An Investigation of Context Clustering for Statistical Speech Synthesis with Deep Neural Network

Bo Chen, Zhehuai Chen, Jiachen Xu, Kai Yu
Key Laboratory of Shanghai Education Commission for
Intelligent Interaction and Cognitive Engineering
SpeechLab, Department of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai, China
{bobmilk, chenzhehuai, msf, kai.yu}@sjtu.edu.cn

Abstract

The state-of-the-art DNN speech synthesis system directly maps linguistic input to acoustic output and voice quality improvement over the conventional MSD-GMM-HMM synthesis system has been reported. DNN-based speech synthesis system does not require context clustering as in GMM-HMM systems and this was believed to be the main advantage and contributor to performance improvement. Our previous work has demonstrated that F0 interpolation, rather than context clustering, is the actual contributor for performance improvement. However, it remains unknown whether the use of unclustered context is a beneficial characteristic of DNN-based synthesis or not. In this paper, this issue is investigated in detail. Decision tree clustered contexts are used as linguistic input for DNN and compared to unclustered context input. A novel approach for inputting context clusters is proposed. Here, the decision tree question indicators are used as input instead of the clustered contexts. Experiments showed that DNN with clustered contexts significantly outperformed DNN with unclustered contexts and the proposed question indicator input approach obtained the best performance. The investigation of this paper reveals the limitation of DNN-based speech synthesis and implies that context clustering is also an important issue for DNN-based speech synthesis with limited training data.

Index Terms: statistical parametric speech synthesis, hidden Markov model, deep neural network, context clustering

1. 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 and F0 parameters, they are modelled simultaneously by separate streams in a multi-stream HMM [3], which uses different state output probability distributions for modelling individual parts of the observation vectors.

Deep Neural Network (DNN), which can model a hierarchical, intricate relationship between input and output layer , has recently been successfully applied in speech recognition[4]. As the inverse of such process, speech synthesis system with DNN as the generative model was built by a few research groups. Deep Belief Network (DBN) with stacked, Restricted Boltzmann Machines (RBMs) [4] 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 modelling the structure in the input data as generative pre-training and finding a region of the weight-space [5]. 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 [6]. Deep neural network is also used to model the conditional probability of the spectral differences between natural and synthetic speech [7].

Zen, et al. [8] analyzed the limitations of the decision tree-based system such as inefficiency in expressing complex context dependencies, fragmenting the training data, and completely ignoring linguistic input features which did not appear in decision trees, and used DNN to overcome these limitations with experiments on a corpus with 33 000 utterances. The two main differences between DNN system [8] and conventional MSD-GMM-HMM system are continuous F0 modelling (CF) [9] and implicit context clustering. Our previous work [10] has demonstrated that F0 interpolation, rather than context clustering, is the actual contributor for performance improvement. Result shows that the ability of F0 modelling is similar between DNN and CF-HMM, while CF-HMM system performs better. Qian, et al. [11] has also examined the aspects of DNN speech synthesis with a moderate size corpus of 5 hours which is more commonly used for parametric speech synthesis training, and suggested that the benefit is very likely from distinctive advantages of DNN, such as DNN is efficient and effective in representing high dimensional and correlated features and modelling highly complex mapping function in a compact manner. However, it remains unknown whether the use of unclustered context is a beneficial characteristic of DNN-based synthesis or not.

To investigate the issue, decision tree clustered contexts are used as linguistic input for DNN and compared to unclustered context input. This paper will show that decision-tree based context clustering still has its importance in DNN speech synthesis. In addition, a novel approach for inputting context clusters is proposed that the decision tree question indicators are used as input instead of the clustered contexts, and achieve the best performance.

The paper is organised as follows. Section II describes the detail of originally proposed DNN framework. Section III describes our methods to apply context clustering in DNN scene. Experiment details of different systems with evaluations are described in Section IV. Finally, Section V concludes the paper.

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 and JiangSu NSF project No. 201302060012 .

