Overview of NIT HMM-based speech synthesis system for Blizzard Challenge 2011

Kei Hashimoto, Shinji Takaki, Keiichiro Oura, and Keiichi Tokuda

Department of Scientific and Engineering Simulation, Nagoya Institute of Technology, Nagoya, JAPAN

Abstract

This paper describes a hidden Markov model (HMM) based speech synthesis system developed for the Blizzard Challenge 2011. In the Blizzard Challenge 2011, we focused on the training algorithm for HMM-based speech synthesis systems. To alleviate the local maxima problems in the maximum likelihood estimation, we apply the deterministic annealing expectation maximization (DAEM) algorithm for training HMMs. By using the DAEM algorithm, the reliable acoustic model parameters can be estimated. In addition, we apply stepwise model selection to the model training. The decision tree based context clustering is used as model selection in HMM-based speech synthesis. By using the stepwise model selection method, decision trees are gradually changed from small trees into large trees for estimating reliable acoustic models. Subjective evaluation results show that the system synthesized the high intelligible speech.

Index Terms: speech synthesis, hidden Markov model, deterministic annealing, model structure

1. Introduction

A statistical parametric speech synthesis system based on hidden Markov models (HMMs) was recently developed. In HMM-based speech synthesis, the spectrum, excitation, and duration of speech are simultaneously modeled by HMMs, and speech parameter sequences are generated from the HMMs themselves [1]. Compared to other synthesis methods, this method has several advantages, 1) under its statistical training framework, it can learn statistical properties of speakers, speaking styles [2], emotions [3], etc,, from the speech corpus; 2) many techniques developed for HMM-based speech recognition can be applied to speech synthesis [4, 5]; 3) voice characteristics of synthesized speech can be easily controlled by modifying acoustic statistics of HMMs [6, 7].

Since acoustic models affect the quality of synthesized speech, the model estimation is one of the most important problem in statistical parametric speech synthesis. Therefore, the appropriate training algorithm is required to estimate the reliable model parameters. The maximum likelihood (ML) criterion has typically been used for training HMMs . The ML criterion guarantees that the ML estimates approach the true values of the parameters. The expectation maximization (EM) algorithm [8] is used to estimate the model parameters maximizing the likelihood for given training data. The EM algorithm provides a simple iterative procedure to obtain approximate the ML estimates of parameters. However, the EM algorithm often suffers from the local maxima problem because it is a hill-climbing approach. To relax this problem, the deterministic annealing EM (DAEM) algorithm has been proposed [9]. In the DAEM algorithm, the problem of maximizing the log-likelihood is reformulated as the problem of minimizing the thermodynamic free energy. The posterior distribution derived in the DAEM algorithm includes a "temperature" parameter which controls the influence of unreliable model parameters. It has been reported that the DAEM algorithm is effective for HMM-based speech recognition [10].

In addition, since model structures affect the model parameter estimation, the model parameters are accurately estimated by selecting appropriate model structures. In HMM-based speech synthesis, the model structures are selected by the decision tree based context clustering [11]. This method constructs a model parameter tying structure which can assign a sufficient amount of training data to each HMM state. In the decision tree based context clustering, it is typically assumed that the state occupancies are not changed by constructed model structures. By using this assumption, the context clustering is efficiently performed. However, the state occupancies are practically changed before and after context clustering. Especially when the model structures are significantly changed, the state occupancies are also significantly changed and the assumption affects the model selection. Once the inappropriate model structures are selected, the model parameters are updated under the inappropriate model structures. As a result, the estimated model parameters are not reliable. To alleviate this problem, we apply stepwise model selection to the training part. In the training procedures using step wise model selection, model structures were gradually changed from small structures into large structures. By gradually changing decision trees to larger in the training procedures, the differences before and after context clustering become small, and the joint optimization of model structures and model parameters are performed. As a result, the reliable acoustic models are obtained.

The rest of this paper is organized as follows. Section 2 describes our base speech synthesis system. Section 3 and 4 introduce new features of our system for the Blizzard Challenge 2011. Subjective listening test results are presented in Section 5. Concluding remarks and future work are presented in the final section.

2. Base system

2.1. HMM-based speech synthesis system

Figure 1 overviews a HMM-based speech synthesis system. It consists of training and synthesis parts.

