A Comparative Study of the Performance of HMM, DNN, and RNN based Speech Synthesis Systems Trained on Very Large Speaker-Dependent Corpora

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

This study investigates the impact of the amount of training data on the performance of parametric speech synthesis systems. A Japanese corpus with 100 hours' audio recordings of a male voice and another corpus with 50 hours' recordings of a female voice were utilized to train systems based on hidden Markov model (HMM), feed-forward neural network and recurrent neural network (RNN). The results show that the improvement on the accuracy of the predicted spectral features gradually diminishes as the amount of training data increases. However, different from the "diminishing returns" in the spectral stream, the accuracy of the predicted F0 trajectory by the HMM and RNN systems tends to consistently benefit from the increasing amount of training data.

Index Terms: speech synthesis, deep neural network, hidden Markov model

1. Introduction

The framework based on the hidden Markov model (HMM) is one of the classical methods for the parametric speech synthesis [1]. Despite the good quality of the synthetic speech given by this method, the HMM-based framework suffers from disadvantages including the data fragmentation caused by decision-trees [2] and limited capabilities of Gaussian distributions in each HMM state [3]. Driven by these drawbacks, various methods using neural networks (NN) have been proposed to either complement [3] or replace the HMM-based framework [2, 4, 5]. Results have shown that NN-based parametric speech synthesizers can yield synthetic speech with better quality.

In fact, pilot speech synthesizers based on neural network can be traced back to 1990s [6, 7, 8]. The recent resurgence of the approaches based on neural networks is possible because of various factors that facilitate the training process of neural networks, including the efficient initialization strategy [9] and more powerful computing devices. Another vital factor may be the large amount of available data for training deep neural networks. It is assumed that "there is no data like more data" for deep neural networks. The impact of the amount of data is well reflected in the speech recognition (ASR) task where a acoustic model similar to that in speech synthesizer must be trained. Amodei et al. showed that the word error rate of a NN-based speech recogniser "decreases by 40% relative for each factor of 10 increase in training set size" [10] and the system achieved the best performance when all the train data of 12000 hours were utilized. Compared with the classical HMM framework, the NN-based ASR systems seems to be more efficient to take advantage of the huge amount of training data [11].

For speech synthesis, explicit explanation in the papers show that existing NN-based approaches typically use training data with duration from 81 minutes [5] up to 16 hours. Although researchers may have utilized more data that that amount for the speech synthesis task, compared with that for the ASR task, the amount of data for speech synthesis seems to be smaller. One reason is that the acoustic model for speech synthesis is not required to be speaker-independent or noise robust. Thus, the amount of training data required to cover speaker and noise variance can be reduced and the data corpora for the speech synthesis task can be smaller. However, the corpus for speech synthesis must be carefully designed to ensure coverage on the phonemic and prosodic events of a target language. The audio data must be recorded in controlled conditions. Besides, speech synthesis corpus usually requires manual data annotation such as annotation of the prosodic events. Thus, the cost in preparing a large corpus for speech synthesis is somehow prohibitive and the corpora can not be larger. Fortunately, if the corpus is carefully designed and prepared, NN-based speech synthesizer can be well-trained based on the 'small' corpora.

In this study, we are curious about the quality of synthetic speech based on larger corpora for the speech synthesis task. Although related work on ASR may argue that the benefit from the increasing amount of training data gradually diminishes [12, 13], we think this study necessary as the speech synthesizer predicts not only the spectral features but also the F0 trajectories. It is interesting to see the results on predicting all the acoustic features and listen to the difference of the quality of the synthetic speech. We train the systems with the duration of the speaker dependent training data ranging from 20 hours up to 100 hours. The trained speech synthesizers are based on the HMM framework, the feed-forward neural network (DNN) and the recurrent neural network (RNN). We admit that deeper analysis on the NN with a huge amount of training data could be more interesting. However, in this initial work, we keep the system as a black box and only show the results of objective measure. One interesting finding is that, while the accuracy of predicted spectral features tends to converge, the improvement on the F0 trajectory modelling tend to be scalable with increasing amount of the training data up to 100 hours.

