An Investigation of Implementation and Performance Analysis of DNN Based Speech Synthesis System

Zhehuai Chen, Kai Yu
Key Lab. of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering
Department of Computer Science and Engineering, Shanghai Jiao Tong University
Email: chenzhehuai@outlook.com, kai.yu@sjtu.edu.cn

***Abstract*— Deep Neural Network (DNN), which can model hierarchical and complex relationship between input and output layer has recently been applied in speech synthesis. However, it is remained uncertain why DNN outperform traditional HMM-based synthesis. This paper describes several implementation details of DNN-based speech synthesis system and compares different impacting factors, e.g, F0 modeling method and adding BAP feature. DNN-based system are further investigated and in particular Continuous F0 HMM (CF-HMM) is taken as the baseline to compare with DNN-based system, as it has more similar input and output features with DNN-based system. Results show the ability of F0 modelling is similar between two systems, while CF-HMM system performs better. It seems that CF-HMM carefully strengthens the model by many technology, while using DNN to model F0 is still rough and needs more research. Another experiment shows that CF-HMM also does better in mcep modelling which needs to be further investigated.**

***Keywords*— Speech Synthesis, DNN, MSD-HMM, CF-HMM**

I. Introduction

Recently, Hidden Markov Model (HMM) based speech synthesis has become the most popular technology in the field [1] . In such system, fundamental frequency (F0), Mel-Cepstral spectral coefficients(Mcep) and Band aperiodical component (BAP) [2] are used as acoustic features of speech. To keep synchronization between spectral parameters andF0 parameters, they are modeled simultaneously by separate streams in a multistream HMM [3], which uses different state output probability distributions for modeling individual parts of the observation vector. In traditional model, multispace probability distributions [4] are used to model the F0 parameters, which use a continuous distribution for voiced frames and a discrete distribution for unvoiced frames. By switching the continuous and discrete space according to the space label associated with each observation, it can model F0 observation vector sequences in a multispace [5]. According to [6], Multi-space Probability Distribution HMM (MSD-HMM) has many disadvantages in F0 modeling, therefore, an improved Continuous F0 HMM (CF-HMM) [7] was proposed to solve them by continuous F0 modeling. Besides, many other methods to solve common over-smoothing problem in HMM-based speech synthesis have been proposed and work reasonably effectively in statistical parametric speech synthesis, e.g, parameter generation algorithm considering Global Varience [8], there are still some limitations [9]. In case of shortcomings, Deep Neural Network was applied as the generative model in speech synthesis system.

Deep Neural Network (DNN), which can model a hierarchical, intricate relationship between input and output layer, with a deep-layered structure, has recently been successfully applied in speech recognition [10]. As the inverse of such process, speech synthesis system with DNN as the generative model was built by a few research groups. Zen, et al. [9] analyzed the limitations of the conventional HMM-based approach and used DNN to overcome these limitations for speech synthesis. It is concluded that DNN-based system, which models the relationship between input linguistic features and their corresponding acoustic features, can outperform the HMM-based approach. Besides, Deep Belief Network (DBN) with stacked, Restricted Boltzmann Machines (RBMs) [10] is used to model joint distribution of linguistic and acoustic features for speech synthesis to reduce over-fitting for the discriminative fine-tuning phase by modeling the structure in the input data as generative pre-training and finding a region of the weight-space [11]. In addition, RBM is directly used to represent the distribution of the spectral envelopes at each HMM state and has been revealed that RBM is better than GMM-HMM which results in a better voice quality in RBM-based speech synthesis [12].

In these DNN-based methods, the key differences between them and the traditional HMM-based method includes model frameworks and its training stages, acoustic feature modeling and organization, and output feature parameter generation methods. With these differences, it is remained uncertain why DNN-based speech synthesis performs better than GMM-HMM speech synthesis:

- In [9], many implementation details haven't been discussed and compared carefully, which can inevitably impact the whole system performance, e.g., the training aspects of DNN [13], the input and output features design, their extraction and preprocessing.
- The previous analysis of DNN-based system performance mainly concerns about naturalness and simply compares the output features from two different system, which is far from thorough analysis of what DNN-based system can actually enhance.
- The comparison between MSD-HMM and DNN seems unreasonable because the output features of these two

This work was supported by the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning, the China NSFC project No. 61222208, JiangSu NSF project No. 201302060012 and MCaiA project.

