The Nitech-NAIST HMM-based speech synthesis system for the Blizzard Challenge 2006

Heiga Zen, Tomoki Toda[†]*, Keiichi Tokuda*

Nagoya Institute of Technology, [†]Nara Institute of Science and Technology
`{zen,tokuda}@ics.nitech.ac.jp`, [†]`tomoki@is.naist.jp`

1. Abstract

The present paper describes an HMM-based speech synthesis system developed by the Nitech-NAIST group for the Blizzard Challenge 2006 (Nitech-NAIST-HTS 2006). To achieve improvements over the 2005 system (Nitech-HTS 2005), new features such as MGC-LSP, MLLT, and full covariance GV pdf are investigated. Subjective listening test results show that combining mel-cepstral coefficients, MLLT and full covariance GV pdf achieved the best score and the performance of the 2006 system is significantly better than that of the 2005 system. Results of the Blizzard Challenge evaluations reveal that Nitech-NAIST-HTS 2006 is still competitive even a relatively large amount of training data is used.

2. Introduction

In January 2005, with a view to allowing closer comparison of corpus-based text-to-speech synthesis systems from labeling, pruning, cost functions, signal processing, and others, an open evaluation named *Blizzard Challenge 2005* [1] was devised. In this challenge, organizers asked participants to use the same speech datasets (CMU ARCTIC databases [2]) to synthesize utterances from a small number of genres. An organized evaluation based on subjective listening tests was also carried out to try to rank the systems and help identify the effectiveness of the techniques [3]. The Nitech group participated in this challenge with a newly designed HMM-based speech synthesis system named Nitech-HTS 2005 [4].

This year, the second challenge was held as the Blizzard Challenge 2006. For this year's challenge, ATR kindly provided a five-hour US English male speech database. Nitech and NAIST jointly developed a new version of HMM-based speech synthesis system (Nitech-NAIST-HTS 2006) and participated in the challenge. To achieve improvements over the 2005 system, we investigated the use of Mel-Generalized Cepstrum-based Line Spectrum Pair (MGC-LSP) [5], Maximum Likelihood Linear Transform (MLLT) [6, 7], and full covariance GV pdf. Nitech-NAIST-HTS 2006. The present paper describes the overview of Nitech-NAIST-HTS 2006 and the results of subjective evaluations.

The rest of the present paper is organized as follows. Section 2 describes the new features integrated into Nitech-NAIST-HTS 2006. Section 3 gives the brief results of the Blizzard Challenge 2006. Finally concluding remarks are presented.

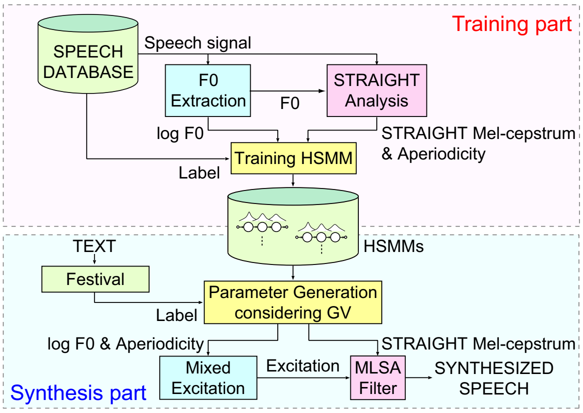

Figure 1: The overview of Nitech-HTS 2005.

3. Nitech-NAIST-HTS 2006

Figure 1 illustrates the overview of Nitech-HTS 2005 [4]. In this system, STRAIGHT-based vocoding [8], Hidden Semi-Markov Model (HSMM) based acoustic modeling [9], and the speech parameter generation algorithm considering Global Variance (GV) [10] were integrated. The structure of Nitech-NAIST-HTS 2006 is almost the same as that of Nitech-HTS 2005. However, to achieve improvements over the 2005 system, we investigate the following new features:

- Using STRAIGHT MGC-LSP parameters instead of STRAIGHT mel-cepstral coefficients as spectral features;
- Applying MLLT to approximate full covariance matrices for state output probability density functions (pdfs);
- Modeling the pdf of GVs using a Gaussian distribution with a full covariance matrix.

In the following sections, brief overview and evaluation of these new features are described.

3.1. Mel-Generalized Cepstrum-Based LSP

The Mel-Generalized Cepstral (MGC) analysis [11] assumes that a speech spectrum $H(z)$ is modeled by the MGC coefficients $c(m)$

as

$$H(z) = \begin{cases} \left(1 + \gamma \sum_{m=0}^{M} c\,(m)\,\tilde{z}^{-m}\right)^{1/\gamma}, & -1 \leq \gamma < 0 \\ \exp \sum_{m=0}^{M} c\,(m)\,\tilde{z}^{-m}, & \gamma = 0 \end{cases} \tag{1}$$

where $\tilde{z}$ is an all-pass transfer function defined by

$$\tilde{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}, \quad |\alpha| < 1. \tag{2}$$

The parameters α and γ control the frequency warping and the weight for pole/zero representation, respectively. It is noted that spectral model of Eq. (1) becomes an all-pole model for $(\alpha, \gamma) = (0, -1)$ and cepstral representation for $(\alpha, \gamma) = (0, 0)$.

