Overview of NIT HMM-based speech synthesis system for Blizzard Challenge 2010

Keiichiro Oura, Kei Hashimoto, Sayaka Shiota, Keiichi Tokuda

Department of Computer Science and Engineering, Nagoya Institute of Technology, Japan

{uratec,bonanza,sayaka}@sp.nitech.ac.jp, tokuda@nitech.ac.jp

Abstract

This paper describes a hidden Markov model (HMM)-based speech synthesis system developed for the Blizzard Challenge 2010. This system employs STRAIGHT vocoding, minimum generation error (MGE) training, minimum generation error linear regression (MGELR) based model adaptation, the Bayesian speech synthesis framework, and the parameter generation algorithm considering global variance. The real-time factor of the speech synthesis system is about 0.3, and its footprint is less than 25 MB. Subjective evaluation results show that the overall speech quality and intelligibility of the systems are better than most other system, especially when a well-labeled speech database can be used.

Index Terms: HMM, speech synthesis, speaker adaptation, HTS, Blizzard Challenge

1. Introduction

A statistical parametric speech synthesis framework based on hidden Markov models (HMMs) was recently developed. In the HMM-based speech synthesis framework, spectrum, pitch, and duration of speech are modeled simultaneously by HMMs, and speech parameter sequences are generated from the HMMs themselves [1]. Compared to other synthesis methods, the HMM-based approach has several advantages, 1) under its statistical training framework, it can automatically learn salient 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]. Furthermore, it can generate smooth and stable speech under a small footprint. As a result, HMM-based speech synthesis gradually became popular both in research and application.

In HMM-based speech synthesis, the maximum likelihood (ML) criterion has typically been used to train HMMs. The optimal model parameters can be obtained by maximizing the likelihood for given training data as

$$\mathbf{\Lambda}_{\mathrm{ML}} = \arg\max_{\mathbf{\Lambda}} P(\boldsymbol{O} \mid S, \mathbf{\Lambda}), \tag{1}$$

where S is a label sequence of training data. Since it is difficult to obtain the model parameter $\mathbf{\Lambda}_{\mathrm{ML}}$ analytically, the model parameters are estimated by using an iterative procedure such as the EM algorithm. In the synthesis part, the speech parameter generation algorithm generates the sequence of speech parameter vectors that maximize its output probability using the model parameters $\mathbf{\Lambda}_{\mathrm{ML}}$ as

$$\hat{\boldsymbol{o}}_{\mathrm{ML}} = \arg\max_{\boldsymbol{o}} P(\boldsymbol{o} \mid s, \mathbf{\Lambda}_{\mathrm{ML}}), \tag{2}$$

where $\boldsymbol{o} = \left[\boldsymbol{o}_1^\top, \boldsymbol{o}_2^\top, \ldots, \boldsymbol{o}_T^\top\right]^\top$ is a speech parameter sequence, and s is a label sequence to be synthesized.

Although the performance of the conventional HMM-based speech synthesis framework is good enough for standard applications, the quality of synthesized speech still can be improved. In recent years, several techniques have been adopted to improve the quality of synthesized speech for HMM-based speech synthesis, including a high quality vocoder Speech Transformation and Representation using Adaptive Interpolation of weGHTed spectrum (STRAIGHT) [8] for spectral analysis, a minimum generation error (MGE) criterion for model training [9], the Bayesian speech synthesis framework [10, 11], and parameter generation algorithm considering global variance (GV) [12]. In NIT's system for the Blizzard Challenge 2010, we build three HMM-based speech synthesis systems using these state-of-the-art techniques.

1. **MGE**: This system uses the MGE criterion for model training. After the basic acoustic models are trained based on the ML criterion, the model parameters are updated several times by the MGE criterion. Here, the Euclidean distance on mel-cepstral coefficients is used as the criterion. We applied this system to EH1, EH3 and MH1 tasks.

