Investigating Very Deep Highway Networks for Parametric 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

The depth of the neural network is a vital factor that affects its performance. Recently a new architecture called highway network was proposed. This network facilitates the training process of a very deep neural network by using gate units to control a information highway over the conventional hidden layer. For the speech synthesis task, we investigate the performance of highway networks with up to 40 hidden layers. The results suggest that a highway network with 14 non-linear transformation layers is the best choice on our speech corpus and this highway network achieves better performance than a feed-forward network with 14 hidden layers. On the basis of these results, we further investigate a multi-stream highway network where separate highway networks are used to predict different kinds of acoustic features such as the spectral and F0 features. Results of the experiments suggest that the multi-stream highway network can achieve better objective results than the single network that predicts all the acoustic features. Analysis on the output of highway gate units also supports the assumption for the multi-stream network that different hidden representation may be necessary to predict spectral and F0 features.

Index Terms: speech synthesis, deep neural network

1. Introduction

Parametric speech synthesis aims at predicting speech acoustic features such as spectral and F0 features based on the input linguistic specification of text (in the case of Text-to-Speech) or conceptual representation of a potential sentence (in the case of Concept-to-Speech) [1]. This parametric method has achieved great performance by leveraging statistical models such as the hidden Markov model (HMM) [2].

Recently, this HMM-based framework was complemented or replaced by various methods using the neural network [3][4]. The claimed advantage of a neural network is its ability to extract structural features from the input data when the network is deep enough [5]. However, for speech synthesis, the comparison between HMM and deep feed-forward neural networks (DNN) with up to 5 hidden layers showed that increasing the depth of the network did not promise better performance for all kinds of acoustic features, especially, for the F0 features [3]. These results may be due to the difficulty in training the deep network since the authors of that paper showed later that a DNN with 7 hidden layers achieved consistently better performance than previous systems if the rectifier linear activation function (ReLU) [6] was utilized to facilitate the training process [7].

Thus, we can infer that, if the difficulty of training DNN can be further alleviated, deeper neural network may be better than 'shallow' models. Although researchers recently showed that a DNN with 5 hidden layers could extract effective phonemic information from the input spectral features for speech recognition [8], it is unknown whether a similar network is sufficient to extract linguistic features from the text for speech synthesis. Additionally, the neural network for speech synthesis computes a single hidden representation and then transforms it into spectral and F0 features at the output layer. Considering the differences between spectral and F0 features, we wonder whether sharing the same hidden representation is the best strategy.

In this paper, we investigate the above questions using a neural network called *highway network* [9]. The highway network utilizes trainable gate units to merge the output of a conventional non-linear transformation layer with the input to this layer, which equivalently builds a highway over the non-linear transformation layer and allows the input information to flow forwards. Similarly, gradients can be propagated backwards through the highway without attenuation, which eases the gradient vanishing problem. For image classification, a very deep highway network with more than 100 hidden layers can be easily trained and has achieved excellent performance [10]. For speech synthesis, the trainable gates in the highway network can be used to inspect the usefulness of the non-linear transformation layer. Typically, the gate units will favor the information on highway if the non-linear transformation is 'useless'.

On the depth of the neural network, our experiments show that a highway network deep enough (with 14 non-linear transformation layers) but not deeper could improve the accuracy of predicted acoustic features than the relatively shallow networks. On the basis of these result, we present a *multi-stream highway network*, where multiple highway networks sharing a common input hidden vector are used to predict spectral and F0 features separately. The analysis on the output of the highway gate units suggests that the spectral and F0 features may not necessarily share the same hidden representation. Experimental results also show that the multi-stream highway network performs better than the single-stream highway network and DNN.

Section 2 of this paper discusses the highway network and Section 3 presents the multi-stream highway network. Section 4 show the experiments, including the influence of the depth on the performance of the highway network, analysis on the gate unit and the performance of the multi-stream highway network.