2. Corpora and Data Preparation

2.1. Corpora

The speaker dependent corpus that we used for this comparative study is a Japanese speech corpus collected for a unit selection system developed at ATR in the past, called 'XIMERA' [14, 15].

The corpus contains a sub-corpus of a male speaker M007 and another of a female speaker F009. The duration of speech data uttered by the male speaker is about 110 hours and that of the female speaker is about 50 hours. The recording took 181 days over a span of 973 days for the male speaker and 95 days over 307 days for the female speaker. Domains of sentences included in the corpus cover newspaper, novel, travel conversation. All the transcriptions are manually verified so that they are consistent with audio files. Speech signals were recorded at 48kHz sampling frequency with 24 bits precision in a sound proof room using the identical microphone over the entire periods. This is one of the largest speaker-dependent speech synthesis corpus that we can obtain in Japan.

2.2. Feature extraction

For the NN-based systems, the target acoustic feature vector includes 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. These acoustic features were extracted for each speech frame by the STRAIGHT vocoder [16] with the F0 adaptive window and 5ms frame shift. The delta and delta-delta components of the acoustic features except the voiced/unvoiced condition were also extracted. Although the dynamic components are theoretically unnecessary for recurrent neural network, we still utilized these features in order to compare the systems performance fairly. The dimension of the target acoustic feature is 259. The acoustic feature vector for the HMM-based systems was prepared in a similar way except that the discontinuous F0 was used instead of the continuous F0 with V/U.

The input linguistic features were automatically extracted from the speech transcription using the front-end modules of an open-source Japanese Text-to-Speech system called Open JTalk [17]. This frond-end conducted normal procedures of text analysis, including grapheme-to-conversion, part-of-speech tagging and morphological analysis based on a Japanese parser called Mecab [18]. The derived quin-phone identity, discrete and numerical syntactic and prosodic information about the text were further encoded into a vector of 389 dimension as the input to the neural network speech synthesizer. The same input information was encoded into a slightly different format as the model context for the HMM-based system.

For the M007 corpus, both the validation and test set consisted of 500 randomly selected utterances. For the F009 corpus, the validation and test set consisted of 260 utterances. The rest data formed the training set.

3. Experiments

3.1. Experimental systems and training recipes

The HMM systems followed the classical recipe of HTS toolkit verison 2.3. Each full-context HMM model contained 5 states and a state transition matrix only for the left-to-right state transition. Multi-space distribution HMM (MSD-HMM) [19] was leveraged to model the F0 feature streams. Monophone MSD-HMMs were estimated at first and then converted into context-dependent HMMs. After that, parameter tying of the context-dependent HMMs was applied with the help of decision tree techniques in order to ensure robust estimation of the model parameters.

The recurrent neural network contained 2 feed-forward layers with 512 nodes perl layer and 2 bi-directional recurrent layers with 256 long-short term memory (LSTM) units per layer. This DBLSTM-RNN structure was adopted based on an existing work using RNN for TTS [4]. The feed-forward neural network contained the 4 hidden layers where the layer size for each layer was $(1024, 512, 512, 512)$. The first layer was set larger in order to control the number of model parameters comparable to that of the DBLSTM-RNN system.

The CURRENNT library [20] was utilized to train both the DBLSTM-RNN and DNN systems. All the NN systems were trained based on random initialization and simple stochastic gradient descent with early stopping. For all the DBLSTM-RNN systems, *parallel sentence training* with 20 utterances in parallel was utilized in order to decrease the training time of one epoch [20]. This parallel sentence training strategy conducts back-propagation for multiple utterances simultaneously and accumulates the gradients from the parallel utterances. Dummy slot is introduced to skip the time slot beyond the duration of the shorter utterances. For all the DNN systems, parallel sentence training was not utilized. Note that, the CURRENNT toolkit is designed mainly for RNN. In default, each utterance is assumed to be one mini-batch (multiple utterances as one mini-batch in the parallel sentence mode). It is not the average but the sum of the gradients that adjusts the model parameter. Thus, we utilized small learning rate (4e-06) for DNN and DBLSTM-RNN systems.