The training part is similar to that used in speech recognition. The main difference is that both spectrum (e.g., mel-cepstral coefficients and their dynamic features) and excitation (e.g., log F_0 and its dynamic features) parameters are extracted from a speech database and modeled by HMMs. Although the

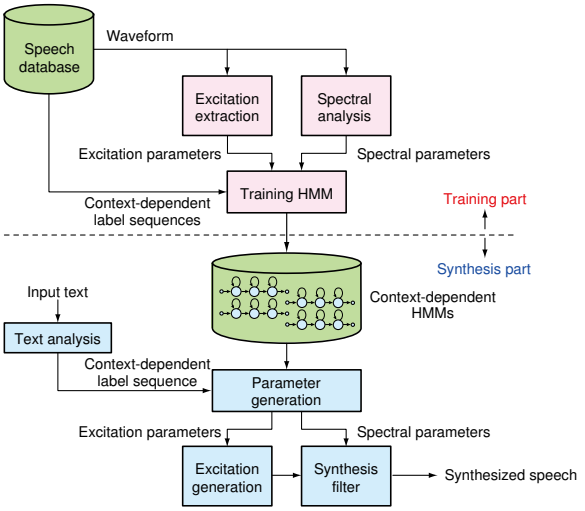

Figure 1: Overview of HMM-based speech synthesis system.

spectrum part can be modeled by continuous HMM, the F_0 part cannot be modeled by continuous or discrete HMM because the observation sequence of F_0 is composed of a one-dimensional continuous value and discrete symbol which represents unvoiced. To model such observation sequence, multi-space probability distributions (MSDs) [12] are used for state-output distributions.

The synthesis part does the inverse operation of speech recognition. First, an arbitrarily given text to be synthesized is converted to a context-dependent label sequence, and then a sentence HMM is constructed by concatenating the context-dependent HMMs according to the label sequence. Second, state durations of the sentence HMM are determined based on the state-duration distributions. Third, the speech parameter generation algorithm generates sequences of spectral and excitation parameters that maximize their output probabilities under the constraints between static and dynamic features [13]. Finally, a speech waveform is synthesized directly from the generated spectral and excitation parameters using a speech synthesis filter. The most attractive part of this system is that voice characteristics, speaking styles, or emotions can easily be modified by transforming HMM parameters using various techniques, such as adaptation [5], interpolation [14], or eigenvoices [15].

2.2. Hidden semi-Markov model

In HMM-based speech synthesis, rhythm and tempo are controlled by state duration probability distributions. One of major limitations of HMMs is that they do not provide an adequate representation of the temporal structure of speech. This is because state duration probabilities decrease exponentially with time. To overcome this problem, the hidden semi-Markov model (HSMM) based speech synthesis framework [4] was used in our system. This framework introduces an HSMM, which is an HMM with explicit state duration probability distributions, into not only the synthesis part but also the training part of the HMM-based speech synthesis system. It makes possible to estimate state output and duration probability distributions simultaneously. The effectiveness of the HSMM-based approach has been reported in [4].

2.3. STRAIGHT vocoding

As a high-quality speech vocoding method, we use STRAIGHT, which is a vocoder type algorithm proposed by Kawahara *et al.* [16]. It consists of three main components; F_0 extraction, spectral and aperiodic analysis, and speech synthesis.

The STRAIGHT automatically extract F_0 with fixed-point analysis [17]. Using the extracted F_0, we use the STRAIGHT method to perform pitch-adaptive spectral analysis combined with a surface reconstruction method in the time-frequency domain to remove signal periodicity. As a spectral parameter, we use the 0th through 49th mel-cepstral coefficients to which the smoothed spectrum analyzed by the STRAIGHT is converted. An aperiodicity measure in the frequency domain [18] is also extracted. As a parameter for constructing a mixed excitation sources in speech synthesis, average values of the aperiodicity measures on 26 frequency bands are used.

2.4. Parameter generation algorithm considering global variance

The HMM-based speech synthesis method generates speech parameters from the HMMs directly, so that an output probability of the parameter is maximized under a constraint on an explicit relationship between static and dynamic features. Consequently, a smoothed parameter trajectory is generated but it is excessively smoothed due to the statistical processing. Therefore, the synthesized speech using over-smoothed parameters sounds muffled. To reduce this effect, we applied a parameter generation algorithm considering global variance (GV) of the generated parameters [19] to both spectral and F_0 parameter generation processes.

