Investigating the impact of a neural network’s depth on spectral and F0 modelling for parametric speech synthesis

Xin WANG†, Shinji TAKAKI†, and Junichi YAMAGISHI†

† National Institute of Informatics, Hitotsubashi 2–1–2, Chiyoda-ku, Tokyo 101–8430, Japan
E-mail: †{wangxin,takaki,jyamagis}@nii.ac.jp

Abstract Deep neural networks have shown to be more effective than shallow networks in various machine learning applications. This paper investigates the performance of very deep neural networks on the speech synthesis task. However, as speech synthesis is a regression problem involving different kinds of target features such as spectral envelopes and F0 trajectories, this investigation focuses on two questions: how deep the network should be for speech synthesis and whether the spectral and F0 modelling benefit consistently as the network becomes deeper.

Based on a new architecture called multi-stream highway network, experiments on an English corpus showed that spectral features can be better predicted if the depth of the network is increased from 4 to 40, but not deeper. On the contrary, F0 trajectories can be well predicted by a network with only 2 hidden layers. Similar observation was also investigated on a large Japanese corpus with up to 50 hour’s speech recordings.

Key words speech synthesis, deep neural network, highway network, F0 modelling

1. Introduction

Text-to-Speech (TTS) aims at enabling the computer to read text messages as human beings. Usually, TTS consists of a text analyzer and an acoustic model, where the first module derives the pronunciation and prosody symbols from the text and the second module further converts those two kinds of linguistic specification into the speech waveform.

Statistical parametric speech synthesis (SPSS) can be used to implement the acoustic model of TTS. SPSS first generates acoustic features such as spectral features and F0 of speech based on the input linguistic specification. Then it uses the vocoder to construct the speech waveform based on the generated acoustic features. SPSS has achieved great performance by using the hidden Markov model (HMM) for generating the acoustic features [1]. Unfortunately, the shortcomings of the HMM-based framework prevent further improvement on SPSS. Recently, various methods based on artificial neural networks have been proposed to enhance or replace HMM for SPSS. Some methods incorporate neural networks with new architectures to model the multi-mode distribution [2], cross-time-dependency [3] or cross-dimension-dependency [4] of the acoustic features, hoping the accuracy of the generated acoustic features can be improved. Other methods focus on using neural networks to model raw and high-dimensional acoustic features such as the short-time FFT spectrum [5]. With specific configurations, all these neural-network-based methods give better results than the HMM-based approach.

A neural network conducts non-linear transformations to convert the input to the output. Its performance is influenced by the depth of the network [6]. However, due to the gradient vanishing and exploding problem, it is not easy to train a network with many hidden layers, and thus, the performance of a very deep neural network on SPSS is still unclear. Our recent work used the so-called highway network to ease the difficulty of training deep networks and showed the performance of networks with up to 14 hidden layers [7]. Particularly, a network with 14 hidden layers achieved better objective results than a normal network with 4 hidden layers. These highway networks were called single-stream networks because spectral and F0 features were generated from the single highway network. Then, we proposed a multi-stream highway network structure, which used partially separated sub-networks to generate the spectral and F0 features. Interestingly, it was shown that sub-networks behaved differently when generating the spectral and F0 features. One hypothesis based on the analysis of the multi-stream highway network is that generation of spectral features requires a deep network while F0 trajectory does not.

In this paper, we will use the multi-stream highway network to testify that hypothesis. While our previous work only compared single-stream highway network with different depth and multi-stream highway networks with different
layer sizes yet the same number of hidden layers, this work directly measures the performance of multi-stream networks with the depth ranging from 2 to 80. Another difference is that, a normalized initialization method is used to initialize the multi-stream highway network, instead of using Gaussian noise. The results directly show that a network with 40 hidden layers but not deeper generates the spectral features with the best performance on an English speech corpus. Meanwhile, the depth doesn’t influence the performance of the network on F0 generation and a multi-stream highway network with only 2 hidden \( \text{tanh} \) layers can generate F0 well. These results confirmed our previous observation on the different behaviours of the network in F0 and spectral feature generation. This work further uses the multi-stream highway structure on a large Japanese corpus including up to 50 hours’ training data and shows similar objective results.

In the rest of the paper, Section 2 introduces the multi-stream highway network. Section 3 and 4 present experiments on the English and Japanese corpora. Section 5 draws the conclusion.