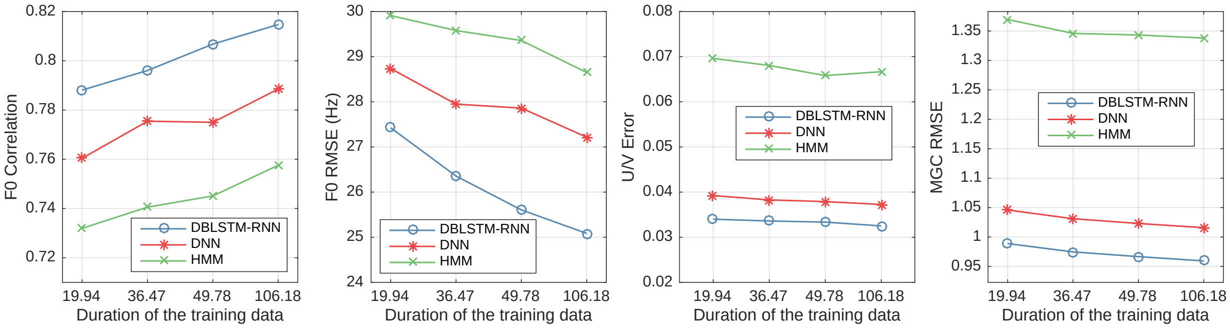
In the synthesis phase, the MLPG algorithm was used to calculate the acoustic feature trajectories. The phoneme-level alignment of the natural voice in the test set was used for all the experimental systems. In other words, the HMM-based systems had to predict the state-level duration for the test set while the DNN and DBLSTM-RNN systems had to generate the temporal dynamic of the acoustic features for each phoneme. At last, post-processing was used to enhance the spectral features and then STRAIGHT vocoder was used to construct the waveforms.

3.2. Objective results

The performance of the system was evaluated based on objective measure. Figure 1 and 2 show the results on the F009 and M007 corpus respectively. Generally, DBLSTM-RNN achieved the best performance in all the cases, followed by DNN and HMM-based systems. The differences between the modelling methods are huge. To give a few examples, F0 correlation of the HMM system using 100 hours of the male speech data is comparable to that of the DNN system using 10 hours of the male speech data. Moreover, F0 correlation of the DNN system using 100 hours of male speech data is comparable to that of the DBLSTM-RNN system using just 10 hours of male speech data. The performance of the DBLSTM-RNN system further improves when we use more data.

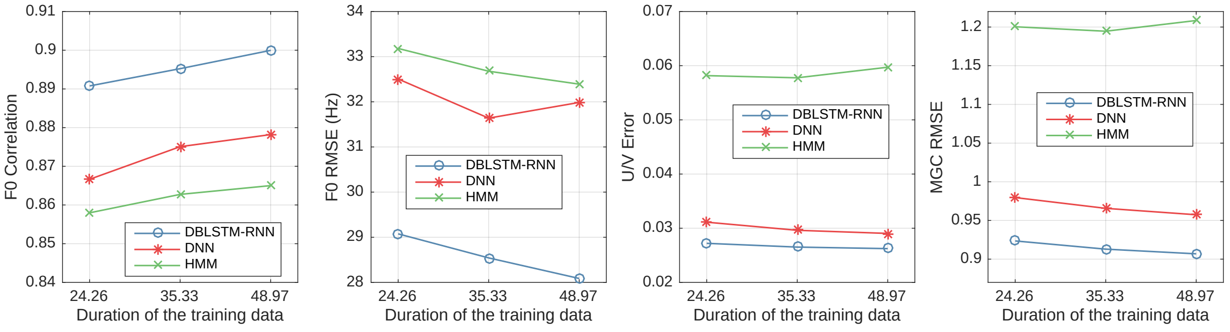
We can infer that the data amount generally has a positive impact on the objective measure results. However, different trends can be found in MGC and F0 features. The RMSE of predicted MGC drops as the data amount increases. But the objective measure tends to converge to a certain level. In comparison, except those of the DNN-based systems, the objective measure of the DBLSTM-RNN and HMM systems improves in a linear scale with the amount of the training data. Particularly, the gap between the F0 RMSE of the DBLSTM-RNN and HMM on the M007 and F009 corpora seems to be enlarged by the increased amount of data.