2. Highway Network

A neural network with a single hidden layer can be easily trained by the back-propagation algorithm. However, the same method does not guarantee a well-trained network when the number of hidden layer increases. This difficulty can be alleviated by several approaches, including the pre-training strategy based on generative models [5][11].

$$\boldsymbol{y} = \mathcal{H}(\boldsymbol{x}) \odot \mathcal{T}(\boldsymbol{x}) + (1 - \mathcal{T}(\boldsymbol{x})) \odot \boldsymbol{x}$$
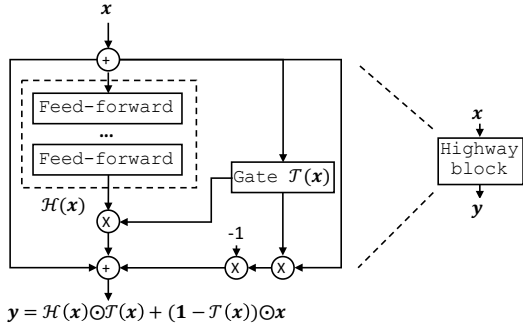
Figure 1: Computation flow in one highway block

Recently, researchers have proposed a new neural network architecture called *highway network* [9]. This new type of network is based on the classical feed-forward neural network. Similarly, each hidden layer in the highway network will transform the input vector $\boldsymbol{x}$ as

$$\mathcal{H}(\boldsymbol{x}) = f(\boldsymbol{W}_H \boldsymbol{x} + \boldsymbol{b}_H), \tag{1}$$

where $f(.)$ is the non-linear activation function, $\boldsymbol{W}_H$ is the transformation matrix and $\boldsymbol{b}_H$ is the bias vector. However, the highway network incorporates a new type of nodes called *gate unit* to compute a control vector as

$$\mathcal{T}(\boldsymbol{x}) = \sigma(\boldsymbol{W}_T \boldsymbol{x} + \boldsymbol{b}_T). \tag{2}$$

The activation function here is the sigmoid function $\sigma(x) = \frac{1}{1+e^{(-x)}}$. The parameters $\boldsymbol{W}_T$ and $\boldsymbol{b}_T$ are trainable.

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

$$\boldsymbol{y} = \mathcal{H}(\boldsymbol{x}) \odot \mathcal{T}(\boldsymbol{x}) + (1 - \mathcal{T}(\boldsymbol{x})) \odot \boldsymbol{x}. \tag{3}$$

Here, the $\odot$ denotes the element-wise multiplication.

In this paper, we use a highway block to represent the transformation from $\boldsymbol{x}$ to $\boldsymbol{y}$, as Figure 1 shows. When the output of the gate approaches zero, $\boldsymbol{x}$ can be directly propagated forwards ($\boldsymbol{y} \approx \boldsymbol{x}$). In this case, the gradient can also be propagated backwards without being attenuated by $\mathcal{H}(\boldsymbol{x})$. Thus, very deep network can be trained using the back-propagation algorithm. The highway block can be more complex by introducing another gate $\mathcal{C}(\boldsymbol{x})$ to replace $(1 - \mathcal{T}(\boldsymbol{x}))$ in Equation 3. It can also be simplified by eliminating all the gates and directly computing the output as $\boldsymbol{y} = \mathcal{H}(\boldsymbol{x}) + \boldsymbol{x}$. This simplified residual network have been used in an image classification task [10].

Note that, $\mathcal{H}(\boldsymbol{x})$ can be the non-linear transformation conducted by multiple hidden layers. The dimension of $\mathcal{H}(\boldsymbol{x})$ should be identical to $\mathcal{T}(\boldsymbol{x})$ and $\boldsymbol{x}$. If the dimension doesn't match, another transformation layer can be incorporated to change the dimension of $\boldsymbol{x}$.

3. Multi-stream Highway Network for the Speech Synthesis Task

3.1. Motivation

Similar to the application in image classification, a highway network can be directly utilized for the speech synthesis task. This network will transform the input linguistic specification by highway blocks and then map the transformed hidden representation into spectral and F0 features as Figure 2 shows. Because
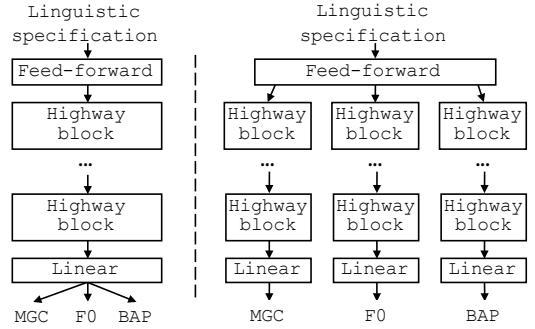

Figure 2: The single-stream (left) and multi-stream (right) highway network for speech synthesis. MGC and BAP denote mel-generalized cepstral coefficients and band aperodicity, respectively.

all the acoustic features are predicted by a single network, we call it a *single-stream* network.