2. Highway Block and the Multi-stream Highway Network

2.1 Difficulty in training deep neural network

A neural network conducts non-linear transformation on the input feature of a data sample for either a classification or a regression task. Usually, random initialization and back-propagation can be used to train the network with a limited number of hidden layers. But the simple training strategy does not guarantee a well-trained network when the network has more hidden layers. Typically, the gradients back-propagated in the network gradually vanishes when they pass through the saturated non-linear activation functions such as \( \text{tanh} \) and \( \text{sigmoid} \) in the hidden layers. Thus, the layers near the input end are trained poorly. This difficulty can be alleviated by several approaches. For example, the non-saturated non-linear function called Rectified linear unit (ReLU) can be used instead of the normal \( \text{tanh} \) or \( \text{sigmoid} \) function [8]. Another approach is to pre-train the network and provide the layers near the input end with better initial parameters [6] [9].

2.2 Highway block

Different from the above methods, the highway network alleviates the gradient vanishing problem by constructing a trainable information pass over the non-linear transformation layers [10]. In the back-propagation stage, this pass allows the gradients to flow backwards without being attenuated by the saturated non-linear function.

The basic building block of a highway network is called a highway block. It can contain a conventional feedforward hidden layer that transforms the input vector \( x \) into a target vector:

\[
\mathcal{H}(x) = f(W_H x + b_H). \tag{1}
\]

Here \( f(.) \) is the non-linear activation function such as \( \text{sigmoid} \) or \( \text{tanh} \), \( W_H \) is the transformation matrix and \( b_H \) is the bias vector.

What’s more, the highway block incorporates a \textit{gate unit} to compute a control vector using the sigmoid function \( \sigma(x) = \frac{1}{1 + e^{-x}} \):

\[
\mathcal{T}(x) = \sigma(W_T x + b_T) \tag{2}
\]

Based on the control vector \( \mathcal{T}(x) \), the highway block merges the output of the hidden layer \( \mathcal{H}(x) \) with the input \( x \) as

\[
y = \mathcal{H}(x) \odot \mathcal{T}(x) + (1 - \mathcal{T}(x)) \odot x, \tag{3}
\]

where \( \odot \) denotes the element-wise multiplication.

A highway block is shown in Figure 3. Note that, multiple feedforward hidden layers can be used in one highway block. The parameters \( W_T \) and \( b_T \) of the highway gate are trainable. When the trained highway gate gives \( \mathcal{T}(x) \approx 0 \), the
input \( x \) to this block can be directly propagated forwards (\( y \approx x \)). Similarly, the gradient can also be propagated backwards without being attenuated in the training stage. Thus, a very deep network can be trained quite easily. More interestingly, \( T(x) \) in the trained highway gate indicates the usefulness of the non-linear transformation \( H(x) \). Typically, \( T(x) \approx 0 \) means that the non-linear transformation doesn’t contribute to the output.

Note that, the dimension of \( H(x) \) should be identical to \( T(x) \) and \( x \). Otherwise, another transformation layer should be used to change the dimension of \( x \).

### 2.3 Multi-stream Highway Network for the Speech Synthesis Task

Multiple highway blocks can be cascaded as a highway network. This network will transform the input into a common hidden representation and then transform the hidden representation into spectral and F0 features. We call this network a *single-stream network*, which is shown in Figure 1.

However, this single-stream structure doesn’t do anything to alleviate the unbalanced dimension of spectral and F0 features. More importantly, since F0 and spectral depicts different aspects of speech signal, they may require different input information and different non-linear transformations. A shared network may have to be large enough to model both spectral and F0 features. Besides, the number of hidden layers can not be customized for spectral and F0 features separately.

Accordingly, we have proposed a multi-stream highway network shown in Figure 3[7]. In this architecture, multiple highway networks are used to generate the spectral and F0 features separately. On one hand, the influence of the unbalanced dimension can be alleviated. On the other hand, the influence of the network’s depth on F0 and spectral feature generation can be measured separately.

### 2.4 Normalized initialization strategy

Although the highway network can alleviate the gradient vanishing problem if the output of the gate unit \( T(x) \approx 0 \), it is irrational to expect that gate units of all layers have a small output. Otherwise, the output of the network will be similar to the input. Thus, we can’t only rely on gate units to address the gradient vanishing problem.

Fortunately, we found that a normalized initialization strategy [11] works for the highway network. Based on this strategy, the weights of the network will be initialized as

\[
W \sim \mathcal{U}[-\frac{\sqrt{6}}{\sqrt{n_j + n_{j+1}}} \sqrt{6}, \frac{\sqrt{6}}{\sqrt{n_j + n_{j+1}}}],
\]

where \( \mathcal{U} \) denotes the uniform distribution, \( n_j \) and \( n_{j+1} \) are the layer size of the \( j \)-th layer and \( j + 1 \)-th layer respectively.