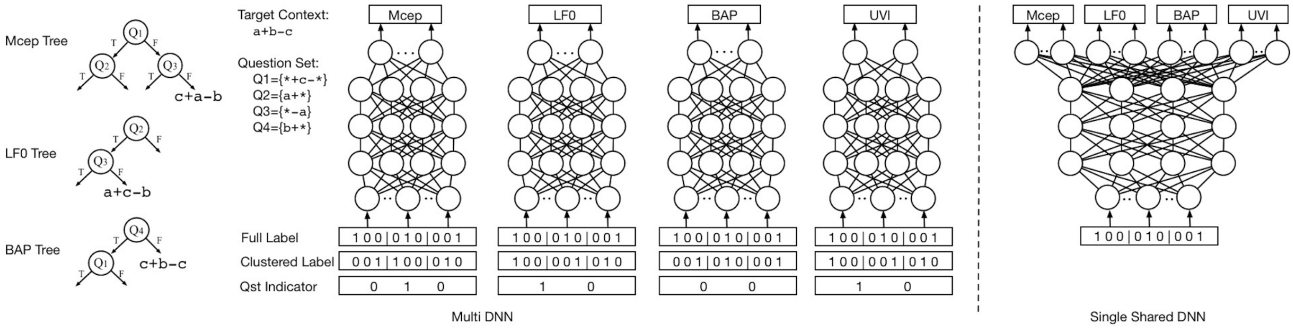

Figure 1: *Structure of the Multi DNN and Single Shared DNN*

2. DNN-Based Speech Synthesis

2.1. Framework of DNN-based speech synthesis system

First, the conventional HMM-based parametric speech synthesis method is briefly reviewed. The F0 and spectral parameters are extracted from the waveforms contained in the training set. Then a set of context-dependent HMMs are estimated to maximize the likelihood function for the training acoustic features. In DNN-based speech synthesis system, rich contexts are used as input features. The input features include linguistic binary answers of context information and numeric values of context number, position, duration etc. [8]. All the linguistic values are packed into a long vector frame-by-frame as the input features.

2.2. Input Feature and its preprocessing

The input feature of each frame contains 2 parts. The first part consists of binary features for categorical contexts, e.g. phone labels and POS labels of the current word. Each dimension of such labels will be converted to a long vector to represent it. For example, first 5 dimensions of the context labels represent the quinphone information. Since the size of phone set in our text analyser is 52(including past phone "nil" before the first centre phone "sil"), each dimension of such labels is converted to a vector of 52 dimensions and the dimension representing the corresponding phone label is set to value 1, while the other dimensions are set to value 0. The second part consists of numerical features for the numerical contexts, e.g. the number of words in the phrase or the position of the current frame in the current phone. Each dimension of such labels can be converted to one dimension of input features, after normalization.

Besides, the modelling unit of DNN is state, same as HMM-based system, thus the state of the current full context label should also be taken into account. The state level information obtained from force-alignment includes 5 different values, {S2, S3, S4, S5, S6}. Therefore, the state information is converted to a vector of 5 dimensions, with one dimension set to value 1 to represent the current state and the other set to value 0.

2.3. Output Feature extraction and data preprocessing

The output acoustic features include 4 parts below. `Mcep`: 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. `LF0`: Fundamental frequency parameters and their time derivatives via continuous F0 modelling [12]. F0 observations in unvoiced regions can be determined by 1-best selection or SPLINE interpolation [13]. `UVI`: Binary voice/unvoice index label. Since F0 is modelled in continuous stream, there is another dimension needed to specify the V/UV. `BAP`: Band aperiodical component parameters [2] and their time derivatives.

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. For better trained DNN performance, each dimension of all three acoustic sequences (static, delta and delta-delta) of `Mcep`, `LF0` and `BAP` should be normalized to zero mean and unity variance except `UVI`.

2.4. Training of DNN

The DNN is trained using stochastic gradient ascent algorithm with momentum to small mini-batches of training cases [4]. In most of the trainings, the mean square error (MSE) function

$$\mathcal{L}_{\texttt{mse}}(\theta) = ||\hat{\mathbf{o}} - \mathbf{o}||_2^2 \tag{1}$$

is used as criteria, where $\hat{\mathbf{o}}$ is the predicted observation and θ is the DNN parameters set. Besides, cross entropy (CE) function

$$\mathcal{L}_{\texttt{ce}}(\theta) = -\sum_s d_s \log P(s|\mathbf{v}, \theta) \tag{2}$$

is also used in the DNN training of this paper that the output of the DNN are softmax posterior probabilities $P(s|\mathbf{v})$ with respect to s indicating the voice/unvoice index, where $\mathbf{v}$ is the input features, θ is the set of parameters of DNN, d_s is 1 for correctly classified `UVI` and 0 for wrongly classified `UVI`.