One GV is calculated from a parameter sequence over the entire of one utterance. It should be noted that only voiced frames are used for calculating GV of F_0 parameters. The probability density on GV is modeled using a Gaussian distribution with a diagonal covariance matrix. In parameter generation, first a parameter trajectory is generated with the speech parameter generation algorithm. Then, the generated trajectory is converted, so that its GV is equal to a mean of the Gaussian distribution. Using this converted trajectory as an initial value, the parameter trajectory is calculated iteratively to maximize a likelihood function with the Newton-Raphson method. This likelihood function consists of the output probability of the parameter sequence and that of its GV.

In order to improve the estimation accuracy of GV models, we use the GV features calculated from only speech region excluding silence and pause regions and estimate the context-dependent GV models instead of a single global GV model. The silence and pause regions are determined by the automatic phone aligner using HSMMs [20] included in the latest HTS.

3. Deterministic annealing EM algorithm in parameter estimation

3.1. EM algorithm

The maximum likelihood criterion has typically been used to train HMMs in HMM-based speech synthesis. In the ML criterion, the optimal model parameters are estimated by maximizing the likelihood for give training data as follows.

$$\begin{aligned} \boldsymbol{\Lambda}_{\mathrm{ML}} &= \arg\max_{\boldsymbol{\Lambda}} \mathcal{L}(\boldsymbol{\Lambda}) \\ &= \arg\max_{\boldsymbol{\Lambda}} \log \sum_{\boldsymbol{q}} P(\boldsymbol{o}, \boldsymbol{q} \mid \boldsymbol{\Lambda}) \end{aligned} \tag{1}$$

where $\boldsymbol{o} = (\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_T)$ and $\boldsymbol{q} = (q_1, q_2, \ldots, q_T)$ are respectively the observation and state sequences, and $\boldsymbol{\Lambda}$ is a set of model parameters. However, it is difficult to obtain the model parameters $\boldsymbol{\Lambda}_{\mathrm{ML}}$ analytically. To overcome this problem, the expectation maximization (EM) algorithm [8] is used in HMM-based speech synthesis. The EM algorithm provides a simple iterative procedure to obtain $\boldsymbol{\Lambda}_{\mathrm{ML}}$.

The objective of the EM algorithm is to estimate a set of model parameters which maximizes the incomplete log-likelihood function. The EM algorithm is described as follows.

EM algorithm

1. Set $\boldsymbol{\Lambda}^{(0)}$ and $k \leftarrow 0$.
2. Iterate the following EM-steps until convergence:

   E-step: Calculate $\mathcal{Q}(\boldsymbol{\Lambda}, \boldsymbol{\Lambda}^{(k)})$

   M-step: $\boldsymbol{\Lambda}^{(k+1)} = \arg\max_{\boldsymbol{\Lambda}} \mathcal{Q}(\boldsymbol{\Lambda}, \boldsymbol{\Lambda}^{(k)})$

   Set $k \leftarrow k + 1$

where k denotes the iteration number. The EM algorithm starts with an initial model parameter set $\boldsymbol{\Lambda}^{(0)}$ and iteratively maximizes the auxiliary function called $\mathcal{Q}$-function.

$$\mathcal{Q}(\boldsymbol{\Lambda}, \boldsymbol{\Lambda}^{(k)}) = \sum_{\boldsymbol{q}} P(\boldsymbol{q} \mid \boldsymbol{o}, \boldsymbol{\Lambda}^{(k)}) \log P(\boldsymbol{o}, \boldsymbol{q} \mid \boldsymbol{\Lambda}) \quad (2)$$

where $P(\boldsymbol{q} \mid \boldsymbol{o}, \boldsymbol{\Lambda})$ is the posterior probability of a state sequence $\boldsymbol{q}$. It can be obtained by the Bayes theorem as follows.

