum_{s \in S} \frac{a_k(s)}{|S|} \log \frac{1}{|S|}}, \]

where $|S|$ is the number of possible values in class $S$. Note that if $S$ refers to a continuous feature such as syllable position, $s$ is an index of a quantized interval. Finally, the average sensitivity of a neuron group $K$ can be defined as $E_{\overline{S}} = \frac{1}{|K|} \sum_{k \in K} E_{S,k}$.

Note that $E_{S,k}$ is maximized only when $a_k(s) = \frac{1}{|S|}$, i.e., when the neuron’s output $h_k(t)$ is constant for any $s \in S$. In this case, the neuron is judged to be insensitive to that feature class $S$. On the contrary, a low $E_{S,k}$ indicates a high sensitivity. This sensitivity measure was originally defined for NNs in speech recognition [34]. The difference is that $\tilde{a}(s)$ in Equation 4 is normalized so that the entropy of unbalanced feature classes can be compared with each other.
4. Experiments

4.1. Corpus and network notation

Experiments used the Blizzard Challenge 2011 Nancy corpus that has 12072 English utterances [35]. Both the test and validation set contained 500 randomly selected utterances. Mel-generalized cepstral coefficients (MGCs) of order 60, continuous F0 trajectory, voiced/unvoiced (V/U) condition, and band aperiodicity (BAP) of order 25 were extracted for each speech frame by using the STRAIGHT vocoder [36]. The Flite toolkit [37] conducted the text-analysis for the entire corpus. The output of Flite were converted into a vector of order 382 as the input \( \mathbf{x} \) to the neural network. This vector encodes common textual features similar to those used in the HMM-based framework [20]. Experiments were conducted on the three types of neural network listed in Table 1. The toolkit for training the neural network was modified on the basis of the CURRENNT library [38].

<table>
<thead>
<tr>
<th>Definition</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>DS</td>
<td>Single-stream deep feedforward network</td>
</tr>
<tr>
<td>HS</td>
<td>Single-stream highway network</td>
</tr>
<tr>
<td>HM</td>
<td>Multi-stream highway network</td>
</tr>
</tbody>
</table>

4.2. Experiments on the depth of the neural network

4.2.1. Network configuration and training recipe

This experiment investigated how the depth of a multi-stream highway network influences its ability to generate different acoustic features. For reference, the single-stream highway and feedforward network were also examined.

All the highway blocks in the single-stream highway network (HS) had a layer size of 382, equal to the dimension of the input vector. Each highway block contained 2 hidden layers based on a \( \tanh \) activation function. The bias of the highway gate was initialized as -1.5 according to the results of a preliminary experiment comparing -1.5, 0.0 and 1.5. The other parameters were initialized using the normalized initialization strategy [39]. The single-stream feedforward network DS had a layer size of 382 and a \( \tanh \) function for all layers. Its model parameters were initialized using the normalized initialization strategy.

The multi-stream highway network HM contained three sub-networks for MGC, F0, and BAP. The layer size of each sub-network was 256. The layer size of the linear projection layer near the input side was 768 (\( \approx 256 \times 3 \)). Each sub-network was configured and initialized in the same way as the HS network.

These three networks were trained with the number of \( \tanh \)-based hidden layers set to \( \{2, 4, 8, 14, 20, 40\} \). The stochastic gradient descent method with early stopping was used for training. Batch normalization was not used [40] as our initial experiment showed that its effect is limited when the normalized initialization is used. The natural alignment on the test data was used in the objective evaluation. The objective measure was calculated on the test data for the last five training epochs of two training trials. The mean and standard deviation of each objective measure were then computed.

4.2.2. Results and analysis

The results are shown in Figure 2. Our first observation is that the MGC RMSE decreased when the depth of HM was increased from 2 to 40. The MGC RMSE curves of the other two experimental networks also decreased similarly. These results suggest that both single- and multi-stream networks are better at generating spectral features if they are deeper.

\(^{2}\)Our previous work [18] used Gaussian noise to initialize the highway network. However, we found that the normalized initialization strategy [39] led to a consistently better performing highway and feedforward networks.
However, the error curves on F0 showed different patterns. While the F0 curves of HS and DS start from a low point and gradually rise with increased network depth, the F0 curve of HM is almost flat. Particularly, HM with just 2 hidden tanh layers performed as well as the HM with 40 hidden layers in terms of F0. These results suggest that, for a network using HM architecture, more hidden layers is better for generating spectral features but not necessarily for F0.