3. Context Clustering for DNN

In conventional MSD-GMM-HMM synthesis systems, acoustic features are modelled in separate streams. For each stream, individual decision trees are built for context clustering. To import the context clustering into DNN synthesis system, the clustering information of different streams should be considered separately. Hence, a Multi-DNN structure is introduced here. Figure 1 shows the structure of a separate Multi-DNN and a shared Single-DNN which is the originally proposed structure [8]. In the Multi-DNN structure, each network is corresponding to one of the 4 types of acoustic features `Mcep`, `LF0`, `BAP` and `UVI`, that the output layer only consists of the related acoustic features. 3 different types of input layers are shown in the Multi-DNN structure. Input layer features from top to bottom are unclustered full linguistic features, clustered full linguistic features and the clustered question indicators which are used by a novel proposed method. And an example of simplified triphone case is illustrated in Figure 1.

3.1. Clustered Context Label

Full linguistic features are also used as input features here. However, the original full context labels are transferred to some selected ones via a state level mapping built from the decision tree based context clustering in the MSD-GMM-HMM scenes. `Mcep`, `BAP` and `LF0` decision trees are built after MSD-GMM-HMM training. A full context label v is randomly selected from each state cluster. V denotes the full context label set that for each component $v' \in V$, v' will arrived at the same cluster as v passing the decision tree. v' is then mapped to v that the input features of v' in this approach are exactly the same as the input features of v in the originally proposed method at that state. Hence, for each network, all the possible input features are the full linguistic features from the selected full context labels.

3.2. Clustered Question Indicators

The input features are binary vectors built via the question set Q used in constructing decision trees in MSD-GMM-HMM training scene. Each question in Q consists of several regular expressions that if the target full context label matches any of the regular expressions, the answer of the question is `True`, otherwise `False`. The input features are presented by the answers that if the answer of a question is `True`, the dimension corresponding to the question is set 1, otherwise 0. Zen has mentioned that they tried to encode numerical features to binary ones by applying questions in their preliminary experiment, but directly using numerical features worked better [8]. Different from the encoding attempt, the key method of the indicators presentation is to transfer the input features from naive linguistic domain to a structured domain. Most importantly, the question sets are selected for different types of acoustic features that only the questions consisted in the corresponding decision trees are selected. Still, the state information is included in 5 dimension indicators.

4. Experiment

4.1. System setting

The experiments were established on a Chinese corpus recorded from a Chinese male speaker "lcxinm". The corpus contained recordings of 3807 phonetically balanced sentences with about 5 hours related speech. TOBI labels were not included in this corpus. 3607 random chosen sentences were used as training set, while another 100 sentences were used for cross validation and the rest 100 sentences were used for objective evaluation. A MSD-GMM-HMM system was build using a modified version of the HTS HMM speech synthesis toolkit [14] version 2.1.1 to align the audio data at state level. The question set used in decision tree based context clustering consisted of 1496 different questions. DNN-based systems were built using a modified version of TNet [15]. According to [11], DNN with 3 hidden layers, each with 1024 nodes, was used as the basic structure. The acoustic features 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 [16]. MDL-based state clustering [17] was performed for each stream to group the parameters of the context-dependent HMMs at state level. The DNN-based system also modelled state level mapping. The duration of each state was obtained by the HMM-based system. The networks were initialized by a RBM with 3 hidden layers while the parameters between the last hidden layer and the output layer were randomly initialized. The RBM was trained by modified linguistic features that each state of a full context label was assigned only one frame to make each label have equal contribution. Global variance (GV) [18] was 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.

4.2. Experiment setting

4.2.1. Unclustered Contexts System

Unclustered contexts system was based on the original proposed DNN synthesis method from Zen's group [8]. Single shared DNN was used in this system with MSE function as the training criteria. The input layer consisted of 452 dimension full linguistic features with 5 dimension state indicators while the output layer consisted of 94 dimension acoustic features.