978-1-4799-2186-7/14/$31.00 ©2014 IEEE

system are not the same (discrete F0 modeling v.s. continuous F0 modeling).

In the paper, implementation details of DNN-based speech synthesis system following framework of [9] is discussed and different methods are tested by comparison experiments. The performance gain and weakness of DNN-based system are further investigated and CF-HMM system [7] is taken as the baseline to compare with DNN-based system, because of its more similar input features and output features with DNN-based system.

The rest of the paper is arranged as follows. Section III outlines the differences between discontinuous and continuous F0 modelling approaches for HMM-based speech synthesis. Section II describes the detail of system implementation and its performance analysis. Section IV presents experiment results of performance gain and model weakness analysis. Finally, section V suggests future work and concludes the paper.

II. DNN-BASED SPEECH SYNTHESIS

A. Framework of DNN-based speech synthesis system

Deep Neural Network (DNN), which can model hierarchical structures between linguistic full context labels and acoustic waveform parameters, was applied in the speech synthesis system. Figure 1 shows a 3-hidden-layer Deep Neural Network, which is the most commonly used structure in this paper and the whole framework of a DNN-based speech synthesis system is similar to HMM-based system except this mapping between input and output features. In DNN-based speech synthesis system, rich contexts are used as input feature. The input features include linguistic binary answers of context information and numeric values of context number, position and duration, etc. [9]. All the linguistic values are packed into a long vector frame-by-frame as the input features. Then the input features are mapped to output features by a trained DNN using forward propagation. The output features are acoustic features like spectral envelope and fundamental frequency (f0). Their dynamic features [14] are also included in the output vector. Besides, a dim of output feature was used to label whether the current frame is voiced or unvoiced (V/UV). Input features and output features are time-aligned frame-by-frame by a MSD-HMM models [6] . The weights of DNN are trained by minimize the errors between the mapped output from a given input feature vector and the target output vector.

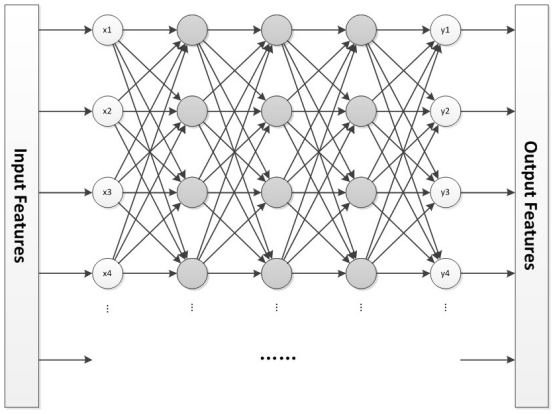

Fig. 1. An example of 3-hidden-layer Deep Neural Network

In synthesis, the input text is converted first into input feature vector through the text analysis and traditional HMM-based model is used to get duration [1], then input feature vectors are mapped to output vectors by a trained DNN using forward propagation. By setting the predicted output features from the DNN as mean vectors and pre-computing global variances of output features from all training data, the speech parameter generation algorithm [8] can generate smooth trajectories of speech parameter features which satistify both the statistics of static and dynamic features. Finally, a waveform synthesis module can output the synthesized waveform.

B. Implementation detail discussion

The implementation detail of DNN synthesis system hasn't been discussed in detail in [9]. However, nearly all components of the system can inevitably impact the whole performance. Therefore, this paper discusses several details below, in order to build a better performance DNN-based system.