Note that HM had a larger total layer size than HS and DS, which may influence HM’s performance on F0 especially when the network was shallow. This possibility is examined in the next section. Nevertheless, the shape of the RMSE curves generally shows how the F0 and spectral feature generation performs differently in the multi-stream network.

4.3. Experiments on the layer size of the neural network

4.3.1. Network configuration and training recipe

This experiment investigated network performance given a fixed depth but different layer sizes. The motivation is to compare different networks with a roughly same number of model parameters. All the neural networks in Table 1 were trained with 14 hidden layers. The depth of 14 was selected because a deeper network with a large layer size required too much GPU memory resource because of the current software implementation. For HS and DS, the trained networks in Section 4.2 with the layer size 382 were included for comparison. Meanwhile, HSs with layer sizes of 482, 582, 782 and 1024 were trained. Then, DSs with layer sizes of 782, 882 and 1024 were trained so that they were comparable with the HSs in terms of model size. The HMs used the configurations shown in Table 2. The training recipe was the same as in Section 4.2.

4.3.2. Results and analysis

According to the results shown in Figure 3, HM’s F0 performance did not change much when the layer size of the F0 sub-network changed. In contrast, HS and DS performed better on F0 as their layer size grew larger. On MGC, all three types of network performed better with a larger layer size.

These results are consistent with those in Section 4.2. For the multi-stream network, a larger network is better at spectral feature modeling than a smaller network, but it might not surpass a shallower network on F0 modeling. A larger single-stream network performs better on both acoustic features. Particularly, its F0 generation performance is similar to that of the multi-stream network. A single-stream network might be dominated by the hidden features for the spectral features. If that is the case, only in a very large network would some of the nodes in the single-stream network be assigned to model the F0. This possibility will be further discussed in Section 5.1.

4.4. Subjective evaluation

The objective evaluation results showed that spectral features can be better modeled by a very deep network while this is not the case for F0. However, it is of interest to know whether the difference is perceptible. Thus, before analyzing the network in detail, we describe a subjective evaluation in this section. We prepared 5 groups of synthetic samples in Table 3 and organized a MUSHRA test [41] in CSTR of the University of Edinburgh. Ten paid native English speakers participated in the test. Among the experimental groups, a comparison of H1, H2, and H3 was used to evaluate the perceptible difference in the generated spectral features, while a comparison of H1, H4, and H5 was used to show the difference in the generated F0s.

Table 2: Network structure of HM networks in Figure 3.

<table>
<thead>
<tr>
<th>Layer size of the sub-network</th>
<th>MGC stream</th>
<th>F0 stream</th>
<th>BAP stream</th>
</tr>
</thead>
<tbody>
<tr>
<td>HM1</td>
<td>256</td>
<td>256</td>
<td>256</td>
</tr>
<tr>
<td>HM2</td>
<td>382</td>
<td>256</td>
<td>256</td>
</tr>
<tr>
<td>HM3</td>
<td>512</td>
<td>382</td>
<td>256</td>
</tr>
<tr>
<td>HM4</td>
<td>768</td>
<td>512</td>
<td>256</td>
</tr>
</tbody>
</table>

Figure 3: Performance of the feedforward (DS), the multi- (HM) and single-stream highway network (HS) on the test set. Each network has 14 hidden tanh layers. Numbers on the curve of HS and DS denote the layer size. Table 2 defines HM1 to HM4. The horizontal axis is the number of model parameters (log scale).
The results are shown in Figure 4. Regarding the spectral features, $H_3$ achieved the best performance, followed by $H_1$ and then $H_2$. A two-sided t-test showed that the average score of $H_3$ was statistically higher than that of $H_1$ ($p \approx 0.000$) and $H_2$ ($p \approx 0.000$). Meanwhile, $H_1$’s average score was higher than that of $H_2$ ($p = 0.049$). These results indicated that a deeper $H_1$ network generated spectral features that were perceived to be better in quality.