4.2.2. Unclustered Contexts Separate System

To confirm that the context clustering, rather than the Multi-DNN structure, mainly contributed to the performance, unclustered contexts separate system was trained in Multi-DNN with 4 separate networks that each of them was related to one of the 4 types of acoustic features `Mcep`, `BAP` and `LF0` and `UVI`. Same input and hidden layer configuration was used as the unclustered contexts system, while the output layers only consisted of the parts corresponding to the acoustic feature type. It should be noted that, the individual `UVI` network turned out to be a classification problem. Hence, 2 dimension output features, indicating "is voice" and "is unvoice", were used and the network was trained with the cross entropy criteria. And the other Multi-DNN systems in this paper followed this `UVI` output setting.

4.2.3. Clustered Contexts System

To examine the performance of the clustered contexts methods, clustered contexts system was trained in Multi-DNN. The input features, which had the same form as the unclustered context system, were generated from the clustered context label via the decision tree based mapping at state level for each stream. The output features only consisted of the parts corresponding to the acoustic feature type. Note that there was no decision tree of `UVI` in conventional MSD-GMM-HMM system, hence the decision tree of `LF0` was used instead in the `UVI` network.

4.2.4. Clustered Question Indicator System

To examine the performance of the clustered question indicators method, the clustered question indicators system was trained in Multi-DNN with 4 separate networks. Input features were generated from the clustered question indicators method. The question set used in each separate network contained the questions appearing in the corresponding decision trees after the state clustering stage in MSD-GMM-HMM training, which was a subset of the complete question set. The number of question indicators of `Mcep`, `BAP`, `LF0` and `UVI` used in the experiments were 662, 639, 1017 and 1017, while 1496 questions were designed in the complete question set. Input features consisted of the question indicators combined with the 5 dimension state indicators, while the output features only consisted of the parts corresponding to the acoustic feature type.

4.3. Objective Evaluation

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 root mean squared error (RMSE, Hz), voiced/unvoiced (V/U) classification errors (VCE) and Mel-frequency cepstral distance (MCD) which calculates the absolute value of the difference between two mel-cepstral coefficients. 100 sentences from the corpus were used in the objective evaluation. It can be seen from Table 1 that clustered contexts system has great improvement on VCE, but there is no obvious difference among systems with unclustered contexts and the indicator approach. Since Multi-DNN system with unclustered contexts didn't outperform Single-DNN system, both the unclustered systems were included in the subjective evaluations.

Table 1: *Objective evaluation on different systems.*

Context	System	RMSE	VCE(%)	MCD
Unclustered	Single-DNN	20.65	7.51	0.21
	Multi-DNN	21.15	7.90	0.21
Clustered	Context	**19.59**	**6.24**	0.21
	Indicators	20.26	7.82	0.21

4.4. Subjective Evaluation

The performance of these systems was investigated further by the following subjective evaluations. Each test consisted of 20 sentences and was provided to 20 Chinese participates.

4.4.1. Test Sentences Selection

The motivation of the sentence selection was to confirm that the clustered methods have better performance on dissimilar sentences and didn't drop quality on normal sentences.

A sentence pool with about 700 sentences was constructed from some famous Chinese proses and latest news from a large news website to make it less similar to the training sentences. Also, we use a objective way to select the most dissimilar sentences from the pool. Full context labels were generated from each sentence. The Euclid distance was calculated between a test label from the pool and every label appeared in the training label list. The smallest distance was denoted as the distance from the test label to the training set. The largest distance of all the labels of a sentence was denoted as the distance from the sentence to the training set. The distances from testing sentences to the training set were sorted from high to low. 10 sentences were selected from the top 20 as the test sentences of subjective evaluation. It could be assumed that these sentences were more dissimilar to the training set. Another 10 sentences were randomly selected from the rest sentences in the pool as test sentences to confirm the quality won't drop on normal contexts label. Since there were no audio files of these sentences from the speaker, objective test on these sentences was not held.

4.4.2. Mean Opinion Score

A mean opinion score (MOS) test was held among the 4 systems. 20 waves were synthesized from the 20 selected sentences by each of the systems. 4 audio files from each sentence were presented to listeners at random order. Every sentence appeared only once in the test. The listeners were asked to score each audio from 1-5 under the guidance that 1 refers to pretty bad that a lot phones are not clear; 2 refers to bad that some phones are not clear; 3 refers to OK that the phones are all clear, but some are slightly wrong; 4 refers to the phones are all correct and clear, but the whole quality is not good enough, 5 refers to the phones are all correct and clear while the whole quality is good enough. It can be seen from Figure 2 that the score of systems with clustered contexts outperform systems with unclustered contexts and the mean score of system with clustered indicators is much higher than the rest. It shows that the system with clustered indicators has much stronger ability to model the relationship between linguistic features and acoustic features, that performs better on contexts from out-corpus sentences.