$$P(\boldsymbol{q} \mid \boldsymbol{o}, \boldsymbol{\Lambda}) = \frac{P(\boldsymbol{o}, \boldsymbol{q} \mid \boldsymbol{\Lambda})}{\sum_{\boldsymbol{q}} P(\boldsymbol{o}, \boldsymbol{q} \mid \boldsymbol{\Lambda})} \quad (3)$$

E-step calculates $\mathcal{Q}$-function, which is the expectation of the log-likelihood with respect to the conditional distribution of $\boldsymbol{q}$ given $\boldsymbol{o}$ under the current estimate of the parameter set $\boldsymbol{\Lambda}^{(k)}$, and M-step computes a parameter set maximizing $\mathcal{Q}$-function. This procedure is iterated until convergence of the expected log-likelihood calculated in the E-step. However, the EM algorithm sometimes suffers from the local maxima problem because it is a hill-climbing approach.

3.2. Deterministic annealing EM algorithm

To relax the local maxima problem in the EM algorithm, the deterministic annealing EM (DAEM) algorithm has been proposed [9]. In the DAEM algorithm, the problem of maximizing the log-likelihood function is reformulated as the problem of minimizing the following free energy function.

$$\begin{aligned} \mathcal{F}(\boldsymbol{\Lambda}) &= -\frac{1}{\beta} \log \sum_{\boldsymbol{q}} P^{\beta}(\boldsymbol{o}, \boldsymbol{q} \mid \boldsymbol{\Lambda}) \\ &= -\sum_{\boldsymbol{q}} f(\boldsymbol{q} \mid \boldsymbol{o}, \boldsymbol{\Lambda}) \log P(\boldsymbol{o}, \boldsymbol{q} \mid \boldsymbol{\Lambda}) \\ &\quad -\frac{1}{\beta} \sum_{\boldsymbol{q}} f(\boldsymbol{q} \mid \boldsymbol{o}, \boldsymbol{\Lambda}) \{-\log f(\boldsymbol{q} \mid \boldsymbol{o}, \boldsymbol{\Lambda})\} \quad (4) \end{aligned}$$

where $1/\beta$ is called as "temperature" in the DAEM algorithm. If $\beta = 1$, the negative free energy $-\mathcal{F}(\boldsymbol{\Lambda})$ becomes equal to the log-likelihood function $\mathcal{L}(\boldsymbol{\Lambda})$. In the deterministic annealing approach, the new posterior distribution f is derived so as to minimize the free energy under the constraint of $\sum_{\boldsymbol{q}} f = 1$. To solve this problem, we can use elementary calculus of variations to take functional derivatives of Eq. (4) with respect to f, and the optimal distribution can be derived as follows.

$$f(\boldsymbol{q} \mid \boldsymbol{o}, \boldsymbol{\Lambda}) = \frac{P^{\beta}(\boldsymbol{o}, \boldsymbol{q} \mid \boldsymbol{\Lambda})}{\sum_{\boldsymbol{q}} P^{\beta}(\boldsymbol{o}, \boldsymbol{q} \mid \boldsymbol{\Lambda})} \quad (5)$$

The DAEM algorithm maximizes the auxiliary function called $\mathcal{U}$-function instead of the $\mathcal{Q}$-function used in the EM algorithm.

$$\begin{aligned} \mathcal{U}_{\beta}(\boldsymbol{\Lambda}, \boldsymbol{\Lambda}^{(k)}) &= \sum_{\boldsymbol{q}} f(\boldsymbol{q} \mid \boldsymbol{o}, \boldsymbol{\Lambda}^{(k)}) \log P(\boldsymbol{o}, \boldsymbol{q} \mid \boldsymbol{\Lambda}) \\ &= \sum_{\boldsymbol{q}} \frac{P^{\beta}(\boldsymbol{o}, \boldsymbol{q} \mid \boldsymbol{\Lambda}^{(k)})}{\sum_{\boldsymbol{q}'} P^{\beta}(\boldsymbol{o}, \boldsymbol{q}' \mid \boldsymbol{\Lambda}^{(k)})} \log P(\boldsymbol{o}, \boldsymbol{q} \mid \boldsymbol{\Lambda}) \end{aligned} \quad (6)$$

The temperature parameter β is gradually increased while iterating the EM-steps at each temperature in the DAEM algorithm. When $1/\beta$ is set to an initial temperature $\beta^{(0)} \approx 0$, the EM-steps may achieve a single global minimum of $\mathcal{F}(\boldsymbol{\Lambda})$. At the initial temperature, the posterior distribution f takes a form nearly uniform distribution. While the temperature is decreasing, the form of f changes from uniform to the original posterior. Finally at the temperature $1/\beta = 1$, the DAEM algorithm is identical with the original EM algorithm. The DAEM algorithm is described as follows.