1) Input Feature and its preprocessing: The input feature to represent linguistic information of each frame contains 2 parts of content:

- Binary features for categorical contexts, e.g., phone labels, POS labels of the current word, and TOBI labels. Each dim of such labels should be converted to a long vector to represent it. For example, first five and 26th dims of the linguistic features represent the phone information of the context. Since the size of phone set in our text analyser is 42, each dim of such labels is converted to a vector of 42 dims and the dim represents the corresponding phone label is set to value 1, while the other dims are set to value 0.
- Numerical features for the numerical contexts, e.g., the number of words in the phrase or the position of current frame in the current phone. Each dim of such labels can be converted to one dim of input features, after normalization.

Besides, the modeling unit of DNN is state, same as HMM-based system, so the state of the current full context label should also be taken into account. The state level information obtained from force-alignment include 5 different values, {S2, S3, S4, S5, S6}. Therefore, the state information is converted to a vector of 5 dims, with one dim set to value 1 to represent the current state and the other set to value 0. After data preprocessing, all the converted values and vectors are joint together to form a long vector frame by frame, containing all linguistic information and packed for DNN input layer.

2) Output Feature extraction and data preprocessing: The out acoustic feature includes 4 parts below:

- Spectral envelop parameters and their time derivatives. The first and second derivative of speech parameter vector sequence form the dynamic feature vectors for a smoother parameter generation. For better trained DNN performance, all three vector sequences (static, delta and delta-delta) should be normalized to zero mean and unity variance.
- Fundamental frequency (F0) parameters and their time derivatives. To model log F0 sequences by a DNN, the

continuous F0 modeling approach [15] was used. In this approach, F0 observations in unvoiced regions can be determined by 1-best selection or SPLINE interpolation [16]. Since SPLINE interpolation might lead to tighter variance which means a better trajectory modelling in V/UV boundary regions [6] , such method is used in feature extraction. All thress F0-related vector sequences should also be normalized after transforming to logarithmic form.

- V/UV index label. Because F0 is modelled in continuous stream, there's another dim needed to specify the V/UV state of each frame for parameter generation.
- Band aperiodical component (BAP) parameters and their time derivatives. Aperiodicity parameters [2] and F0 are used to generate the source signal for synthetic speech by controlling relative noise levels and the temporal envelope of the noise component of the mixed mode excitation signal. This mixed excitation model has been shown to give significant improvements in the quality of HMM-based synthesized speech [17] . Therefore, such acoustic features are also taken into DNN output features after normalization.

The output acoustic parameters are packed together as a long vector frame by frame and aligned with the input features, for the DNN output layer.

3) Training of DNN: The DNN is trained using stochastic gradient ascent algorithm with momentum to small minibatches of training cases [10] . One iteration of gradient descent updates the parameters W,b as follows:

$$W_{ij} = W_{ij} - \alpha \frac{\partial}{\partial W_{ij}} C(W,b) \quad (1)$$

$$b_i = b_i - \alpha \frac{\partial}{\partial b_i} C(W,b) \quad (2)$$

where $C(W,b)$ is the cost function and α is the learning rate predefined. Besides, it is more efficient to compute the derivatives on a small, random "minibatch" of training cases, rather than the whole training set, before updating the weights in proportion to the gradient [10]. This stochastic gradient descent method can be further improved by using a momentum coefficient ϵ [10], that smooths the gradient computed for minibatch below.

$$\Delta W_{ij}(t) = \epsilon \Delta W_{ij}(t-1) - \alpha \frac{\partial}{\partial W_{ij}(t)} C(W,b) \quad (3)$$

C(W,b) includes two kinds of criterion. Acoustic parameter determination is a regression problem. The commonly-used criterion in DNN training of that part is mean squared error. However, determining V/UV index is a classification problems. Therefore, Cross-entropy is used as the criterion of its training. In case of over-fitting in DNN training, 10% of the total data is chosen as the test set in the training process to measures the discrepancy between target vectors and the predicted output from updated model. According to [13] , random initialization of DNN model is chosen. Three hidden layers with 512 or 1024 nodes for each laysers is used as the framework of the DNN.

C. Performance analysis discussion

The DNN-based method is different with HMM-based method mainly in aspects below.