A drawback of the single-stream structure is the unbalanced dimension of spectral and F0 features [12]. Another drawback is that the same hidden representation is utilized to predict the spectral and F0 features. Training this single-stream network with spectral and F0 features as targets is a Multitask Learning (MTL) task [13]. Although the theory of MTL argues that shared representation can improve the generalization ability of the model for every task involved, a prerequisite for this advantage is that those tasks should be related with each other. For example, MTL is beneficial when the system predicts perception-based spectral features and normal spectral features simultaneously [14]. For Text-to-Speech (TTS), the spectral features are highly correlated with the identity of segmental units of the input text. However, the F0 features are not only influenced by the segmental units (e.g. lexical stress) but also by the supra-segmental aspect such as the syntactic structure and discourse context of the text [15]. While the identity of segmental units can be easily retrieved from the input linguistic specification, linguistic information related to F0 prediction is not directly accessible. Thus, complex computation may be required to extract the useful information from the text. Besides, TTS systems usually provide the speech synthesizer with the prosodic features. These automatically inferred noisy features may also require additional transformation. Thus, different hidden representations may be necessary to predict F0 and spectral features.

3.2. The structure of a multi-stream network

To examine the above argument, we present a multi-stream highway network shown in Figure 2. Near the input end, a linear projection layer transforms the input vector into another vector of a certain dimension. Then, multiple highway networks are utilized to predict the spectral and F0 features separately.

By separating the highway network for each data stream, the influence of the unbalanced dimension can be alleviated. Besides, the multi-stream structure disentangles the non-linear transformation for spectral and F0 streams. Because the gate units are trainable, the number of non-linear transformation layers for each feature stream can be dynamically adjusted in a data-driven approach. Thus, it's more flexible than the single-stream structure. Besides, analysis on the output of the gate units in different streams can show whether F0 and spectral features should share a similar hidden representation.
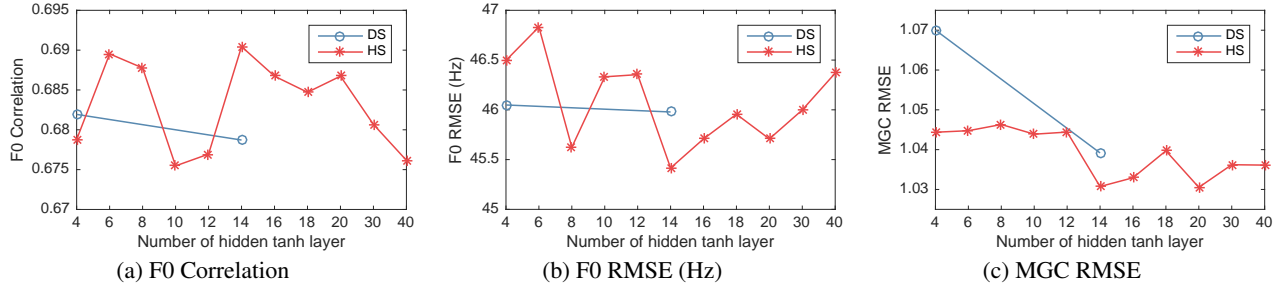
Figure 3: Results of the preliminary test on single-stream feed-forward network (*DS*) and highway network (*HS*) on the test set. Each highway block in *HS* contains two hidden transformation layers with the $tanh$ activation function.