DAEM algorithm

1. Set $\beta \leftarrow \beta^{(0)} (\beta^{(0)} \approx 0)$.
2. Set $\boldsymbol{\Lambda}^{(0)}$ and $k \leftarrow 0$.
3. Iterate the following EM-steps until convergence:

   E-step: $\mathcal{U}_{\beta}(\boldsymbol{\Lambda}, \boldsymbol{\Lambda}^{(k)})$

   M-step: $\boldsymbol{\Lambda}^{(k+1)} = \arg\max_{\boldsymbol{\Lambda}} \mathcal{U}_{\beta}(\boldsymbol{\Lambda}, \boldsymbol{\Lambda}^{(k)})$

   Set $k \leftarrow k + 1$
4. Increase β.
5. If $\beta < 1$, repeat from step 3.

3.3. Optimization of state sequences

In HMM-based speech synthesis , the DAEM posterior distribution f can be calculated by the forward-backward algorithm. The numerator of the posterior distribution in Eq. (5) is written as follows.

$$P^{\beta}(\boldsymbol{o}, \boldsymbol{q} \mid \boldsymbol{\Lambda}) = P^{\beta}(\boldsymbol{o} \mid \boldsymbol{q}, \boldsymbol{\Lambda}) P^{\beta}(\boldsymbol{q} \mid \boldsymbol{\Lambda}) \quad (7)$$

where $P(\boldsymbol{o} \mid \boldsymbol{q}, \boldsymbol{\Lambda})$ and $P(\boldsymbol{q} \mid \boldsymbol{\Lambda})$ indicate state output and transition probabilities, respectively. It can be observed that Eq. (7) has the same form as the likelihood function of HMMs. Therefore, the expectations with respect to the DAEM posterior distribution f can be calculated by replacing the state output and transition probabilities with $P^{\beta}(\boldsymbol{o} \mid \boldsymbol{q}, \boldsymbol{\Lambda})$ and $P^{\beta}(\boldsymbol{q} \mid \boldsymbol{\Lambda})$, respectively. When the temperature parameter is set to the initial temperature $\beta^{(0)} \approx 0$, the state output and transition distributions of all models are represented by using the same parameter. Therefore, the reliable model parameters can be estimated without the phone boundary information when the DAEM algorithm is applied.

4. Stepwise model selection

It is well known that contextual factors affect speech. Therefore, context-dependent acoustic models [21, 22] are widely used in HMM-based speech synthesis. Although a large number of context-dependent acoustic models can capture variations in speech data, too many model parameters lead to the over-fitting problem. Consequently, maintaining a good balance between model complexity and the amount of training data is very important for obtaining high generalization performance. The decision tree based context clustering [11] is an efficient method for dealing with the problem of data sparseness, for both estimating robust model parameter of context-dependent acoustic models and obtaining predictive distributions of unseen contexts. This method constructs a model parameter tying structure which can assign a sufficient amount of training data to each HMM state.

Since model structures affect the model parameter estimation, the model parameters are accurately estimated by selecting appropriate model structures. In the decision tree based context clustering, it is typically assumed that the state occupancies are not changed by the split nodes. By using this assumption, the context clustering is efficiently performed. However, the state occupancies are practically changed before and after context clustering. Especially when the model structures are significantly changed, the state occupancies are also significantly changed and the assumption affects the model selection. In addition, when hidden semi-Markov models are used as acoustic models, the state occupancies are significantly changed because the model structure of state duration model is changed. Once the inappropriate model structures are selected, the model parameters are estimated under the model structures. As a result, the estimated model parameters are not reliable. To alleviate this problem, we apply stepwise model selection in the training. In the stepwise model selection method, decision trees are gradually changed from small trees into large trees for estimating reliable acoustic models. By changing from small trees into large trees in the training procedures, the differences before and after context clustering become small, and the joint optimization of the model structures and the state occupancies are performed.