- Both GMM-HMM and DNN model the structures between linguistic labels and acoustic parameters for speech synthesis, while Deep Neural Network is a more long-span and highly-complex mapping and Hidden Markov Model is more shallow and carefully-designed. In training stage, HMM states and decision tree decompose the training data into small partitions and the model parameters are updated independently with the corresponding data, while all the weights of DNN are updated by looping through all training data [13].
- The most traditional GMM-HMM system models different kinds of features in different streams and models F0 stream in a discontinuous method (MSD-HMM). While DNN-based system models each dim of features in parallel and models F0 in a continuous method (SPLINE interpolation [15]). Besides, DNN-based system does normalization to feature data before model training.
- Both GMM-HMM system and DNN-based system use a variation of parameter generation algorithm [18] on the output feature parameters to synthesize speech waveform. However, while doing parameter generation, GMM-HMM system uses technologies to overcome over-smoothing problem, e.g, considering the Global Variance (GV [8]). And DNN-based system simply uses pre-computed mean and variance of all training data to de-normalize the output feature data.

With several differences between 2 kinds of systems, to analyze DNN-based system performance, it is common to compare the F0 and spectrum performances respectively with GMM-HMM system. It is concluded that DNN-based systems can outperform the HMM-based systems in [9]. However, the comparison seems to be inaccurate.

A multi-space probability distribution (MSD) HMM system is always used as a baseline version of HMM-based system in the previous comparison. But the continuous F0 modeling approach [15] is used in DNN-based system. In another word, it might be the continuous F0 modeling technology (e.g. SPLINE interpolation) but not DNN modelling power that leads to the better performance compared to MSD-HMM system.

In the paper, CF-HMM system [7] is taken as the baseline to compare with DNN-based system. Because of its more similar input features and output features with DNN-based system, the comparison seems to be more resonable and convincing. Besides, the more carefully-designed framework of model might bring about better results compared to MSD-HMM and even DNN.

III. Comparison of discontinuous and continuous F0 modelling approaches for HMM-based system

As described in section I, there are two different fundamental assumptions of the F0 observations:

- Discontinuous F0 modelling. Multi-space probability distribution HMM (MSD-HMM) is the most widely used solution for discontinuous F0 modelling [4]. It assumes discontinuous F0 and observable voicing labels. It can be derived that the state output distribution of MSD-HMM is:

$$p(o|s) = p(l, f_+|s) = \begin{cases} P(U|s) & \text{l=U} \\ P(V|s)N(f_+|s, V) & \text{l=V} \end{cases} \quad (4)$$

Where f_+ is the discontinuous F0 model and $N(\cdot)$ is referred to a Gaussian distribution. Due to the discontinuity, it is not convenient to calculate dynamic features of F0 at the boundary between voiced and unvoiced regions. This common implementation limits the power of HMM to model F0 trajectory.

- Continuous F0 modelling. Continuous F0 modelling is proposed to improve the F0 trajectory modelling. By generating real F0 values for unvoiced regions and assuming hidden voicing labels, the Continuous F0 modelling is obtained [7]. In continuous F0 modelling, F0 was determined by some continuous F0 modelling methods, e.g, SPLINE interpolation and N-best [15]. Then an independent data stream is introduced to explicitly model voicing labels, referred to as Continuous F0 modelling with Independent Voicing label and F0 value [6]. To strengthen the correlation between voicing label stream and F0 value stream, 2 streams was refined to only one stream used to simultaneously model both observable voicing labels and continuous F0 values, referred to as Continuous F0 modelling with Joint Voicing label and F0 value [7]. The state output distribution is

$$p(o|s) = p(l, f|s) = P(l|s)p(f|s, l) \quad (5)$$

Such model allows voicing labels to affect the forward-backward state alignment process, and it will naturally strengthen the voicing label modelling. It is revealed that this model can achieve best improvement in the naturalness of synthesised speech by continuous F0 modelling [7].