4. Experiments

4.1. Corpus preparation and system notation

The speech database for experiments contains 12072 English utterances (16 hours) recorded by a female speaker in a neutral news reading style. Both the test and validation set contained 500 randomly selected utterances. Mel-generalized cepstral coefficients (MGC) of order 60, a one-dimensional continuous F0 trajectory, the voiced/unvoiced (V/U) condition, and band aperiodicity of order 25 were extracted for each speech frame by the STRAIGHT vocoder [16]. The delta and delta-delta components of the acoustic features except the voiced/unvoiced condition were also extracted. The Flite toolkit [17] was used for all the systems to conduct the grapheme-to-phoneme conversion and prosodic prediction for both the training and test sets. These phonemic, syntactic and prosodic feature were encoded into a vector of 382 dimension as the input to the neural network. In this paper, the MGC is referred to as the spectral features.

Three kinds of systems listed in Table 1 were involved in experiments. The toolkit for training the neural network was modified on the basis of the CURRENNT library [18]. All the experimental systems were trained based on the stochastic gradient descent strategy without pre-training.

4.2. Preliminary test on the depth of the highway network

This experiment tested *HS* systems with 2 to 20 highway blocks. Every block contained 2 hidden transformation layers with the $tanh$ activation function. Thus, the deepest *HS* included 40 transformation layers in total. Layer size of the transformation and gate layers was set to 382 in order to avoid the transformation on the dimension of the input data. Bias of the gate was initialized as -1.5 while other parameters were randomly initialized. As the author of [9] notes, the gates with -1.5 as the initial bias tends to deliver the gradients to the non-linear transformation layers near the input end during the first several training epochs. This strategy ameliorates the gradient vanishing problem. We also trained systems based on the initial value of 0.0 and 1.5, and the results supported the choice of -1.5.

Objective results on the test set are shown in Figure 3. The comparison among *HS* groups shows that the RMSE on the MGC generally decreased with the increasing number of highway blocks. However, the RMSE gradually increased after the number of highway blocks was larger than 14. This result suggests that, although deeper network is helpful in spectral acoustic features modelling, the depth can not be increased infinitely without over-fitting to the data.

As reference systems, *DS* systems with 4 and 14 hidden layers were trained. The layer size of the shallow *DS* was set to 512 following our previous experiments while that of the deep

Table 1: Experimental systems involved in experiments.

Notation	Definition
HS	Single-stream highway network
HM	Multi-stream highway network
DS	Single-stream deep feed-forward network

DS was set to 382. The results demonstrates that *DS* with 14 hidden layers performed similarly as the best *HS* system. Note that the *DS* system with 14 hidden layers was trained based on a special initialization strategy [19]. It can be inferred that either a better initialization strategy or a carefully designed network such as the highway network can help the training process of the deep network.

On F0 prediction, *HS* with 14 transformation layers performed better than other networks. But generally the improvement was not consistent with the depth of the network. Particularly, highway networks with 14 to 20 non-linear transformation layers can be worse than the *DS* system with only 4 hidden layers. This gap is quite different from that in the MGC prediction.

4.3. Experiments on multi-stream highway networks

Based on the results of the preliminary experiment, we tested the performance the multi-stream structure *HM* against *HS* and *DS* when the number of hidden transformation layer for all systems was fixed as 14. Following the argument in Section 3, *HM* adopted three highway networks for MGC, F0 and BAP features. To compare the performance, experimental systems with different sizes of the hidden (and gate) layer were trained. Because *HM* adopted multiple sub-networks, the layer size of each sub-network can be adjusted more flexibly. Thus, as shown in Table 2, *HM* systems with different configurations of layer size were trained.

The results are shown in in Figure 4. The comparison between *DS* and *HS* indicates that the highway network can perform better than the traditional feed-forward network when the number of model parameters is comparable. Interestingly, the RMSE on MGC increased when the layer size of *HS* was increased to 1024. But the performance on F0 prediction did not degrade.