In HMM-based speech synthesis, the minimum description length (MDL) criterion is widely used as the criterion for model selection [23]. The context clustering based on the MDL criterion constructs the decision trees which minimize the objective function $\mathcal{F}_{\mathrm{MDL}}$.

$$\mathcal{F}_{\mathrm{MDL}} = -\log P(\boldsymbol{o} \mid \boldsymbol{\Lambda}) + \alpha D N log \mathcal{T}(S_0) \qquad (8)$$

where D and N are the dimension of observation vectors and the number of clusters, $\mathcal{T}(S_0)$ is the amount of training data assigned to the root node of the decision tree, and α is the tuning parameter to control the size of the selected decision tree. In the training of our system, the stepwise model selection is performed by gradually decreasing the tuning parameter α. The context clustering was separately applied to distributions of the spectrum, F_0, aperiodicity measures, and state duration.

5. Blizzard Challenge 2011 evaluation

We used 10,000 utterances, which were selected according to the alignment likelihood, as training data. Speech signals were sampled at a 48 kHz rate and windowed by an F_0-adaptive Gaussian window with a 5 ms shift. Feature vectors comprised 231-dimensions: 49-dimension STRAIGHT [16] mel-cepstral coefficients (plus the zero-th coefficient), log F_0, 26 band-filtered aperiodicity measures, and their dynamic and acceleration coefficients. We used 5-state left-to-right context-dependent multi-stream MSD-HSMMs [4, 12] without skip transitions as acoustic models. Each state output distribution was composed of spectrum and F_0 streams. The spectrum stream was modeled by single multi-variate Gaussian distributions with diagonal covariance matrices. The F_0 stream was modeled by a multi-space probability distribution consisting of a Gaussian distribution for voiced frames and a discrete distribution for unvoiced frames. Each state duration distribution was modeled by a one-dimensional Gaussian distribution.

In the training part of our system, we trained the MSD-HSMMs according to the following procedure.

Training procedure

1. Estimate monophone models using DAEM algorithm. The number of temperature parameter updates was ten, and the number of model parameter updates at each temperature was five.
2. Model structures are selected by using the MDL criterion and the tuning parameter α.
3. Estimate context-dependent models using EM algorithm. The number of model parameter updates was five.
4. Decrease the tuning parameter.
5. Repeat from step 2.

In this training procedure, we did not use the given phone boundary information because the DAEM algorithm was applied. The decision tree-based context clustering technique was separately applied to distributions of the spectrum, F_0, aperiodicity measures, and state duration. The model selection was iterated three times. Then, the tuning parameter decreased as follows: 4, 2, 1.

In the synthesis part, we applied the parameter generation algorithm considering global variance (GV) to all parameter generation processes. In order to improve the estimation accuracy of GV models, we used the GV features calculated from only speech region excluding silence and pause regions and estimated the context-dependent GV models instead of a single global GV model. The decision tree-based context clustering technique was also applied to the context-dependent GV models. The decision tree was automatically selected by the MDL criterion. In this system, only sentence-level contextual features (e.g., number of phonemes in a sentence) were used.

5.1. Experimental results

To evaluate naturalness and similarity, 5-point mean opinion score (MOS) tests were conducted. The scale for the naturalness was 5 for "completely natural" and 1 for "completely unnatural". The scale for the similarity was 5 for "sounds like exactly the same person" and 1 for "sounds like a totally different person" compared to a few natural example sentences from the reference speaker. To evaluate intelligibility, the subjects were asked to transcribe US address and semantically unpredictable sentences (SUS), and the average word error rates (WER) were calculated from these transcripts.

Figure 2 shows the evaluation results on naturalness. Figure 3 shows the evaluation results on speaker similarity. Figure 4 shows the evaluation results on intelligibility. In these figure, "A", "B", "C", "D", and "F" correspond as follows.

- A: natural speech.