Regarding FO among the three groups with different configurations, $H_1$ had a lower score than $H_4$ ($p = 0.003$), even though $H_1$ used FO generated by a network with 10 more hidden layers. Similarly, $H_5$ was not significantly different from $H_4$ ($p = 0.088$) even though $H_5$ used FO from the deepest network with 40 hidden layers. Interestingly, although the objective evaluation of FO in Figure 2 did not show huge differences among networks with different depths, the subjective evaluation showed a perceptible difference between $H_1$ with 14 hidden layers and the other cases. This could be because different models generate quite different FO trajectories even though the average RMSE and CORR metrics on the generated FO trajectories are not so different. However, this result at least shows that a deeper network based on $H_1$ does not guarantee improvement in the generated FO trajectories, unlike the case for spectral features.

On the basis of the first MUSHRA test, the second test examined the overall performance of $DS$, $HS$ and $HM_1$ with depths 4 and 40. The baseline depth was selected as 4 because it was commonly used for baseline feedforward neural networks. Here, the FO and spectral features of one synthetic sample were generated by the same network.

The results of the second test are shown in Figure 5. Additionally, a two-sided Wilcoxon signed rank test is used to measure the significance of the difference and the results are shown in Table 4. For all types of experimental network, the synthetic samples generated by the network with 40 hidden layers had a higher average score than those of the same type of network with just 4 hidden layers. According to Table 4, increasing the depth to 40 led to a statistically significant improvement for $DS$ and $HS$. This result is consistent with the objective evaluation and suggests that using a very deep network with more than 10 hidden layers is better than using a network with 4 hidden layers. Even the classical feedforward network is improved if it is sufficiently deep. When the network is very deep, both single- and multi-stream highway networks are good choices.

Interestingly, $HM$ had a higher average score than $HS$ when they each had 4 hidden layers. However, $HS$ outperformed $HM$ when the number of hidden layers was 40 for each network. One reason is that, when the network is shallow, the multi-stream structure takes advantage of the larger layer size. However, when the network is deep enough, even a single-stream network with a smaller layer size has enough capacity to model the spectral features as well as the FO features.
5. Analysis of the highway network

5.1. Activity of the highway network

The experiments in Section 4 showed that the depth of a multi-stream network had different effects on the generation of F0 and spectral features. To explain this result, we used the histograms introduced in Section 3 to analyze the network. Specifically, we used the input feature vectors of a specific phoneme from the test data as an excitation and plotted the histogram over the output of the highway gate. The histograms for different phonemes were similar and only the results for phoneme /a/ are shown here.

Figures 6 (a) and (b) show the histograms for HM_1 with 14 hidden layers. The histogram of the first block in the MGC sub-network, which is plotted in figure (b.1), showed an unbalanced binomial distribution, meaning that most of the gates were closed (T(x) \approx 0) while some gates were open (T(x) \approx 1). In the second block, the binomial distribution was kept to some degree. For blocks near the output side, the shape of the histogram gradually changed into a bell shape, which indicates that these blocks conducted complex transformations by summing the non-linear transformation output and the input.

The histograms for the F0 sub-network showed different patterns from those for the MGC sub-network. These histograms, especially near the end of the network, were spike-shaped. The width of the spike - variance of the data - was much smaller than that for MGC. Note that the spike in the histogram was located around 0.2, and this location was determined by the initial value of the gate units (\frac{1}{2}+\exp(1.5)). Nevertheless, these blocks generally avoided non-linear transformation and simply delivered their input to the next block. Therefore, most of the blocks in the F0 sub-network were inactive.

Similar results were also observed in the histograms for HM_1 with 20 and 40 hidden layers, which are shown in Figure 6 (c-f). Particularly, the blocks near the output side in the deeper F0 sub-network had a sharper spike in the histogram. This result was consistent for all phonemes and can be found on the author’s website. Because an inactive highway block tends to directly deliver the input data to the output side, adding these inactive blocks may not introduce additional feature transformation. This may explain why adding highway blocks did not improve the overall performance for F0 generation in the multi-stream highway network.