Figure 3(a) compares the predicted F0 trajectory of one randomly picked sample generated by the DBLSTM-RNN systems. For the voiced segment between the 250th and 350th frame, the DBLSTM-RNN trained with the full training set



(a) (b) (c) (d)

Figure 1: Performance of the DBLSTM-RNN, DNN and HMM on the M007 corpus.



(a) (b) (c) (d)

Figure 2: Performance of the DBLSTM-RNN, DNN and HMM on the F009 corpus.

predicted more accurately than other systems. In the other part of Figure 3(a), the advantage of this system is not obvious or even worse. In a word, increasing the amount of training data does not promise the improvement for every segment of the synthetic speech. In Figure 3(b), the comparison shows that the F0 trajectory predicted by the DNN-based system deviates from the natural one more than the HMM-based system. The peculiar performance of the DNN systems on F0 modelling can also be observed in Figure 1 and 2. One possible reason is the weak ability of the DNN to model temporal variation of the F0 trajectory.

Comparison across the corpora is also informative. If we compare the F0 correlation of M007 with that of F009, we can see that systems trained on the smaller F009 corpus have better correlations than systems trained on the larger M007 corpus. Besides, the perceived quality of the synthetic female speech is better than the synthesized male speech. One possible reason may be the variance of the physical and mental condition of the male speaker during the long recording period.

4. Conclusion

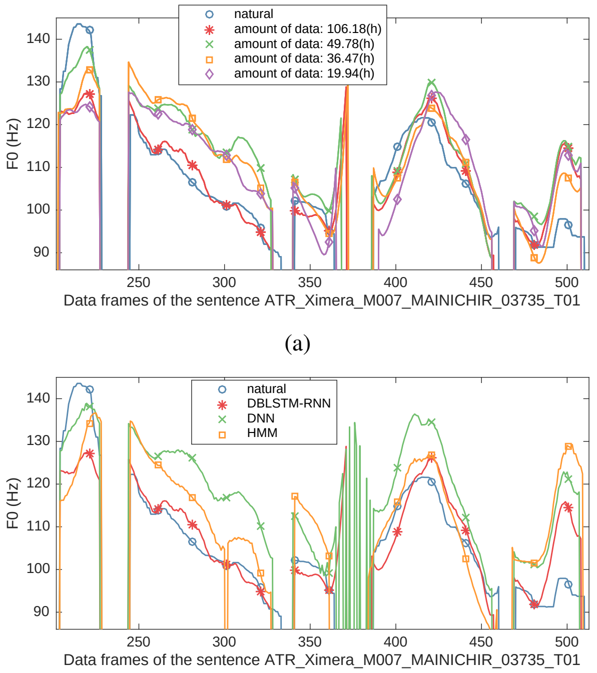
This initial work investigated the impact of the amount of training data on the objective performance of speech synthesisers based on HMM, DNN and DBLSTM-RNN networks. A female voice speech corpus with 50 hours' recordings and a male voice speech corpus with 100 hours' data were utilized. The results showed that the accuracy of the predicted MGC features improved yet converged gradually when the amount of the training data was increased. On the other hand, from the perspective of data modelling, prediction of the F0 features in both RNN and HMM systems seemed to surprisingly benefit from the increased data amount consistently even then the full training data was used. However, as the one of the reviewer points out, we also have to realize that a better F0 model is not guaranteed by the blindly increasing the amount of training data to the F0 model based on frame-level trajectory. Better design and extraction of the input and output features can be more beneficial.

The subjective evaluation was partially finished. We will update the subjective evaluation results on the website [1]. The synthetic samples can also be accessed from the website.

Based on the large data corpus, more interesting work will be conducted. First, deeper models based on a new type of neural network called highway network [21] could be leveraged. We think the deeper model may show different performance with the large amount of data. Second, the multi-stream highway network may also give different results on F0 trajectory prediction after separating the F0 modelling stream from the MGC one [22] .