To sum up, continuous F0 modelling can bring about several advantages below [6] and is one of the state-of-art HMM-based speech synthesis system:

- Probability mass can be shared between voiced and unvoiced parts. Therefore, voiced observations near V/U boundaries from can be used in the estimation of the unvoiced distribution and vice versa. This affects the estimation accuracy near V/U boundaries and it makes the system robust to F0 extraction errors.
- The voicing label model is consistent at V/U boundaries. Because only one stream used in voicing labelling, it can be overcome that the unvoiced regions for the delta and delta-delta streams are in compatible with those for the static stream in MSD-HMM modelling.
- The redundant voicing parameters associated with the delta and delta-delta streams in MSD-HMM can be decreased. Thus, when the minimum description length (MDL) criterion [19] or any similar complexity metric is used to control the state clustering process, the less free parameters will result in more robust and accurate context-dependent F0 modelling.

IV. Experiments and Results

A. Experimental setup

The DNN-based system and HMM-based system described above have been evaluated on two CMU ARCTIC speech synthesis data sets [20]. A U.S. female English speaker, slt, and a U.S. male English speaker, awb, were used. Each data set contains recordings of the same 1132 phonetically balanced sentences totalling about 0.95 hours of speech per speaker. To obtain objective performance measures, 90% sentences from each data set were randomly selected as the training set for all experiments, and the remainder were used to form a test set and CV set.

HMM-based systems were built using a modified version of the HTS HMM speech synthesis toolkit version 2.0.1 [21]. DNN-based systems were built using a modified version of TNet [22] The speech features used were 24 Mel-Cepstral spectral coefficients, the logarithm of F0, and aperiodic components in five frequency bands (0 to 1, 1 to 2, 2 to 4, 4 to 6 and 6 to 8 KHz). All features were extracted using the STRAIGHT programme [23]. MDL-based state clustering [24] was performed for each stream to group the parameters of the context-dependent HMMs at state level. And DNN-based system also models state level mapping. All the duration of each state is obtained by the HMM-based system.

During the synthesis stage, global variance (GV) [8] is used in the speech parameter generation algorithm to reduce the well-known over-smoothing problem of HMM based speech synthesis and make DNN-based system more robust.

B. Implementation detail

Several implementation details are discussed in this section. Objective and subjective measures are used to evaluate the performance of systems with different implementations. Synthesis quality is measured objectively in terms of distortions between natural test utterances of the original speaker and the synthesized speech frame-by-frame. The objective measures are F0 distortion in the root mean squared error (RMSE, Hz) [6], voiced/unvoiced (V/U) swapping errors [6] and Mel-frequency cepstral distance (MSD) which calculates the absolute value of the difference between two mel-cepstral coefficients. The subjective measure includes mean opinion score (MOS) test and AB preference test each with 10 listeners participated in. All the tests include 5 sentences each with 2 voices synthesized from a female database (slt) and a male database (awb).

1) Structure of DNN and its training aspects: Different numbers of layers and different numbers of nodes in each hidden layer are tested in this paper. Table I shows the result that 512 or 1024 nodes for each 3 layers yield very similar performances in all three objective measures. And Table II, which tests system performance using 2-5 hidden layers each with 1024 nodes also gets similar result. Besides, if the hidden layer structure is too deep, the result can even be deteriorated because the hardness of model training grows with the deepness of structure. Besides, different sizes of mini-batch also yields similar performance in Table III. So 1024*3 structure with mini-batch=256 is used in the latter experiments.

node number	Female			Male		
	RMSE	VCE (%)	MSD	RMSE	VCE (%)	MSD
1024	12.81	6.59	0.21	13.44	4.69	0.17
512	12.35	6.36	0.20	13.79	4.92	0.17

TABLE I
OBJECTIVE MEASURES OF NUMBER OF NODES IN 3 HIDDEN LAYER

layer number	Female			Male		
	RMSE	VCE (%)	MSD	RMSE	VCE (%)	MSD
2	12.30	6.20	0.22	13.77	4.88	0.17
3	12.81	6.59	0.21	13.44	4.69	0.17
4	12.80	6.80	0.21	13.88	4.92	0.17
5	20.50	15.36	0.40	23.79	15.92	0.35