2. **BAYES**: This system is based on the Bayesian speech synthesis framework. The estimation of posterior distributions, model selection, and speech parameter generation are consistently performed based on the Bayesian criterion. Since the Bayesian approach can construct more robust model than the ML approach, we applied this system to EH2 and MH2 tasks which consists of smaller training data than those in EH1 and MH1 tasks.

3. **MGELR**: This system is a speaker adaptation system. The average voice model is trained by the ML-based speaker adaptive training (SAT) method [13]. The transform matrices are trained by the MGE criterion. The minimum generation error linear regression (MGELR) method [14] shows better performance than maximum likelihood linear regression (MLLR) method. We applied this system to ES1 and MS1 tasks.

All systems uses the hidden semi-Markov models (HSMMs) [4] as the acoustic models, STRAIGHT vocoding, and the parameter generation algorithm considering GV.

The rest of the paper organized as follows. Section 2 describes NIT's baseline system. In sections 3 and 4, we briefly review the MGE training and the MGELR based model adaptation, and Bayesian speech synthesis, respectively. In section 5, evaluation results are shown. Our conclusions are given in section 6.

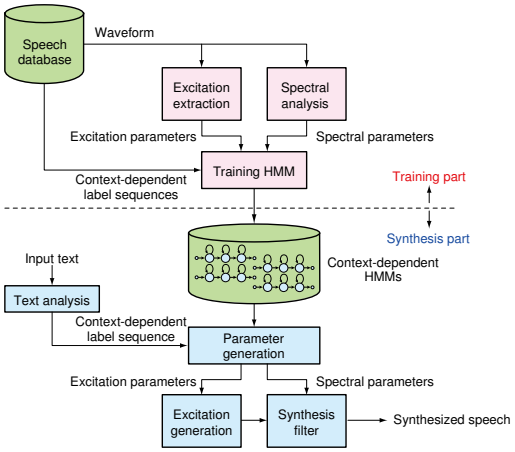

Figure 1: Overview of HMM-based speech synthesis system

2. Basic system

2.1. HMM-based speech synthesis system

Figure 1 shows an overview of the basic HMM-based speech synthesis system [1]. It consists of training and synthesis parts.

In the training part, spectral and excitation parameters are extracted from a speech database, and each feature vector consists of spectrum and excitation parameter vectors: the spectrum parameter vectors are composed of mel-cepstral coefficients, their delta, and delta-delta, and the excitation parameter vectors composed of logarithmic fundamental frequency (log F_0) values and aperiodicity measurements, their delta, and delta-delta. Although the spectrum part can be modeled by continuous HMM, the F_0 part cannot be modeled by continuous or discrete HMM since 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, the feature vectors are modeled by context-dependent multi-space probability distribution (MSD) HMMs [15].

In the synthesis stage, first an arbitrarily given text to be synthesized is converted to a context-dependent label sequence and a sentence HMM is constructed by concatenating the context-dependent HMMs according to the label sequence. Secondly, state durations maximizing their probabilities were determined. Thirdly, mel-cepstral coefficients and log F_0 sequences maximizing their output probabilities for a given state sequence are generated by speech parameter generation algorithm (case 1 in [16]). Finally, speech waveform is synthesized directly from the generated mel-cepstral coefficients and log F_0 sequences using Mel Log Spectrum Approximation (MLSA) filter.

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 limitation, in the HMM-based speech synthesis system, each state duration probability distribution is explicitly modeled by a single Gaussian distribution. They are estimated from statistics obtained in the last iteration of the forward-backward algorithm, and then clustered by the decision tree-based context clustering [17, 18]. In the synthesis part, we construct a sentence HMM corresponding to an arbitrarily given text and determine state durations which maximize their probabilities. Then, a speech parameter sequence is generated for the given state sequence by the speech parameter generation algorithm. However, there is an inconsistency between training and synthesis: although speech is synthesized from HMMs with explicit state duration probability distributions, HMMs are trained without them. To overcome this inconsistency, hidden semi-Markov model (HSMM) based speech synthesis has been proposed [4]. 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 re-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.* [8]. It consists of three main components, i.e., F_0 extraction, spectral and aperiodic analysis, and speech synthesis.