In contrast, the highway blocks in the MGC sub-network are generally active even in the network with 20 highway blocks. This result may explain the better performance of HM_1 in MGC generation as the depth increased from 2 to 40. Interestingly, Figure 6(e) also shows that the last block (b.20) in the MGC sub-network has a sharp histogram. This
the MGC and F0 sub-networks. Intuitively, we would expect increased model capacity for both F0 and spectral features. This indicates that the assumption is supported to some extent by the histogram of the deepest layers in Figure 6(e). In this case, we cannot differentiate the histograms for MGC and F0. Interestingly, the histograms of HS resemble those in the MGC sub-network of HM1. This resemblance suggests that HS mainly focused on MGC modeling instead of F0. If this is true, it explains why HS’s performance on F0 was worse than that of HM.

Results in Section 3 showed that HS performed better on F0 when the network was deeper. One reason may be that a larger network has additional capacity for modeling F0. This assumption is supported to some extent by the histogram of the deepest HS network shown in Figure 6(h). Compared with b.20 in Figure 6(e), b.20 of HS didn’t produce a sharp histogram. This indicates that the HS network took full advantage of the increased model capacity for both F0 and spectral features.

### 5.2. Sensitivity of the network to input textual features

The previous section showed different levels of activity in the MGC and F0 sub-networks. Intuitively, we would expect a highway block to be active when there is complex mapping between the input and target feature. It is possible that the blocks for F0 were inactive because the target F0 had fewer dimensions than those of the spectral features. It is also possible that the input features were not informative for F0 generation.

Given the trained networks, we next examined the second possibility by analyzing the sensitivity of neurons to the input textual features. The HM1 network trained with 14 hidden layers was analyzed based on the measure $E_{sk}$ introduced in Section 3. This measure was calculated over the test set for every highway gate neuron and textual feature class. Additionally, the average $E_{sk}$ over the sub-network was calculated for each textual feature class $S$. The feature classes with the highest and lowest $E_{sk}$ value are listed in Table 5.

Table 5: The rank of input feature in terms of the sensitivity level for $HM_1$ with 14 hidden layers. A smaller value of $E_k$ indicates a higher sensitivity level. Note that $E_k$ is the average $E_{sk}$ of all the neurons in each sub-network and was calculated on the test data. For example, $E_k$ of the phoneme identity for the MGC sub-network of $HM_1$ is the average value of all the black curves in Figure 7(a). Due to the limited space, only the top and bottom five features are shown here.

<table>
<thead>
<tr>
<th>in the MGC sub-network</th>
<th>in the F0 sub-network</th>
</tr>
</thead>
<tbody>
<tr>
<td>feature class $S$</td>
<td>$E_k$</td>
</tr>
<tr>
<td>Ph phoneme identity</td>
<td>0.9923</td>
</tr>
<tr>
<td>Top five entries</td>
<td>Position of phoneme in syllable (backward)</td>
</tr>
<tr>
<td>Position of stressed syllables in phrase</td>
<td>0.9974</td>
</tr>
<tr>
<td>Number of stressed syllables remained in phrase</td>
<td>0.9975</td>
</tr>
<tr>
<td>Bottom five entries</td>
<td>Position of phrase in utterance</td>
</tr>
<tr>
<td>Number of words in phrase</td>
<td>0.9999</td>
</tr>
<tr>
<td>Number of phrases in utterance</td>
<td>0.9999</td>
</tr>
<tr>
<td>Number of syllables in previous phrase</td>
<td>1.0000</td>
</tr>
<tr>
<td>ToBI boundary tone</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

suggests that $HM_1$ might not require 20 highway blocks on the corpus. Furthermore, the inactive blocks indicate that the highway blocks near the output side didn’t extract new hidden representations. Instead, they just refined the output of the previous block, which is consistent with hypothesis of ‘unrolled iterative estimation’ for deep highway networks [42].

The above analysis concerns the multi-stream highway network. Next, we plotted the histograms for HS with 14 hidden layers in Figure 6(g). In this case, we cannot differentiate the histograms for MGC and F0. Interestingly, the histograms of HS resemble those in the MGC sub-network of $HM_1$. This resemblance suggests that HS mainly focused on MGC modeling instead of F0. If this is true, it explains why HS’s performance on F0 was worse than that of HM.