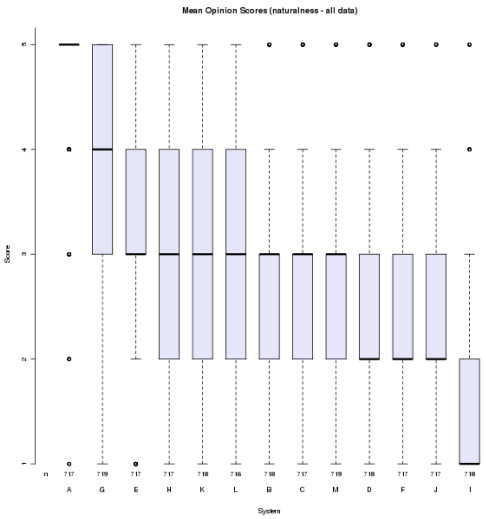

Figure 2: Results of MOS on naturalness.

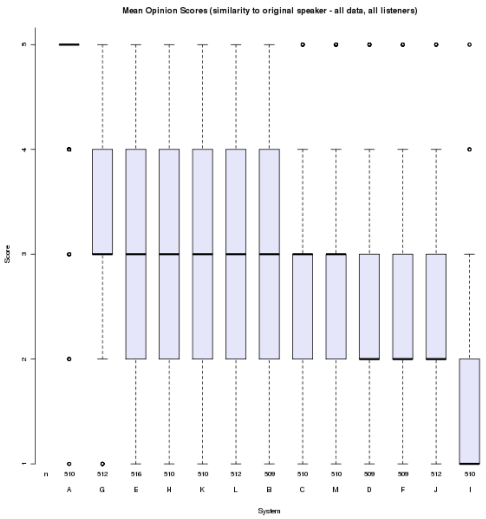

Figure 3: Results of MOS on speaker similarity.

- B: a Festival benchmark system. This system is a standard unit-selection voice built using the same method as used in the CSTR entry to Blizzard 2007.
- C: a benchmark speaker-dependent HMM-based system. This system is built using a similar method to the HTS entry to Blizzard 2005.
- D: the same as System C, except using 48kHz sample rate data.
- F: the 2011 NIT HMM-based speech synthesis system.

The results of listening tests showed that our system "F" was as good as both the benchmark unit-selection system "B" and the HMM-based system "D" in naturalness of speech. However, our system "F" was worse than the benchmark unit-selection system "B" in terms of speaker similarity. The speaker similarity was the acknowledged weakness of the HMM-based speech synthesis method especially in the case of large speech databases. In terms of intelligibility, our system "F" outperformed the benchmark unit-selection system "B" although our system was as good as the HMM-based system "C" and "D". [24] also showed that a HMM-based speech synthesis system was significantly more intelligible than a unit-selection based speech synthesis system. These results indicate that our system "F" can generate the high intelligible speech although the naturalness and speaker similarity do not reach high enough levels. It seems that the buzziness of speech cause these results. Therefore, we need to improve the excitation model and feature extraction.

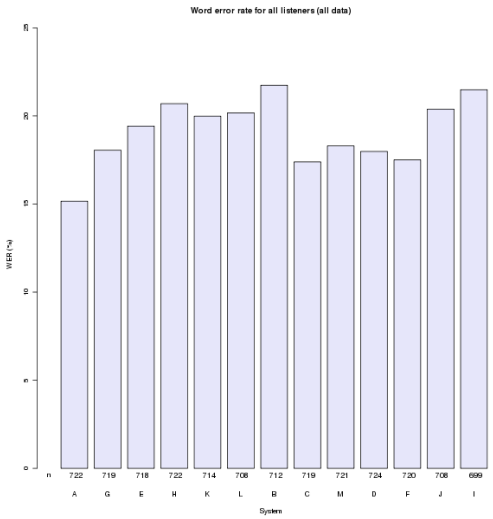

Figure 4: Results of WER.

6. Conclusion

We described HMM-based speech synthesis system developed at the Nagoya Institute of Technology (NIT) for the Blizzard Challenge 2011. We experimented with the DAEM algorithm and stepwise model selection in the training part and the parameter generation considering context-dependent global variance (GV) excluded silence and pause segments in the synthesis part. The results of listening tests showed that our system was as good as both the benchmark unit-selection system and the HMM-based system in naturalness of speech. However, our system was worse than the benchmark unit-selection system in terms of speaker similarity. The speaker similarity was the acknowledged weakness of the HMM-based speech synthesis method especially in the case of large speech databases. In terms of intelligibility, our system competed with natural speech and outperformed the benchmark unit-selection system although there was no significant difference. These results indicate that our system can generate the high intelligible speech although the naturalness and speaker similarity do not reach high enough levels. It seems that the buzziness of speech cause these results. Therefore, we need to improve the excitation model and feature extraction.