The STRAIGHT automatically extract F_0 with fixed-point analysis [19]. We adopt a two-stage algorithm to alleviate errors of the F_0 extraction, e.g., halving and doubling. Firstly, we perform the F_0 extraction for all training data for each speaker in which a search range is set to 55-480 Hz. Taking account of a histogram of the extracted F_0s, we roughly estimate an F_0 range of each speaker. Then, F_0s are again extracted in the speaker-specific range.

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 40-th STRAIGHT mel line spectrum pair (mel-LSP) coefficients and 0th through 39th mel-cepstral coefficients to which the smoothed spectrum analyzed by the STRAIGHT is converted. An aperiodicity measure in the frequency domain [20] is also extracted. As a parameter for constructing a mixed excitation sources in speech synthesis, average values of the aperiodicity measures on five frequency bands, 0-1, 1-2, 2-4, 4-6, and 6-8 kHz are used for 16k sampling data and 0-3, 3-6, 6-12, 12-18 and 18-24 kHz are used for 48k sampling data.

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 [21] 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 distri-

bution 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 accuracy of GV estimation, the GV Gaussian probability density function (pdf) is changed from a single global distribution to a context-dependent one. In a similar way to HMM observation density tying, the decision-tree based context clustering technique is applied to the context-dependent GV pdfs to tie their parameters. The number of leaf nodes of the decision trees is automatically determined by the MDL criterion [22]. In this paper, to simplify the implementation, only sentence-level contextual features (e.g., number of phonemes in a sentence) were used. Furthermore, to improve the estimation accuracy of the GV vector, the GV vector is calculated from only speech region excluding silence and pause regions from the calculation, based on automatic segmentation. Since HSMMs are used as acoustic models in our system, the silence and pause regions are estimated by using WFST-based aligner [23].

3. Minimum generation error criterion

3.1. Minimum generation error training

In general, the aim of HMM-based speech synthesis is to generate the speech as close to the natural speech as possible, i.e., the generation error should be as small as possible. The parameter generation algorithm [16] is applied to obtain the speech parameter vector sequence $\boldsymbol{o}$ which maximizes $P(\boldsymbol{o} \mid \lambda, \boldsymbol{q})$, where λ and $\boldsymbol{q}$ are a given HMM and the state sequence, respectively. In order to keep the smooth property of the generated parameter sequence, the dynamic features including delta and delta-delta coefficients are used. For a state sequence $\boldsymbol{q}$ of a given speech parameter vector sequence $\boldsymbol{o}$, the generated vector sequence $\hat{\boldsymbol{c}}(\lambda, \boldsymbol{q})$ can be calculated. We assume the distance between original and generated data as $D(\boldsymbol{c}, \hat{\boldsymbol{c}}(\lambda, \boldsymbol{q}))$. Without loss of generality, we denote as $\hat{\boldsymbol{c}}(\lambda, \boldsymbol{q})$ as $\hat{\boldsymbol{c}}_{\boldsymbol{q}}$. The generation error $\hat{e}(\boldsymbol{c}, \lambda)$ for a feature vector sequence $\boldsymbol{c}$ is calculated by using the Euclidean distance $D(\boldsymbol{c}, \hat{\boldsymbol{c}}_{\boldsymbol{q}})$ as

$$\hat{e}(\boldsymbol{c}, \lambda) = D(\boldsymbol{c}, \hat{\boldsymbol{c}}_{\boldsymbol{q}}) = \| \boldsymbol{c} - \hat{\boldsymbol{c}} \|^{\top}. \tag{3}$$

It should be noted that the distance measure can be replaced by other measure which is more suitable for the real application. Under the definition of generation error, we incorporated the parameter generation into the HMM training procedure for generation error calculation. In order to minimize the generation errors, the GPD algorithm is applied. The HMM parameters were optimized to minimize the total generation errors of training data.

