Investigating Very Deep Highway Networks for Parametric Speech Synthesis

Xin Wang[♯‡], **Shinji Takaki**[♯], **Junichi Yamagishi**[♯‡†]
[♯]National Institute of Informatics, Tokyo, Japan
[‡]The Graduate University for Advanced Studies, Kanagawa, Japan
[†]The Centre for Speech Technology Research, The University of Edinburgh, Edinburgh, U.K.

1 Introduction

This work uses *highway networks* [1] to explore two questions:

- What's the performance of a network with more than 10 hidden layers for speech synthesis?
- Is it always the best strategy to share the same hidden states for generating different kinds of acoustic features?

The reasons to use highway networks:

- While a deep network is difficult to train, a *highway network* makes it easier by constructing an information highway over non-linear layers
- While the hidden states of a conventional network is hard to interpret, the output of the *highway gate* in a highway network shows plain information about the behavior of the network

2 The highway block and the highway network

Given an input vector $\boldsymbol{x}$, a highway block generates a output vector $\boldsymbol{y}$ as:

$$\boldsymbol{y} = \mathcal{H}(\boldsymbol{x}) \odot \mathcal{T}(\boldsymbol{x}) + (\boldsymbol{1} - \mathcal{T}(\boldsymbol{x})) \odot \boldsymbol{x}, \quad (1)$$
$$\mathcal{H}(\boldsymbol{x}) = f(\boldsymbol{W}_H \boldsymbol{x} + \boldsymbol{b}_H), \quad (2)$$
$$\mathcal{T}(\boldsymbol{x}) = \sigma(\boldsymbol{W}_T \boldsymbol{x} + \boldsymbol{b}_T), \quad (3)$$

where $\mathcal{H}(\boldsymbol{x})$ is the hidden state (or the hidden feature) computed by conventional feedforward layers, and $\mathcal{T}(\boldsymbol{x})$ denotes the output of the highway gate. $\sigma(.)$ in $\mathcal{T}(.)$ is a sigmoid function, and $f(.)$ in $\mathcal{H}(.)$ can be a $tanh$ function.

- When $\mathcal{T}(\boldsymbol{x}) \approx \boldsymbol{0}$, $\boldsymbol{y} \approx \boldsymbol{x}$ and gradients will not be attenuated by $f(.)$
- When $\mathcal{T}(\boldsymbol{x}) \approx \boldsymbol{1}$, $\boldsymbol{y} \approx \mathcal{H}(\boldsymbol{x})$ is the output of a normal feedforward layer

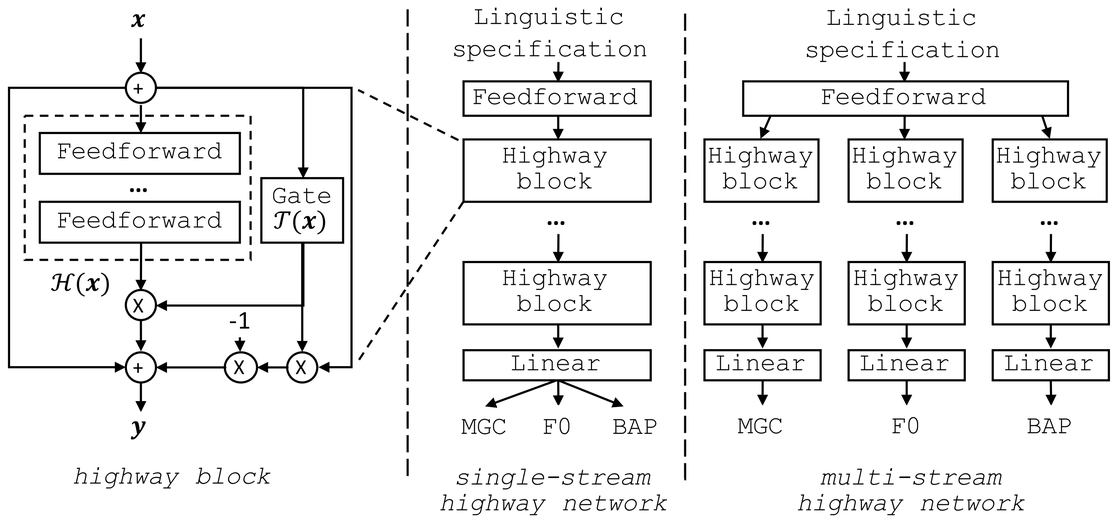

Fig.1 A highway block and the highway networks

A highway network contains several highway blocks:

- A single-stream highway network (**HS**) generates Mel-generalized cepstral (MGC), F0 and band aperiodicity (BAP) from a single highway network
- A multi-stream highway network (**HM**) uses separate highway networks to generate MGC, F0 and BAP

Note, **DS** denotes the single-stream feedforward neural network.

3 Corpus and Systems

- Corpus: the Blizzard Challenge 2011 corpus (the Nancy voice)
- Input: a 382-dim vector given by Flite (phoneme, pitch accent, etc.)
- Target: a 259-dim vector given by STRAIGHT
  - MGC, delta, delta-delta (180-dim in total)
  - U/V, continuous F0, delta, delta-delta (4-dim in total)
  - BAP, delta, delta-delta (75-dim in total)
- Initialization strategy:
  - **HS** and **HM**: the plain random initialization for $\boldsymbol{W}_H$ and $\boldsymbol{b}_H$
  - **HS** and **HM**: 0 for $\boldsymbol{W}_T$, -1,5 for $\boldsymbol{b}_T$
  - **DS**: the 'Normalized Initialization' [2]
- Training Strategy: the plain stochastic gradient descent algorithm

Toolkit and synthesis speech samples: http://tonywangx.github.io/
Comments & suggestions are welcome. Contact: wangxin@nii.ac.jp

4 Experiment on the depth of highway networks

- 14 hidden layers is a better choice than 4 hidden layers on this corpus
- 40 hidden layers is not significantly better

Note: Layer size is 382 for all systems

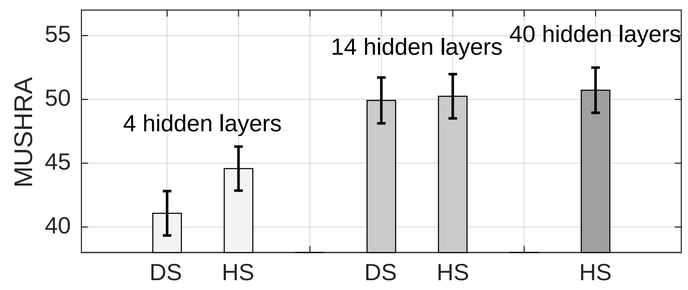

Fig.2 A MUSHRA test (with natural speech)

5 Experiment on the multi-stream highway network

Although the perceived difference is insignificant, **HM** achieved better objective results on F0 generation than **HS** and **DS**.

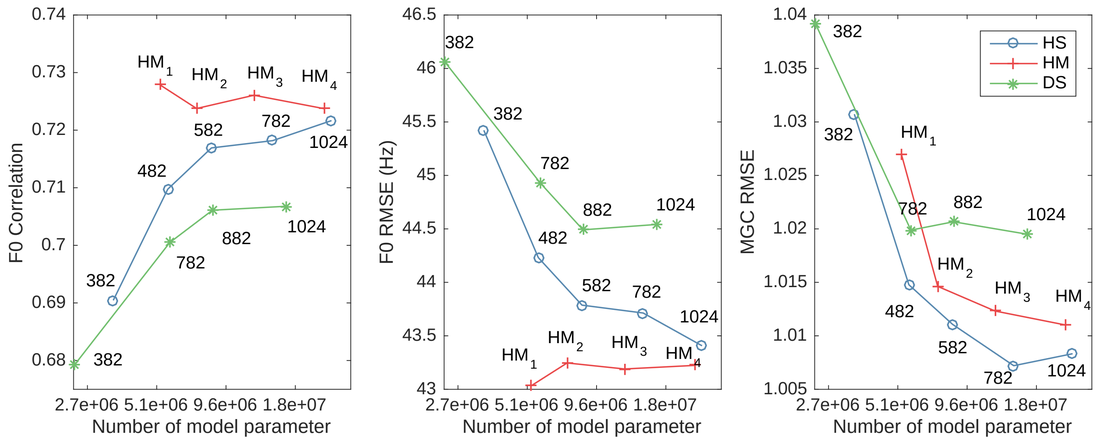

Fig.3 Performance of networks with 14 hidden layers. The number near **HS** and **DS** denotes their layer size ($\{382, 782, 882, 1024\}$).