7. Acknowledgements

The research leading to these results was partly funded from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 213845 (the EMIME project http://www.emime.org) and the Strategic Information

and Communications R&D Promotion Programme (SCOPE) of the Ministry of Internal Affairs and Communication, Japan. A part of this research was supported by Japan Society for the Promotion of Science (JSPS) Research Fellow.

8. 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," *Proceedings of Eurospeech 1999*, pp. 2347–2350, 1999.

[2] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis," *IEICE Transactions on Information & Systems*, vol. E88-D, no. 3, pp. 502–509, 2005.

[3] R. Tsuzuki, H. Zen, K. Tokuda, T. Kitamura, M. Bulut, and S. Narayanan, "Constructing emotional speech synthesizers with limited speech database," *Proceedings of ICSLP 2004*, vol. 2, pp. 1185–1188, 2004.

[4] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Hidden semi-Markov model based speech synthesis," *Proceedings of ICSLP*, pp. 1185–1180, 2004.

[5] J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training," *IEICE Transactions on Information & Systems*, vol. E90-D, no. 2, pp. 533–543, 2007.

[6] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Speaker adaptation for HMM-based speech synthesis system using MLLR," *Proceedings of ESCA/COCOSDA Third International Workshop on Speech Synthesis*, pp. 273–276, 1998.

[7] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Adaptation of pitch and spectrum for HMM-based speech synthesis using mllr," *Proceedings of ICASSP 2001*, pp. 805–808, 2001.

[8] A. Dempster, N. Laird, and D. Rubin, "Maximum-likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[9] N. Ueda and R. Nakano, "Deterministic annealing em algorithm," *Neural Networks*, vol. 11, pp. 271–282, 1998.

[10] Y. Itaya, H. Zen, Y. Nankaku, C. Miyajima, K. Tokuda, and T. Kitamura, "Deterministic annealing em algorithm in acoustic modeling for speaker and speech recognition," *IEICE Transactions on Information & Systems*, vol. E88-D, no. 3, pp. 425–431, 2005.

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

[12] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Multi-space probability distribution HMM," *IEICE Transactions on Information & Systems*, vol. E85-D, no. 3, pp. 455–464, 2002.

[13] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," *Proceedings of ICASSP 2000*, pp. 936–939, 2000.

[14] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Speaker interpolation in HMM-based speech synthesis system," *Proceedings of Eurospeech 1997*, pp. 2523–2526, 1997.

[15] K. Shichiri, A. Sawabe, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Eigenvoices for HMM-based speech synthesis," *Proceedings of ICSLP 2002*, pp. 1269–1272, 2002.

[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] H. Kawahara, H. Katayose, A. Cheveigne, and R. Patterson, "Fixed point analysis of frequency to instantaneous frequency mappintg for accurate estimation of f_0 and periodicity," *Proceedings of Eurospeech 1999*, pp. 2781–2784, 1999.

[18] 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," *Proceedings of MAVEBA*, pp. 13–15, 2001.

[19] T. Toda and K. Tokuda, "Speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *Proceedings of Interspeech 2005*, pp. 2801–2804, 2005.

[20] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "A fully consistent hidden semi-Markov model-based speech recognition system," *IEICE Transactions on Information & Systems*, vol. E91-D, no. 11, pp. 2693–2700, 2008.

[21] K. Lee, "Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 38, no. 4, pp. 599–609, 1990.

[22] J. Odell, "The use of context in large vocabulary speech recognition," *PhD dissertation, Cambridge University*, 1995.

[23] K. Shinoda and T. Watanabe, "Acoustic modeling based on the MDL criterion for speech recognition," *Proceedings of Eurospeech 1997*, pp. 99–102, 1997.

[24] M. Wolters, K. Isaac, and S. Renals, "Evaluating speech synthesis intelligibility using amazon mechanical turk," *Proceedings of SSW7*, pp. 136–141, 2010.