A log spectral distortion (LSD) was adopted to replace the Euclidean distance to define the generation error between the original and generated LSPs [24] in MGE training, and the quality of synthesized speech was improved. However, MGE training with Euclidean distance was used for Blizzard Challenge 2010 because high computational cost is required for MGE training with LSD.

3.2. MGELR-based model adaptation

In the MGELR-based model adaptation, we incorporate the parameter generation into model adaptation process to calculate the generation errors of adaptation data, and then optimize the parameters of transformation matrices so as to minimize the total generation errors of adaptation data. After model transformation, the generation error for a feature vector sequence $\boldsymbol{c}$ in adaptation data is defined as Eq. (3) where $\hat{\boldsymbol{c}}_{\boldsymbol{q}}$ is the generated feature vector sequence using the transformed models, which is calculated as

$$\hat{\boldsymbol{c}}_{\boldsymbol{q}} = \hat{\boldsymbol{R}}_{\boldsymbol{q}}^{-1} \hat{\boldsymbol{r}}_{\boldsymbol{q}}, \tag{4}$$

where

$$\hat{\boldsymbol{R}}_{\boldsymbol{q}} = \boldsymbol{W}^{\top} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{q}}^{-1} \boldsymbol{W}, \tag{5}$$

$$\hat{\boldsymbol{r}}_{\boldsymbol{q}} = \boldsymbol{W}^{\top} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{q}}^{-1} \hat{\boldsymbol{\mu}}_{\boldsymbol{q}}. \tag{6}$$

The whole model training and adaptation procedure based on the MGELR algorithm is implemented as follows:

1. Train the source voice model using the source speech database.
2. Conduct the MLLR-based model adaptation, and initialize the transformation matrices.
3. Obtain the optimal state alignments for all adaptation data using the MLLR-adapted HMMs.
4. Iteratively optimize the parameters of transformation matrices based on MGELR algorithm.
5. Apply the optimized transformation matrices to the source voice model.

4. Bayesian speech synthesis

The Bayesian approach considers the posterior distribution of any variables. That is, all the variables introduced when models are parameterized, such as model parameters and latent variables, are regarded as random variables, and their posterior distributions are obtained based on the Bayes theorem. The difference between the Bayesian and ML approaches is that the target of estimation is the distribution function in the Bayesian approach whereas it is the parameter value in the ML approach. Based on this posterior distribution estimation, the Bayesian approach can generally achieve more robust model construction than the ML approach. A framework of speech synthesis based on the Bayesian approach was recently proposed [10, 11]. The Bayesian approach assumes that a set of model parameters $\boldsymbol{\Lambda}$ is a random variable, while the ML approach estimates constant model parameters. In the Bayesian approach, the speech parameter is generated by the predictive distribution as follows

$$\begin{aligned} \hat{\boldsymbol{o}}_{\text{Bayes}} &= \arg\max_{\boldsymbol{o}} P(\boldsymbol{o} \mid s, \boldsymbol{O}, S) \\ &= \arg\max_{\boldsymbol{o}} P(\boldsymbol{o}, \boldsymbol{O} \mid s, S). \end{aligned} \tag{7}$$