TABLE II

OBJECTIVE MEASURES OF NUMBER OF LAYERS EACH WITH 1024 NODES

mini-batch	Female			Male		
	RMSE	VCE (%)	MSD	RMSE	VCE (%)	MSD
128	12.90	6.85	0.21	13.67	4.82	0.17
256	12.81	6.59	0.21	13.44	4.69	0.17
512	12.70	6.82	0.21	13.52	4.85	0.17

TABLE III

OBJECTIVE MEASURES OF DIFFERENT SIZES OF MINI-BATCH

2) Continuous F0 modeling method: There are several methods to model the F0 in unvoiced region. In this paper, SPLINE interpolation and N-best [6] are compared in Table IV. Result shows that SPLINE interpolation method brings about better F0 distortion with a little worse V/UV errors. Because human beings might be more sensitive to the pitch distortion [6], SPLINE interpolation method is more suitable for DNN-based system.

F0 Model	Female		Male	
	RMSE	VCE (%)	RMSE	VCE (%)
Interpolation	12.40	6.27	13.26	4.96
N-best	12.50	6.10	14.05	4.39

TABLE IV

OBJECTIVE MEASURES OF DIFFERENT F0 MODELLING METHODS IN DNN

3) Adding BAP feature: In HMM-based system, band aperiodical component (BAP) is used to improve quality [2]. We test BAP performance gain in DNN-based system. Subjective measure experiment is also conducted, and Figure 2 shows the result. There's statiscally-significant better listener preference in system with BAP feature (p-values: 0.0000000016 for female and 0.00000041 for male at 95% confidence level).

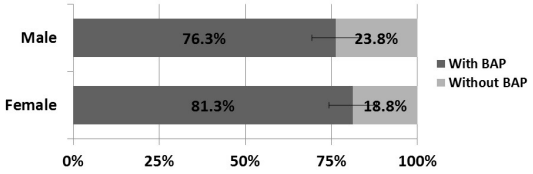

Fig. 2. subjective measures of adding BAP features into DNN-based system

C. Thorough performance comparison

1) MSD-HMM vs. CF-HMM vs. DNN: In this paper, 3 different synthesis systems (MSD-HMM, CF-HMM [7], DNN) are compared, Table V shows the objective measure result. It can be concluded that the F0 modeling performance of DNN-based system statistically-significant better than MSD-HMM system, but a little worse than CF-HMM system. The MSD of 3 system is almost the same. Figure 3 shows the MOS test result. There are 5 systems in the test set, including 2 reference sets, which one is natural speech and the other is vocoded speech to determine the effects of vocoder artifacts on the assessment [6]. The MOS result agrees with objective measures, which reveals that DNN-based system is better than MSD-HMM but still worse than CF-HMM.

To further compare the performance of HMM-based system and DNN-based system, AB preference test is conducted.

System model	Female			Male		
	RMSE	VCE (%)	MSD	RMSE	VCE (%)	MSD
MSD-HMM	16.02	5.24	0.20	15.11	3.52	0.18
CF-HMM	10.56	6.51	0.20	12.17	4.77	0.18
DNN	12.40	6.27	0.22	13.26	4.96	0.17