As Table 5 shows, the neurons in the MGC sub-network were the most sensitive to the phoneme identity and position of the phoneme in the syllable. The $E_{sk}$ of each neuron with respect to these two feature classes are shown in Figure 7(a) and (b). The results are not surprising because the spectral features correlate mainly with the segmental information of the text. The results also showed that neurons in the F0 sub-network were sensitive to the segmental information. However, although F0 does correlate with segmental features [43, 44], we cannot see any input features above the word level that were highly ranked in Table 5. What’s more, the ToBI boundary tone [45] was ranked as one of the least sensitive features. This result may be because the input features over the whole corpus, including the ToBI boundary tone, were automatically annotated by the
text-analyzer. These input features may be too noisy to provide useful information for F0 generation.

A comparison of the curves for the MGC and F0 sub-networks in Figure 7 generally indicates that the neurons in the F0 sub-network were less sensitive to all classes of input features. In addition, it shows that most of the neurons in the F0 sub-network were insensitive even to the top two feature classes. This may explain why most histograms of the highway blocks in the F0 sub-network had a bell shape. In contrast, the neurons in the MGC sub-network showed high sensitivity at least to the phoneme identity, including the neurons near the network’s output. This is consistent with the shape of the histograms in the MGC sub-network.

In general, the results based on the sensitivity measure are consistent with the histograms of the MGC sub-network. This indicates that the network should be deep enough to fully transform the input features. However, the sufficient depth varies for different types of acoustic feature. Typically, the network for F0 can be shallower than the one for MGC.

6. Discussion

Experiments and analysis in this paper didn’t measure the correlation between F0 and spectral features directly. However, the experiments’ results suggest that F0 and spectral features cannot be assumed to be highly correlated for any corpus. They also suggest that the network structure should be carefully chosen to model spectral and F0 features together.

The experiments on the multi-stream highway network indicated that the F0 generation didn’t gain from using a deeper network. Analysis in Section 5.2 suggested that one reason may be the uninformative input features. Particularly, the automatically inferred ToBI boundary tones turned out to be uninformative. Of course, this result depends on the language and the front-end, and it does not deny the effectiveness of ToBI boundary tone in other scenarios. For example, we analyzed a multi-stream highway network trained on a large Japanese corpus and found that various information, such as the pitch accent type and inflected form of the word, were more useful than the phoneme identity since these features can be easily acquired from the surface text or retrieved from the lexicon. This result can be found online: http://tonywangx.github.io. However, even though the input features can be manually corrected, some of these features are not highly related to the realization of F0. As F0 is argued to rely on contextual information at various levels, it may be necessary to introduce more effective context features. Another approach is to consider the hierarchical property of the F0 and using hierarchical feature representation of F0 and spectral features as reported by this work.

One interesting direction for further work is to introduce the highway connection to RNN. Recent studies have proposed highway RNNs with complex structures. Based on a recent analysis of highway networks, we proposed a simple structure to combine the highway network and RNN. Preliminary results suggest that this structure performed relatively better than a deep RNN with a similar model size. Furthermore, its training time was less than that of a deep RNN. To control the length and content of this paper, the highway RNN is described in a separate report.

7. Conclusion

This work showed the effect of a neural network’s depth on the performance of F0 and spectral feature modeling for parametric speech synthesis. The comparison of multi-stream highway networks with different depths showed that the quality of generated spectral features improved when the network’s depth was increased up to 40. However, the generated F0 from the network with 4 hidden layers was not significantly worse than that from a deeper network. These results are quite different from those from the commonly used single-stream highway and feedforward network, both of which must be sufficiently deep to accurately model both F0 and spectral features as reported by this work.

The above results were further justified by histogram-based analysis of the highway gate. Analysis on multi-stream highway networks showed that the highway blocks in the F0 sub-network tended to carry their input directly to the next layer, while blocks of spectral sub-network conducted more complex non-linear transformations. This analysis supported our experimental evaluation on MGC and F0 modeling.

In a nutshell, a network should be deep enough to model the spectral features; on F0, a network can be shallow if it is one part of a multi-stream network. For the commonly used single-stream network, the depth should be sufficiently large. When such a deep network is required, the highway network may be more suitable than the plain feedforward networks.

8. Acknowledgement

The work presented in this paper was partially supported by EPSRC EP/J002526/1 (CAF), by the Core Research for Evolutional Science and Technology (CREST) from the Japan Science and Technology Agency (JST) (uDialogue project), by MEXT KAKENHI Grant Numbers (26280066, 26540092, 15H01686, 15K12071, 16H06302, 16K16096) and by The Telecommunications Advancement Foundation Grant.

References