It can be seen that Eq. (7) directly represents the problem of speech synthesis; that is, speech feature sequence $\boldsymbol{o}$ is generated from given training feature sequences $\boldsymbol{O}$ with labels S and labels to be synthesized s. The marginal likelihood of $\boldsymbol{o}$ and $\boldsymbol{O}$

is defined by

$$
\begin{aligned}
&P(\boldsymbol{o}, \boldsymbol{O} \mid s, S) \\
&\quad = \sum_{\boldsymbol{z}} \sum_{\boldsymbol{Z}} \int P(\boldsymbol{o}, \boldsymbol{z}, \boldsymbol{O}, \boldsymbol{Z}, \boldsymbol{\Lambda} \mid s, S) d\boldsymbol{\Lambda} \\
&\quad = \sum_{\boldsymbol{z}} \sum_{\boldsymbol{Z}} \int P(\boldsymbol{o}, \boldsymbol{z} \mid s, \boldsymbol{\Lambda}) P(\boldsymbol{O}, \boldsymbol{Z} \mid S, \boldsymbol{\Lambda}) P(\boldsymbol{\Lambda}) d\boldsymbol{\Lambda},
\end{aligned} \tag{8}
$$

where $\boldsymbol{z}$ and $\boldsymbol{Z}$ are sequences of HMM states for a speech parameter sequence $\boldsymbol{o}$ and the training data $\boldsymbol{O}$, $P(\boldsymbol{\Lambda})$ is a prior distribution for model parameter $\boldsymbol{\Lambda}$, $P(\boldsymbol{o}, \boldsymbol{z} \mid s, \boldsymbol{\Lambda})$ is the likelihood of synthesis data $\boldsymbol{o}$, and $P(\boldsymbol{O}, \boldsymbol{Z} \mid S, \boldsymbol{\Lambda})$ is the likelihood of the training data $\boldsymbol{O}$. The model parameters are integrated out in Eq. (8) so that the effect of over-fitting is mitigated. However, it is difficult to solve the integral and expectation calculations, when a model includes latent variables. To overcome this problem, the variational Bayesian method has been proposed as a tractable approximation method of the Bayesian approach and it has shown good generalization performance in many applications [25].

The variational Bayesian method maximizes a lower bound of log marginal likelihood $\mathcal{F}$ instead of the true marginal likelihood. A lower bound $\mathcal{F}$ is defined by using Jensen's inequality:

$$
\begin{aligned}
&\log P(\boldsymbol{o}, \boldsymbol{O} \mid s, S) \\
&\quad = \log \sum_{\boldsymbol{z}} \sum_{\boldsymbol{Z}} \int P(\boldsymbol{o}, \boldsymbol{z}, \boldsymbol{O}, \boldsymbol{Z}, \boldsymbol{\Lambda} \mid s, S) d\boldsymbol{\Lambda} \\
&\quad = \log \sum_{\boldsymbol{z}} \sum_{\boldsymbol{Z}} \int Q(\boldsymbol{z}, \boldsymbol{Z}, \boldsymbol{\Lambda}) \frac{P(\boldsymbol{o}, \boldsymbol{z}, \boldsymbol{O}, \boldsymbol{Z}, \boldsymbol{\Lambda} \mid s, S)}{Q(\boldsymbol{z}, \boldsymbol{Z}, \boldsymbol{\Lambda})} d\boldsymbol{\Lambda} \\
&\quad \geq \sum_{\boldsymbol{z}} \sum_{\boldsymbol{Z}} \int Q(\boldsymbol{z}, \boldsymbol{Z}, \boldsymbol{\Lambda}) \log \frac{P(\boldsymbol{o}, \boldsymbol{z}, \boldsymbol{O}, \boldsymbol{Z}, \boldsymbol{\Lambda} \mid s, S)}{Q(\boldsymbol{z}, \boldsymbol{Z}, \boldsymbol{\Lambda})} d\boldsymbol{\Lambda} \\
&\quad = \left\langle \log \frac{P(\boldsymbol{o}, \boldsymbol{z}, \boldsymbol{O}, \boldsymbol{Z}, \boldsymbol{\Lambda} \mid s, S)}{Q(\boldsymbol{z}, \boldsymbol{Z}, \boldsymbol{\Lambda})} \right\rangle_{Q(\boldsymbol{z}, \boldsymbol{Z}, \boldsymbol{\Lambda})} \\
&\quad = \mathcal{F},
\end{aligned} \tag{9}
$$