Compared with *HS*, *HM* achieved better performance on F0. This result can partially supports that F0 prediction in the single-stream structure may be affected by the unbalanced dimension of spectral and F0 features. However, *HM*'s performance on MGC stream can not surpass *HS* when the total number of parameters is comparable. One possible reason is that the size of the network for MGC stream in *HM* is always smaller than the size of *HS*. If *HS* devotes most of the model capacity to model MGC features, its performance on MGC prediction

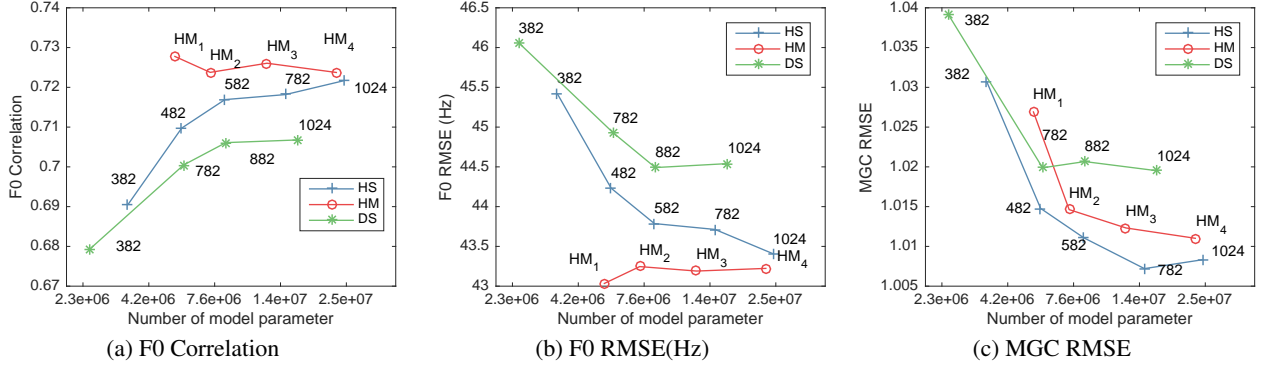(a) F0 Correlation   (b) F0 RMSE(Hz)   (c) MGC RMSE

Figure 4: Performance of the multi-stream highway network (*HM*) against the single-stream highway (*HS*) and feed-forward network (*DS*) on the test set. All systems contained 14 hidden transformations layers. The number associated with every *HS* and *DS* system is the layer size of the network. Definition of HM_1 to HM_4 is listed in Table 1. For reference, the number of parameter of the *DS* with 4 hidden layers and 512 units per layer is 1.1e+06.

Table 2: The network structure of *HM* systems in Figure 4.

Notation	Layer size of the sub-network		
	MGC stream	F0 stream	BAP stream
HM_1	256	256	256
HM_2	382	256	256
HM_3	512	382	256
HM_4	768	512	256

is expected to be better than *HM*. Note that, a fair comparison on MGC prediction is impossible because the exact number of hidden units devoted by *HS* for MGC modelling is unknown.

However, the performance of *HM* on MGC prediction can be improved by increasing the layer size of the MGC sub-network, which is shown by comparison between HM_1 and HM_2. Increasing the layer size from HM_2 to HM_3 resulted in in further improvement. However, when the layer size of F0 sub-network was increased to 512, HM_4's performance on the F0 prediction degraded, which may suggest the layer size below 512 is sufficient for F0 modelling on the utilized corpus. In general, the network structure of *HM* can be adjusted in a flexible way. If the layer size is carefully chosen, *HM* can achieve better overall performance than *HS*.

Subjective evaluation on the experimental systems is currently undergoing. Results will be updated on the website [1].

4.4. Analysis on the multi-stream highway network

Compared with the single-stream network, the performance of the *HM* systems on spectral and F0 streams showed different patterns. To interpret the results, we analyzed the histogram of the output of the gate units in HM_1. Figure 5 shows the results for all the test data frames corresponding to the phoneme /a/. Particularly, Figure 5(a)-(b) shows the output of the gate units in the MGC stream after the first and last training epoch. Because the bias of the gates was set to -1.5 and model parameters were randomly initialized, the output of the gate units were around 0.2 ($\approx \frac{1}{1+exp(1.5)}$) before the training process began. Then, after the first training epoch, the output of the gate was still around 0.2 as Figure 5(a) shows. Note that the variance of the gate output in block 1 was larger than other blocks be-

---

[1]Subjective evaluation results, synthetic samples, scripts and CURRENNT for highway network training can be found on http://tonywangx.github.io/.

cause the input to this block was not bound by the shared linear projection layer preceding it. However, after the last training epoch as Figure 5(b) shows, the gate output in this block approximated a binomial distribution, which indicates that it may derive a sparse representation for the spectral features.