TABLE V

OBJECTIVE MEASURES OF DIFFERENT SPEECH SYNTHESIS SYSTEM

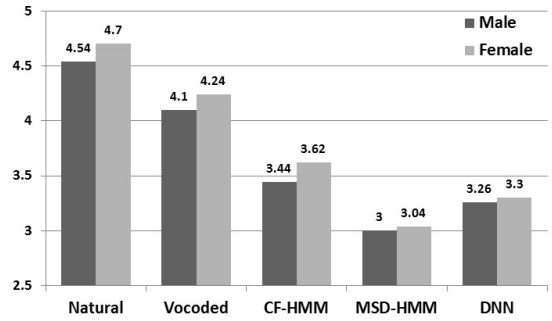

Fig. 3. subjective measures of different speech synthesis system

Figure 4 firstly compare the acoustic preference between MSD-HMM system and DNN-based system. Result shows that DNN-based system can be greatly more pleased to human beings and the preference of it is shown to be significant at 95% confidence level (p-values: 0.0011 for female and 0.000013 for male). Figure 5 compare the CF-HMM system and DNN-based system and shows that CF-HMM system can be a little better than DNN-based system but not significantly(p-value is 0.16). Summarizing results before, it can be concluded that, $DNN \approx CFHMM > MSDHMM$. Besides, Different sizes of database are tested in this paper to find out how much the performance can be gained with more training data. Table VI shows the result. In the test, 1 hour and 5 hours data from the same speaker are test as the training data for 3 systems respectively. It can be reveal that the performance of CF-HMM system gains largely with more training data, while other systems improve a little.

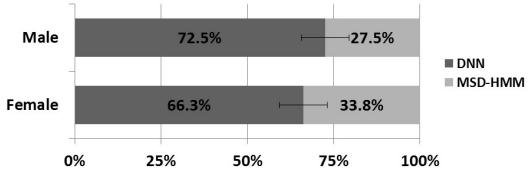

Fig. 4. Comparison between MSD-HMM and DNN modelling

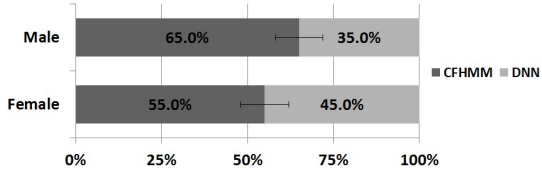

Fig. 5. Comparison between CF-HMM and DNN modelling

System model	1 hours			5 hours		
	RMSE	VCE (%)	MSD	RMSE	VCE (%)	MSD
MSD-HMM	28.25	6.71	0.22	25.00	5.01	0.20
CF-HMM	28.53	7.98	0.22	17.52	6.10	0.20
DNN	29.82	7.61	0.21	27.27	6.79	0.33

TABLE VI

OBJECTIVE MEASURES IN DIFFERENT SIZES OF TRAINING DATA

2) F0 modeling ability analysis: To further inspect whether DNN can better the F0 modelling performance, we take CF-HMM as baseline and directly test the F0 modelling result by providing both systems the mcep features extracted from

natural speech data. In other words, we synthesize speech with F0 generated from DNN/CF-HMM system and mcep extracted from real speech, and do tests on the synthetic speeches. Figure 6 shows the subjective measure result. It can be concluded that the ability of F0 modelling is similar between 2 systems, while CF-HMM system performs better (p-values is 0.83 at confidence level 95% which is not significant). We suspect that it is because that CF-HMM carefully strengthen the model by many technology [6] [7], while using DNN to model F0 is still rough and need more researches.

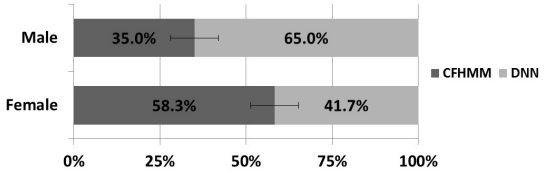

Fig. 6. Comparison of f0 modelling ability between CF-HMM and DNN

3) Spectrum modeling ability analysis: By similar method, we test the mcep modelling ability between CF-HMM and DNN. Figure 7 shows the result. It reveals that CF-HMM does better in mcep modelling and there's still long way for DNN-based system to go, especially in spectrum modelling.

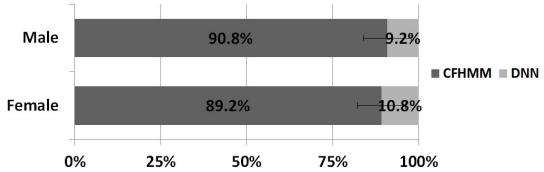

Fig. 7. Comparison of mcep modelling ability between CF-HMM and DNN

V. Conclusion and Future Work

This paper has described several implementation details of DNN-based speech synthesis system and compared different methods by experiments which can notably impact performance. The performance gain and weakness of DNN-based system are further investigated and CF-HMM system is taken as the baseline to compare with DNN-based system, Because of its more similar input features and output features with DNN-based system. Result shows the ability of F0 modelling is similar between 2 systems, while CF-HMM system performs better. The reason seems that CF-HMM carefully strengthen the model by many technology, while using DNN to model F0 is still rough and need more researches. Another experiment shows that CF-HMM does better in mcep modelling and there's still long way for DNN-based system to go, especially in spectrum modelling.