where $\langle \cdot \rangle_Q$ denotes a calculation of expectation with respect to Q, and $Q(\boldsymbol{z}, \boldsymbol{Z}, \boldsymbol{\Lambda})$ is an approximate distribution of the true posterior distribution $P(\boldsymbol{z}, \boldsymbol{Z}, \boldsymbol{\Lambda} \mid \boldsymbol{o}, \boldsymbol{O}, s, S)$. The VB method uses the assumption that probabilistic variables associated with $\boldsymbol{z}, \boldsymbol{Z}, \boldsymbol{\Lambda}$ are statistically independent of the other variables as

$$
Q(\boldsymbol{z}, \boldsymbol{Z}, \boldsymbol{\Lambda}) = Q(\boldsymbol{z})\, Q(\boldsymbol{Z})\, Q(\boldsymbol{\Lambda}). \tag{10}
$$

In the VB method, posterior distributions $Q(\boldsymbol{z})$, $Q(\boldsymbol{Z})$ and $Q(\boldsymbol{\Lambda})$ are introduced to approximate the true posterior distributions. The optimal posterior distributions can be obtained by maximizing the objective function $\mathcal{F}$ with the variational method

$$
Q(\boldsymbol{z}) = C_{\boldsymbol{z}} \exp \left\langle \log P(\boldsymbol{o}, \boldsymbol{z} \mid s, \boldsymbol{\Lambda}) \right\rangle_{Q(\boldsymbol{\Lambda})}, \tag{11}
$$

$$
Q(\boldsymbol{Z}) = C_{\boldsymbol{Z}} \exp \left\langle \log P(\boldsymbol{O}, \boldsymbol{Z} \mid S, \boldsymbol{\Lambda}) \right\rangle_{Q(\boldsymbol{\Lambda})}, \tag{12}
$$

$$
\begin{aligned}
Q(\boldsymbol{\Lambda}) = C_{\boldsymbol{\Lambda}} P(\boldsymbol{\Lambda}) &\exp \left\langle \log P(\boldsymbol{o}, \boldsymbol{z} \mid s, \boldsymbol{\Lambda}) \right\rangle_{Q(\boldsymbol{z})} \\
&\times \exp \left\langle \log P(\boldsymbol{O}, \boldsymbol{Z} \mid S, \boldsymbol{\Lambda}) \right\rangle_{Q(\boldsymbol{Z})},
\end{aligned} \tag{13}
$$

where $C_{\boldsymbol{z}}$, $C_{\boldsymbol{Z}}$ and $C_{\boldsymbol{\Lambda}}$ are normalization terms of $Q(\boldsymbol{z})$, $Q(\boldsymbol{Z})$ and $Q(\boldsymbol{\Lambda})$, respectively. These posterior distributions can be updated effectively by iterative calculations similar to the EM algorithm used in the ML approach.

From Eq. (7), the optimal speech parameter sequence for Bayesian speech synthesis can be generated by maximizing the marginal likelihood. Thus, the optimal speech parameter sequence $\hat{\boldsymbol{o}}$ can be generated by maximizing the lower bound $\mathcal{F}$ in Eq. (9) because the VB method guarantees that the log marginal likelihood is approximately the lower bound $\mathcal{F}$.

In the model selection, the VB method can select appropriate model structure, even when there are insufficient amounts of data, because it does not use an asymptotic assumption. In the VB method, since prior distributions of the model parameters affect the estimation of posterior distributions and model selection, the determination of prior distributions is an important problem for estimating of appropriate acoustic models. In this paper, a prior distribution determination technique using the cross validation [26] is apply to the context clustering. Using prior distributions determined by the cross validation, it is expected that a higher generalization ability is achieved and an appropriate model structure can be selected in the context clustering without any tuning parameters.