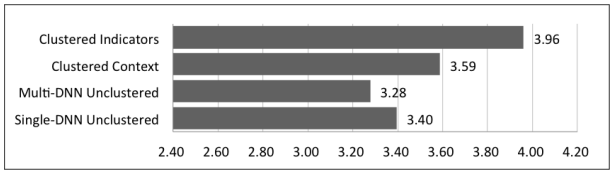

Figure 2: *Mean Opinion Score of different systems*

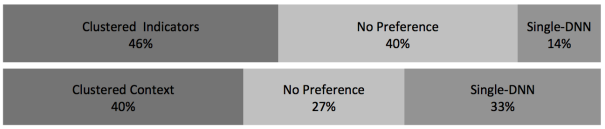

Figure 3: *Clustered Systems v.s. Unclusted Systems*

4.4.3. Preference Tests

Since the Single-DNN outperforms Multi-DNN with unclustered contexts in objective evaluation and MOS test, the preference tests were held between Single-DNN system and Multi-DNN systems with clustered information to verify that the clustered approaches significantly improved the performance. Each of the tests is designed as follows. 20 waves were synthesised from the 20 selected sentences by each system. For each sentence, 2 audio files were presented to each listener at random order. The listeners could listen to the audios multiple times and select from the 3 choices: the first is better, the second is better or neutral. Each sentence appeared twice in a test. Every listener was provided $20\times2\times2 = 80$ waves. $20\times2\times20 = 800$ preferences were gathered in all in one test. The neutral choices is divided in half to each preference choice to compute the confidence score. The result in Figure 3 shows that the systems with clustered indicators (p-value$<$1e-9) and clustered contexts (p-value=0.018) are significantly better than system with unclustered contexts. We believe that context clustering still has its advantage in DNN speech synthesis with limited data.

5. Conclusion

This paper has described the investigation of whether context clustering is a beneficial characteristic of DNN-based speech synthesis with limited data. The method of applying context clustering into DNN was proposed. Also, we proposed a novel approach for inputting context cluster which is the clustered question indicators. Experiments have been conducted for different methods with objective and subjective evaluations. Result showed that clustered systems outperformed the unclustered systems significantly and the clustered question indicators system obtained the best performance. In summary, context clustering is an important issue for DNN-based speech synthesis with limited data. We feel optimistic that structured DNN will achieve great improvement in the future.

6. References

[1] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "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*. University of Cambridge, Department of Engineering, 1993.

[4] 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.

[5] 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.

[6] 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.

[7] L.-H. Chen, T. Raitio, C. Valentini-Botinhao, J. Yamagishi, and Z.-H. Ling, "Dnn-based stochastic postfilter for hmm-based speech synthesis," 2014.

[8] 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*. IEEE, 2013, pp. 7962–7966.

[9] K. Yu and S. Young, "Continuous f0 modeling for hmm based statistical parametric speech synthesis," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 5, pp. 1071–1079, 2011.

[10] Z. Chen and K. Yu, "An investigation of implementation and performance analysis of dnn based speech synthesis system," in *Signal Processing (ICSP), 2014 12th International Conference on*. IEEE, 2014, pp. 577–582.

[11] Y. Qian, Y. Fan, W. Hu, and F. K. Soong, "On the training aspects of deep neural network (dnn) for parametric tts synthesis," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 3829–3833.

[12] 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 *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 3773–3776.

[13] A. A. Privalov, "Convergence of cubic interpolation splines to a continuous function," *Mathematical Notes*, vol. 25, no. 5, pp. 349–359, 1979.

[14] 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.

[15] K. Veselỳ, L. Burget, and F. Grézl, "Parallel training of neural networks for speech recognition," in *Text, Speech and Dialogue*. Springer, 2010, pp. 439–446.

[16] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigné, "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, no. 3, pp. 187–207, 1999.

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

[18] 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.