The binomial distribution can observed in the second highway block. In the following blocks, the gate output approximated the distribution of a bell shape. This trend indicates that the highway blocks near the output of the network conduct a complex transformation based on the weighted sum of the input and output of the non-linear transformation layers. Note that, the histograms for different phonemes are similar, although each individual dimension of the gate output showed different distribution across phonemes.

The distribution of gates' output in the F0 stream, as Figure 5(c)-(d) shows, was different from that in the MGC stream. In block 1, only a few dimensions of the gates' output approached 1.0. In the following highway blocks, the gate output was dominated by the mode near 0.2, and the bias of those highway blocks after model training was still similar to the initial value. These results indicated that the highway blocks were only slightly tuned. But this dorment network seemed to be more effective for F0 modelling than the single-stream highway and normal feed-forward network. Thus, these results suggest that the different hidden representations may be beneficial to model different kinds of acoustic features.

The histogram of the gate output of *HS* is shown in Figure 6. Although we can not differentiate the hidden representation for spectral and F0 streams, the distribution of gate output resembled the spectral stream in Figure 5(b). Because the number of the dimension of the spectral features was much larger than that of the F0 features, this observation may support the assumption that the hidden representation for the spectral features dominates the neural network based on the single-stream structure.

5. Conclusion

By leveraging gate units that control information flow over the conventional hidden transformation layer, the highway network provides a good way for training a deep neural network. In this paper, we investigated the use of the highway network for the speech synthesis task. Experimental results show that a highway network with 14 non-linear transformation layers can
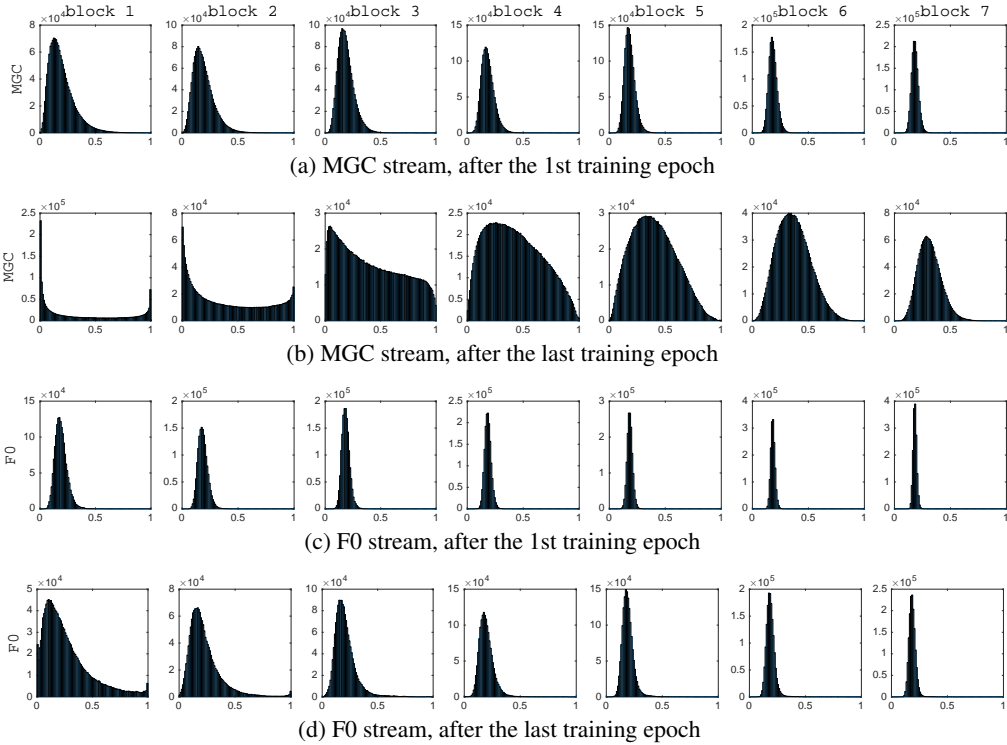
(a) MGC stream, after the 1st training epoch