5. Experiments

5.1. Experimental conditions

Seven systems were constructed for nine tasks in Blizzard Challenge 2010. Experimental conditions for each task are listed in Table 1. Speech signals were windowed with an F_0-adaptive Gaussian window with a 5-ms shift. The feature vectors consisted of STRAIGHT mel-Cepstrum/mel-LSP coefficients, log F_0, aperiodicity measures, and their dynamic and acceleration coefficients. We used 5-state left-to-right context-dependent multi-stream MSD-HSMMs. Each state had a single Gaussian pdf with a diagonal covariance matrix.

5.2. Experimental results

Tables 2-7 show average Mean Opinion Scores (MOSs) and average Word Error Rates (WERs) of natural speech (NATURAL), the best and the worst of other participants (BEST and WORST), and our system (NIT) respectively. From these tables, it can be seen that the NIT system kept the low WERs for all SUS tests. Especially, the ES1, ES3, and MS1 systems achieved the best WERs in all participants. The MGE training and MGELR adaptation seem to work well. Although the Blizzard Challenge rules allow participants to add pronunciations for out-of-vocabulary words found in the test set to their lexicon, we did not add them due to our limited human resources. The NIT systems for Mandarin tasks kept high MOSs. It seems that the labels given by organizers were fortunately accurate. However, there are significant differences between natural speech and all other systems from the point of view of MOSs.

6. Conclusions

We described HMM-based speech synthesis system developed at the Nagoya Institute of Technology (NIT) for Blizzard Challenge 2010. We built three HMM-based speech synthesis systems incorporating several state-of-the-art techniques, including the STRAIGHT vocoder, the MGE training, the MGELR based model adaptation, the Bayesian speech synthesis framework, and the parameter generation algorithm considering GV. The results of listening tests showed that our systems indicated the better performance of intelligibility than most other systems.

Table 1: Experimental conditions.

Task	Our method	Lang	Task	Training data	Sampling rate	Spectrum features
EH1	MGE training	Eng	MOS & SUS	6.0 hours	16 kHz	39 mel-Cepstrum
EH2	Bayesian approach	Eng	MOS & SUS	1.5 hours	16 kHz	39 mel-Cepstrum
ES1	MGELR adaptation	Eng	MOS & SUS	0.1 hours	16 kHz	39 mel-LSP
ES2	MGE training	Eng	SUS with noise	6.0 hours	16 kHz	39 mel-Cepstrum
ES3	MGE training	Eng	MOS & SUS	6.0 hours	48 kHz	49 mel-Cepstrum
MH1	MGE training	Chn	MOS & SUS	8.3 hours	16 kHz	39 mel-Cepstrum
MH2	Bayesian approach	Chn	MOS & SUS	1.0 hours	16 kHz	39 mel-Cepstrum
MS1	MGELR adaptation	Chn	MOS & SUS	0.1 hours	16 kHz	39 mel-LSP
MS2	MGE training	Chn	SUS with noise	8.3 hours	16 kHz	39 mel-Cepstrum

Table 2: Naturalness (English).

Name	EH1	EH2	ES1	ES3
NATURAL	4.8	4.8	4.8	4.9
BEST	4.2	3.9	3.1	3.9
WORST	1.4	1.7	1.6	2.2
NIT	2.8	2.9	2.5	2.0

Table 3: Similarity to original speaker (English).

Name	EH1	EH2	ES1	ES3
NATURAL	4.8	4.8	4.8	4.4
BEST	4.2	3.5	2.8	3.8
WORST	1.6	1.7	1.6	2.1
NIT	2.7	2.7	2.1	2.2

Table 4: Word error rates for SUS test (English).

Name	EH1	EH2	ES1	ES2	ES3
NATURAL	0.12	0.12	0.11	0.51	0.14
BEST	0.15	0.18	0.20	0.41	0.19
WORST	0.28	0.35	0.40	0.76	0.34
NIT	0.16	0.22	0.18	0.61	0.17

Table 5: Naturalness (Mandarin).