Fine analysis and different methods for F0 modelling are also interesting. As [23] shows, perception of the intonation may not degrade significantly after stylising the F0 trajectory with straight lines. Thus, the training criterion of NN to fit the F0 trajectory exactly may be too harsh. Another question is that, NN systems trained with the mean-square-error criterion tend to predict the average of the target features in the training data [24]. For speech synthesis, the predicted average F0 pattern could result in a good objective measure when the distribution of the target F0 features is multi-modal. But it may be less preferred by the listeners than samples generated from either

[1] http://tonywangx.github.io



(a)

(b)

Figure 3: Samples of synthetic F0 trajectory on the M007 voice. Figure (a) compares the DBLSTM-RNN systems trained with different amount of training data. Figure (b) compares the DBLSTM-RNN, DNN and HMM trained using the full M007 data.

mode of the distribution. We need further analysis on the synthetic samples and NN systems.

5. Acknowledgements

This work was partially supported by EPSRC through Programme Grant EP/I031022/1 (NST) and EP/J002526/1 (CAF) and by the Core Research for Evolutional Science and Technology (CREST) from the Japan Science and Technology Agency (JST) (uDialogue project). Shinji Takaki was supported in part by the NAVER Labs.

6. References

[1] 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.

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

[3] Z.-H. Ling, L. Deng, and D. Yu, "Modeling spectral envelopes using restricted Boltzmann machines and deep belief networks for statistical parametric speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, pp. 2129–2139, 2013.

[4] Y. Fan, Y. Qian, F. Xie, and F. K. Soong, "TTS synthesis with bidirectional LSTM based recurrent neural networks," *INTERSPEECH-2014*, pp. 1964–1968, 2014.

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

[6] C. Tuerk and T. Robinson, "Speech synthesis using artificial neural networks trained on cepstral coefficients." in *EUROSPEECH*, 1993, pp. 1713–1716.

[7] O. Karaali, G. Corrigan, and I. Gerson, "Speech synthesis with neural networks," in *Proc. of World Congress on Neural Networks*, 1996, pp. 45–50.

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

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

[10] D. Amodei, R. Anubhai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, J. Chen, M. Chrzanowski, A. Coates, G. Diamos *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," *arXiv preprint arXiv:1512.02595*, 2015.

[11] X. Huang, J. Baker, and R. Reddy, "A Historical Perspective of Speech Recognition," *Commun. ACM*, vol. 57, no. 1, pp. 94–103, jan 2014. [Online]. Available: http://doi.acm.org/10.1145/2500887

[12] K. Wei, Y. Liu, K. Kirchhoff, C. Bartels, and J. Bilmes, "Submodular subset selection for large-scale speech training data," in *ICASSP-2014*. IEEE, 2014, pp. 3311–3315.

[13] Y. Wu, R. Zhang, and A. Rudnicky, "Data selection for speech recognition," in *ASRU-2007*. IEEE, 2007, pp. 562–565.

[14] H. Kawai, T. Toda, J. Ni, M. Tsuzaki, and K. Tokuda, "Ximera: A new tts from atr based on corpus-based technologies," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[15] H. Kawai, T. Toda, J. Yamagishi, T. Hirai, J. Ni, N. Nishizawa, M. Tsuzaki, and K. Tokuda, "Ximera: A concatenative speech synthesis system with large scale corpora," *IEICE Trans. Inf. Syst.(Japanese Edition)*, pp. 2688–2698, 2006.

[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] The HTS Working Group, "The Japanese TTS System "Open JTalk"," 2015. [Online]. Available: http://open-jtalk.sourceforge.net/

[18] T. Kudo, "MeCab: Yet Another Part-of-Speech and Morphological Analyzer." [Online]. Available: http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

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

[20] 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.

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

[22] X. Wang, S. Takaki, and J. Yamagishi, "Investigating very deep highway networks for parametric speech synthesis," in *SSW-9 (accepted)*, 2016.

[23] C. d'Alessandro and P. Mertens, "Automatic pitch contour stylization using a model of tonal perception," *Computer Speech & Language*, vol. 9, no. 3, pp. 257 – 288, 1995.

[24] C. M. Bishop, "Mixture density networks," Aston University, Tech. Rep., 2004.