(b) MGC stream, after the last training epoch

(c) F0 stream, after the 1st training epoch

(d) F0 stream, after the last training epoch

Figure 5: Histogram of the output of gates units in highway blocks of the multi-stream highway network HM_1 (7 highway blocks). (a)-(b) show the gates in the sub-network for MGC; (c)-(d) for F0. block 1 is near the input layer while block 7 is linked to the output. The data were generated given all the input data of phoneme /a/ in the test set.
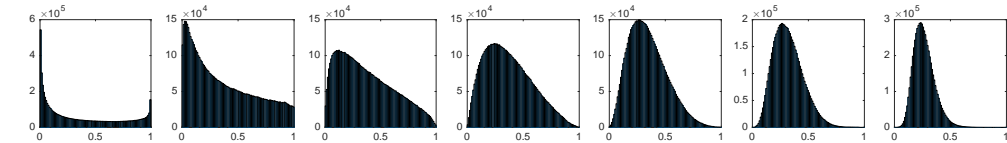

Figure 6: Histogram of the gates output $\mathcal{T}(\boldsymbol{x})$ of network HS with 14 hidden layers and 1024 layer size after the last training epoch. The data were generated for all the frames of phoneme /a/ in the test set .

achieve better performance than a feed-forward network with 4 to 5 hidden layers.

The highway network can also be utilized to analysis the performance of the neural network. Typically, the distribution of the gate output in the highway network indicates that different non-linear transformations may be preferred to derive hidden representations for spectral and F0 prediction. Accordingly, a multi-stream highway network, in which separate highway networks are utilized for predicting spectral and F0 features, can achieve better performance for F0 modelling while yield similar performance on the spectral part.

It is not surprising that different representations are required to predict spectral and F0 features. However, although we assume that more complex transformations may be preferred for the F0 stream, experimental results only showed a dormant network for the F0 prediction. This dormant network may be due to the ineffective input features. For the future work, the sensitivity measure defined in [20] may tell what kind of information contributes to the F0 modelling task.

In this work, we only considers the highway network based on the feed-forward neural network. It is also possible to combine the highway with recurrent neural network as what Zhang et al. did for speech recognition task [21].

6. Acknowledgements

7. References

[1] P. Taylor, *Text-to-Speech Synthesis*. Cambridge University Press, 2009.

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

[4] Z. H. Ling, S. Y. Kang, H. Zen, A. Senior, M. Schuster, X. J. Qian, H. M. Meng, and L. Deng, "Deep learning for acoustic modeling

in parametric speech generation: A systematic review of existing techniques and future trends," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 35–52, 2015.

[5] Y. Bengio, "Learning deep architectures for ai," *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, Jan. 2009. [Online]. Available: http://dx.doi.org/10.1561/2200000006

[6] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *International Conference on Artificial Intelligence and Statistics*, 2011, pp. 315–323.

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

[8] T. Nagamine, M. L. Seltzer, and N. Mesgarani, "Exploring how deep neural networks form phonemic categories," in *INTERSPEECH-2015*, 2015.

[9] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," *CoRR*, vol. abs/1505.00387, 2015. [Online]. Available: http://arxiv.org/abs/1505.00387

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[11] G. E. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science 28*, vol. 313, no. 5786, pp. 504–507, 2006.

[12] S. Kang and H. M. Meng, "Statistical parametric speech synthesis using weighted multi-distribution deep belief network," in *INTERSPEECH-2014*, 2014, pp. 1959–1963.

[13] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

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

[15] J. Cole, "Prosody in context: a review," *Language, Cognition and Neuroscience*, vol. 30, no. 1-2, pp. 1–31, 2015.

[16] H. Kawahara, I. Masuda-Katsuse, and A. 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.

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

[18] 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.

[19] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *International conference on artificial intelligence and statistics*, 2010, pp. 249–256.

[20] K. C. Sim, "On constructing and analysing an interpretable brain model for the dnn based on hidden activity patterns," in *ASRU-2015*, 2015, pp. 22–29.

[21] Y. Zhang, G. Chen, D. Yu, K. Yao, S. Khudanpur, and J. Glass, "Highway long short-term memory rnns for distant speech recognition," in *ICASSP-2016*. IEEE, 2016, pp. 5755–5759.