Name	MH1	MH2	MS1
NATURAL	4.5	4.6	4.7
BEST	3.9	3.6	3.0
WORST	1.4	1.5	1.7
NIT	3.9	3.4	2.9

Table 6: Similarity to original speaker (Mandarin).

Name	MH1	MH2	MS1
NATURAL	4.3	4.3	4.6
BEST	3.9	3.6	3.0
WORST	1.5	1.5	1.7
NIT	3.1	3.4	2.9

Table 7: Word error rates for SUS test (Mandarin).

Name	MH1	MH2	MS1	MS2
NATURAL	0.17	0.19	0.20	0.58
BEST	0.21	0.23	0.35	0.45
WORST	0.46	0.57	0.76	0.92
NIT	0.22	0.31	0.29	0.65

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), and the Strategic Information and Communications R&D Promotion Programme (SCOPE), Ministry of Internal Affairs and Communication, Japan.

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," Proc. of Eurospeech, 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 Trans. Information and Systems, vol. E88-D, no. 3, pp.z 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," Proc. of ICSLP, 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," Proc. of ICSLP, vol. 2, pp. 1397–1400, 2004.

[5] J. Yamagishi, and T. Kobayashi, "Adaptive training for hidden semi-Markov model," Proc. of ICASSP, vol. 2, pp. 1213–1216, 2004.

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

[7] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Adapta-

tion of pitch and spectrum for HMM-based speech synthesis using MLLR," Proc. of ICASSP, pp. 805–808, 2001.

[8] 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.

[9] Y. J. Wu and R. H. Wang, "Minimum generation error training for HMM-based speech synthesis," Proc. of ICASSP, vol. 1, pp. 889–892, 2006.

[10] K. Hashimoto, H. Zen, Y. Nankaku, T. Masuko, and K. Tokuda, "A Bayesian approach to HMM-based speech synthesis," Proc. of ICASSP, pp. 4029–4032, 2009.

[11] K. Hashimoto, Y. Nankaku, and K. Tokuda, "A Bayesian approach to hidden semi Markov model based speech synthesis," Proc. of Interspeech, pp. 1751–1754, 2009.

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

[13] T. Anastasakos, J. McDonough, R. Schwartz, and K. Tokuda, "A Compact Model for Speaker-Adaptive Training," Proc. of ICSLP, vol. 2, pp. 1137–1140, 1996.

[14] L. Qin, Y. J. Wu, Z. H. Ling, R. H. Wang, and L. R. Dat, "Minimum generation error linear regression based model adaptation for HMM-based speech synthesis," Proc. of ICASSP, pp. 3953–3956, 2008.

[15] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling," Proc. of ICASSP, pp. 229–232, 1999.

[16] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features," Proc. of ICASSP, vol. 1, pp. 660–663, 1995.

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

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

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

[20] 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," Proc. of MAVEBA, pp. 13–15, 2001.

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

[22] K. Shinoda and T. Watanabe, "Acoustic modeling based on the MDL criterion for speech recognition," Proc. of Eurospeech, vol. 1, pp. 99–102, 1997.

[23] K. Oura, H. Zen, Y. Nankaku, A. Lee, and K. Tokuda, "A fully consistent hidden semi-markov model-based speech recognition system," IEICE Trans. Information and Systems, vol. E91-D, no. 11, pp. 2693–2700, 2008.

[24] Y. J. Wu and K. Tokuda, "Minimum generation error training with direct log spectral distortion on LSPs for HMM-based speech synthesis," Proc. of Interspeech, pp. 577–580, 2008.

[25] H. Attias, "Inferring parameters and structure of latent variable models by variational Bayes," Proc. of UAI 15, 1999.

[26] K. Hashimoto, H. Zen, Y. Nankaku, A. Lee, and K. Tokuda, "Bayesian context clustering using cross valid prior distribution for HMM-based speech recognition," Proc. of Interspeech, pp. 936–939, 2008.