In summary, Deep Neural Network is a more long-span and highly-complex mapping while Hidden Markov Model is more shallow and carefully-designed. Currently, no enough evidence shows that the modeling ability of a hierarchical complex structure has outperformed that of a shallow but carefully-designed and optimized one. How we can analyze the modeling ability and proficiency between them and then realize these potentials is a topic for future investigation.

References

[1] T. YoshimuraÝ, K. TokudaÝ, T. MasukoÝÝ, T. KobayashiÝÝ, and T. KitamuraÝ, "Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis," 1999.

[2] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight." in *MAVEBA*, 2001, pp. 59–64.

[3] S. J. Young and S. Young, *The HTK hidden Markov model toolkit: Design and philosophy*. Citeseer, 1993.

[4] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling," in *International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 1999.

[5] 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.

[6] K. Yu and S. Young, "Continuous F0 Modeling for HMM Based Statistical Parametric Speech Synthesis," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 19, pp. 1071–1079, 2011.

[7] ——, "Joint modelling of voicing label and continuous F0 for HMM based speech synthesis," in *International Conference on Acoustics, Speech, and Signal Processing*, 2011, pp. 4572–4575.

[8] T. Tomoki and K. Tokuda, "A speech parameter generation algorithm considering global variance for hmm-based speech synthesis," *IEICE TRANSACTIONS on Information and Systems*, vol. 90, no. 5, pp. 816–824, 2007.

[9] H. Ze, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, May 2013, pp. 7962–7966.

[10] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[11] S. Kang, X. Qian, and H. Meng, "Multi-distribution deep belief network for speech synthesis," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8012–8016.

[12] Z.-H. Ling, L. Deng, and D. Yu, "Modeling spectral envelopes using restricted boltzmann machines for statistical parametric speech synthesis," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7825–7829.

[13] Q. Yao, F. Y, H. W, and S. Frank K, "On the training aspects of deep neural network (dnn) for parametric tts synthesis," 2014.

[14] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, pp. 52–59, 1986.

[15] K. Yu, T. Toda, M. Gasic, S. Keizer, F. Mairesse, B. Thomson, and S. Young, "Probablistic modelling of F0 in unvoiced regions in HMM based speech synthesis," in *International Conference on Acoustics, Speech, and Signal Processing*, 2009, pp. 3773–3776.

[16] A. A. Privalov, "Convergence of cubic interpolation splines to a continuous function," *Mathematical Notes*, vol. 25, pp. 349–359, 1979.

[17] H. Zen, T. Toda, M. Nakamura, and K. Tokuda, "Details of the Nitech HMM-Based Speech Synthesis System for the Blizzard Challenge 2005," *Ieice Transactions*, vol. 90-D, pp. 325–333, 2007.

[18] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for hmm-based speech synthesis," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, vol. 3. IEEE, 2000, pp. 1315–1318.

[19] K. Shinoda and T. Watanabe, "Acoustic modeling based on the MDL principle for speech recognition," 1997.

[20] J. Kominek and A. W. Black, "CMU ARCTIC databases for speech synthesis."

[21] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black, and K. Tokuda, "The hmm-based speech synthesis system (hts) version 2.0," in *Proc. of Sixth ISCA Workshop on Speech Synthesis*, 2007, pp. 294–299.

[22] K. Vesely, L. Burget, and F. Grzl, *Parallel Training of Neural Networks for Speech Recognition*, 2010.

[23] H. Kawahara, I. Masuda-katsuse, and A. D. Cheveign, "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.

[24] S. J. Young, J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proceedings of the workshop on Human Language Technology*, 1994, pp. 307–312.