### 3. Experiments on the English corpus

#### 3.1 Results of the previous experiments

In our previous work on highway networks [7] where all networks were initialized using Gaussian noise, we found that

- a single-stream highway network achieved the best objective performance when it had 14 hidden \( \tanh \) layers;
- in a multi-stream highway network with 14 hidden \( \tanh \) layers for all sub-networks, most of the gate units in the F0 sub-network gave \( T(x) \approx 0 \) while gate units in the sub-network for spectral features did not, which suggested that the sub-network for spectral features should be deep while that for F0 can be shallow.

Experiments here are different from the previous ones:

- the multi-stream highway networks with up to 80 hidden \( \tanh \) layers are tested so that the sufficient depth for speech synthesis can be found while the influence of depth on spectral features and F0 can be evaluated separately;
- the normalized initialization strategy is used so that the highway network can be better trained.

#### 3.2 Corpus and experimental systems

The corpus and configuration were the same as our previous work. The corpus is the Blizzard Challenge 2011 Nancy corpus [12] with 16 hours’ speech recording. Among the 12072 utterances, both the test and validation set contained 500 randomly selected utterances. Given the speech waveforms, mel-generalized cepstral coefficients (MGCs) of order 60, continuous F0 trajectory in a warped domain (1127 \* \log(1 + f/700)), voiced/unvoiced (V/U) condition, and band aperiodicity (BAP) of order 25 were extracted for each speech frame by using the STRAIGHT vocoder [13]. The dimension of the acoustic feature vector is 259. The Flite toolkit [14] conducted the text-analysis for the entire corpus. The output of Flite were converted into a vector of order 382 as the input \( x_t \) to the neural network.

For the experiment, we only included the multi-stream highway network. Based on our previous work, we set the layer size of the sub-network as 256 for MGC, F0 and BAP. Each highway block contained two hidden layers based on \( \tanh \) function. The number of highway blocks of the experiment systems ranged from 1 to 40, i.e. the number of hidden \( \tanh \) layers ranged from 2 to 80. The normalized initialization strategy was used to train all the systems. A modified CURRENNT toolkit [15] was used for the experiment [31].

#### 3.3 Results and analysis

The objective evaluation results are shown in Figure 2. All the objective measures are averaged over the results of

(11) The code and synthetic speech samples can be found on http://tonywangx.github.io
the last five training epochs. The standard deviation of the objective measure is also shown.

The first observation is that the F0 RMSE and correlation don’t show a clear pattern with the increased number of hidden layers. However, using 8 hidden tanh layers (HM$_8$) achieved better performance on F0 than the network with more hidden layers or larger layer sizes. In other cases, the difference between different systems is insignificant. This result is consistent with the observation that the highway blocks for F0 generation near the output end are inactive and the effective number of hidden layers should be less than 14. A F0 trajectory can be generated quite well even with a relatively shallow network. Note that, this conclusion is made on the multi-stream structure where the sub-network for F0 is not heavily influenced by the MGC sub-network.

Contrary to the result on F0, the accuracy on MGC generation dropped consistently when the number of hidden tanh layers was increased from 2 to 40. This result is also consistent with our previous observation that MGC generation may require more than 14 hidden layers on the same corpus. However, Figure 2 also shows that the RMSE on MGC generation increased when the network had more than 40 hidden layers. To explain the result, the gross error on the training and validation set is plotted in Figure 3. Obviously, although the error measure of HM$_{60}$ on the training set reached a low level as other deep networks, its error measure on the validation set was higher the other systems. Thus, HM$_{60}$ may overfit to the data.

In conclusion, on the Blizzard 2011 corpus, the most effective depth is 8 for F0 while 40 for MGC. While F0 generation doesn’t benefit from a very deep network, the increased depth improves the generated MGC when the depth is below 40. A deeper network doesn’t improve the generated MGC feature.

### Table 1

<table>
<thead>
<tr>
<th>Layer size of the sub-network for the English experimental systems based on the multi-stream highway network (HM). n ∈ [2, 80] denotes the number of hidden tanh layers.</th>
</tr>
</thead>
<tbody>
<tr>
<td>MGC stream</td>
</tr>
<tr>
<td>HM$_n$</td>
</tr>
</tbody>
</table>

### Figure 2

Objective measure of the multi-stream highway network (HM) with different depth on the English corpus. The depth, or the number of hidden tanh layers, is shown as the subscript.