Tab.1 Layer size of **HM**'s sub-networks.

	MGC	F0	BAP
$\mathbf{HM}_1$	256	256	256
$\mathbf{HM}_2$	382	256	256
$\mathbf{HM}_3$	512	382	256
$\mathbf{HM}_4$	768	512	256

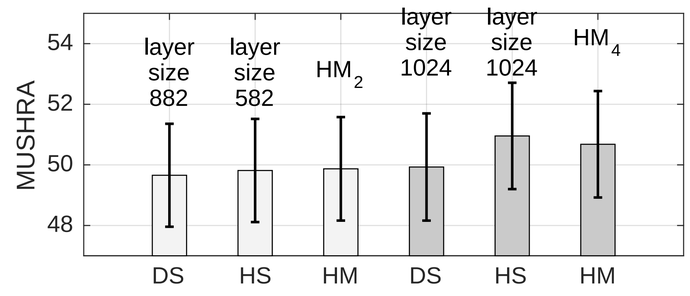

Fig.4 A MUSHRA test (14 hidden layers for all)

Why is **HM** better on F0 modeling?

- F0 and MGC may rely on different hidden representations. According to Fig.5, the generation of F0 requires less non-linear transformation
- The hidden states of **HS** (and possibly **DS**) is mainly for MGC generation.

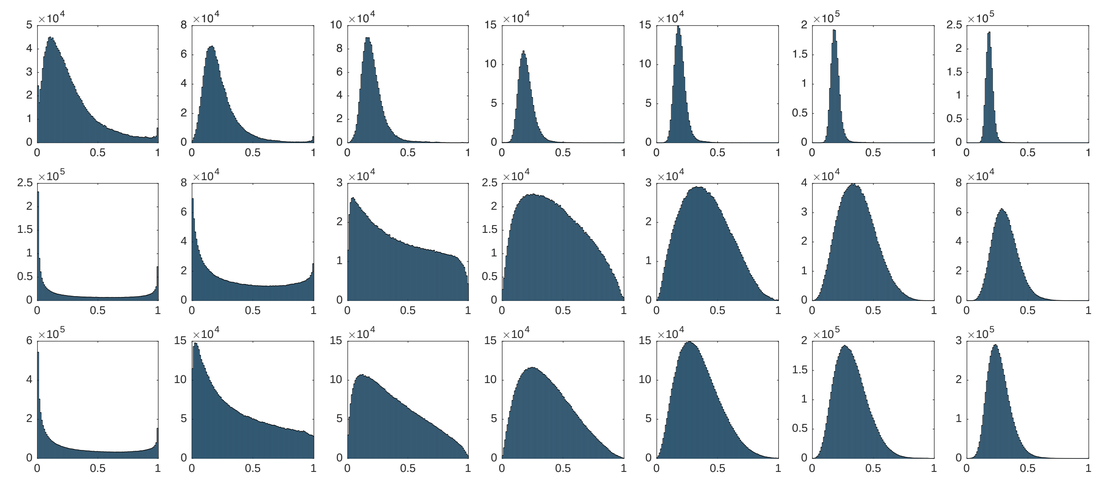

Fig.5 Histogram of $\mathcal{T}(\boldsymbol{x})$ for **HM**'s F0 stream (the first row), MGC stream (the second row), and **HS**'s single stream (the third row).

6 Conclusion

- A network should be deep enough but do not need to be deeper. The effective depth is 14 for the used corpus and experimental configuration;
- Generating F0 separately from MGC can improve the accuracy of the generated F0, particularly for a smaller network.

References

[1] Srivastava, R. K., Greff, K., & Schmidhuber, J. (2015). Training very deep networks. In Proc. NIPS (pp. 2377-2385).

[2] Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In Proc. AISTATS (pp. 249-256).