### Figure 3

The gross RMSE on the training data (train.) and validation data (val.) of the HM$_n$ systems with n ∈ [20, 40, 60, 80].

## 4. Experiments on the Japanese corpus

### 4.1 Review of the previous experiments

Our previous experiments on large speaker-dependent Japanese corpora mainly compared the performance of feedforward and recurrent networks trained using different amount of data. The results showed that performance of both kinds of networks improves with more training data, particularly on F0 generation. Experiments in this section compares the multi-stream highway network with the recurrent and feedforward neural networks so that the results on the English corpus can be testified again and the scalability of the highway network can be shown.

### 4.2 Corpus and experimental systems

The same Japanese corpus and feature configuration in
our previous work were adopted for this experiment. The Japanese corpus is a speech corpus provided by ATR for the ‘XIMERA’ unit-selection system [16], [17]. For experiment here, we only used the subset of a female speaker F009, which contains about 50 hours’ recording. The same set of acoustic features as for the English experiments were extracted. The input linguistic feature vector of 389 dimension were automatically extracted by using an open-source text-analyzer from Open JTalk [18]. The validation and test set consisted of 260 utterances. The rest data formed the training set.

For the experiment, we used HM networks with 4, 14 and 20 hidden tanh layers. Similar to our previous work, we testified the networks’ performance using different amount of training data. As reference, we incorporated previously trained systems based on feedforward and recurrent networks [19]. These systems are listed in Table 2.

<table>
<thead>
<tr>
<th>Notation</th>
<th>Network Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>HM4</td>
<td>256 for each stream, 4 hidden tanh layers</td>
</tr>
<tr>
<td>HM14</td>
<td>256 for each stream, 14 hidden tanh layers</td>
</tr>
<tr>
<td>HM20</td>
<td>256 for each stream, 20 hidden tanh layers</td>
</tr>
<tr>
<td>RNN</td>
<td>bi-directional recurrent network with 2 recurrent layers of size 256 using long short term memory (LSTM) units and 2 feedforward layers of size 512</td>
</tr>
<tr>
<td>DNN</td>
<td>feedforward network with 1 feedforward layer of layer size 1024 and 3 feedforward layers of layer size 512</td>
</tr>
</tbody>
</table>

4.3 Results and analysis

Due to the prohibitive GPU resources required to train very large networks on huge amount of data, we did not train the HM14 and HM20 systems using the whole data corpus. On the MGC objective measure, HM4 gave a lower RMSE than DNN no matter how large the amount of training data was. The comparison between HM4 and HM14 showed that a deeper network improved the accuracy of generated MGC, which is consistent with the results on the English corpus. However, the gap between HM14 and RNN was still large.

On the F0 objective measure, all the HM systems achieved better performance than DNN. But the difference between these HM systems is small, particularly on the F0 correlation measure. This result indicates that F0 modelling for the Japanese corpus doesn’t require a deep network, which is again consistent with the result on the English corpus. It is also interesting to note that more than half of gap between RNN and DNN has been eliminated by HM. However, HM is still worse than RNN on both MGC and F0 modelling. One possible reason is that highway network based on feedforward network ignores the dependency among different time steps while the recurrent connection and LSTM unit in RNN models both long and short term dependency.

In a word, results on the Japanese corpus showed that F0 doesn’t require a deep network. By using the multi-stream structure, a network based on the multi-stream structure can be better than the normal feedforward network with the same number of hidden non-linear transformation layers.

5. Conclusion

This paper follows our previous work on using multi-stream highway network and explored the influence of the network’s depth on the accuracy of acoustic feature generation. By comparing the performance of multi-stream highway network with the number of hidden layers ranging from 2 to 80, the experiments on an English corpus showed that the network doesn’t need to be deep in order to generate F0 with good performance. However, the network must be deep
enough to convert the input to the spectral features such as MGC. For the used English corpus, the necessary depth for MGC generation is 40.

This paper further testifies the performance of multi-stream network on a large Japanese corpus. By comparing the multi-stream networks and the feedforward network with the same number of hidden \textit{tanh} layers, the experiment confirmed the assumption of the multi-stream network, i.e. F0 should be separately generated from the MGC features.

Future work will examine the performance of single-stream and normal feedforward network with the number of hidden layers up to 80. Besides, the reason for the different performance on F0 and MGC will be explored.

References