If $\gamma \neq 0$, Eq. (1) can be arranged as

$$H\,(z) = \frac{\tilde{K}}{\{C\,(\tilde{z})\}^{-1/\gamma}} \tag{3}$$

$$C\,(\tilde{z}) = 1 + \gamma \sum_{m=1}^{M} c'\,(m)\,\tilde{z}^{-m}, \quad -1 \leq \gamma < 0 \tag{4}$$

where

$$\tilde{K} = \{1 + \gamma c(0)\}^{1/\gamma} \tag{5}$$

$$c'(m) = \frac{c(m)}{1 + \gamma c(0)}, \quad m = 1, 2, \ldots, M. \tag{6}$$

Considering $C\,(\tilde{z})$ as an LPC polynomial, MGC-LSP parameters [5] can be obtained using the same technique for extracting LSP parameters.

MGC-LSP parameters has better quantization/interpolation property than LSP or mel-cepstral coefficients [5], and were successfully applied to speech coding [12]. This characteristics could also be beneficial for the HMM-based speech synthesis because statistical modeling and speech parameter generation processes have close relationship to quantization and interpolation, respectively. Actually, Marume et al. have applied MGC-LSP parameters to spectral representation in the HMM-based speech synthesis and reported significant improvements over the mel-cepstral coefficients [13]. However, they evaluated the performance of MGC-LSP parameters in the basic HMM-based speech synthesis system [14]. In the present paper, we evaluate MGC-LSP parameters on the latest system.

3.2. Maximum Likelihood Linear Transform

In Nitech-HTS 2005, the speech parameter generation algorithm considering GV [10] was used. This algorithm iteratively optimizes the following objective function with respect to a speech parameter vector sequence $\boldsymbol{c}$ (static features only):

$$\mathcal{L} = w \log P\,(\boldsymbol{W}\boldsymbol{c} \mid \boldsymbol{q}, \lambda) + \log P\,(v\,(\boldsymbol{c}) \mid \lambda_v) \tag{7}$$

where λ is a sentence HSMM, $\boldsymbol{q}$ is a state sequence, $\boldsymbol{W}$ is a regression window matrix which appends delta and delta-delta features to $\boldsymbol{c}$, w is a weight for the state output probability, $v\,(\boldsymbol{c})$ is a GV[1] of $\boldsymbol{c}$, and λ_v denotes the parameters of a GV pdf.

Whereas the use of this algorithm dramatically reduces the buzziness in synthesized speech, it sometimes generates artificial sound. One of the possible reasons is that each dimension of static features (cepstral coefficients) would be optimized independently. There are apparently correlations between cepstral features. It is important to keep relationships between cepstral features properly in optimization process. However, currently each state output pdf is modeled by a Gaussian component with a diagonal covariance matrix. Hence, correlations between features are ignored. Although the use of full covariance models may solve this problem, it requires a huge amount of training data.

Recently, various structured precision (inverse covariance) matrix models have been proposed. They well-approximate full covariance models, and allow us to capture correlations between features using relatively small number of parameters. Maximum Likelihood Linear Transform (MLLT) [6] is an instance of the structured precision matrix model. In this model, each precision matrix of Gaussian distribution for state output pdfs is represented as

$$\boldsymbol{\Sigma}_j^{-1} = \boldsymbol{A}^\top \boldsymbol{\Lambda}_j \boldsymbol{A}, \tag{8}$$

where $\boldsymbol{\Sigma}_j$ is a covariance matrix of the j-th Gaussian distribution, $\boldsymbol{A}$ is a global transformation matrix which is tied across all Gaussian distributions, and $\boldsymbol{\Lambda}_j$ is a distribution-specific diagonal matrix whose leading diagonal elements $\Lambda_j^{(ii)} = 1/\sigma_{ji}^2$ are the inverse variances in the transformed space. This formulation allows us to estimate the transformation matrix $\boldsymbol{A}$ and the variances in the transformed space $\boldsymbol{\Lambda}_j$ efficiently using iterative closed-form update formulas [7].

3.3. Full Covariance GV Pdf

In Eq. (7), the GV pdf is usually modeled by a Gaussian distribution as

$$P\,(v(\boldsymbol{c}) \mid \lambda_v) = \mathcal{N}\,(v(\boldsymbol{c}) \mid \boldsymbol{\mu}_v, \boldsymbol{\Sigma}_v)\,, \tag{9}$$

where $\boldsymbol{\mu}_v$ and $\boldsymbol{\Sigma}_v$ are a mean vector and a diagonal covariance matrix, respectively. As discussed in the previous section, the use of a diagonal covariance matrix cannot capture correlations between features. To address this problem, full covariance model or its approximations can be used. In this case we cannot adopt MLLT because currently only one Gaussian distribution is used for modeling GVs. Here, we try to use a factor analysis (FA) to approximate $\boldsymbol{\Sigma}_v$.

3.4. Evaluations

According to informal subjective listening test results, we chose mel-cepstral coefficients, MLLT, and full covariance GV pdf for Nitech-NAIST-HTS 2006. In the present paper, we evaluate these techniques in a formal subjective listening tests.

The 4273 utterances released from the Blizzard Challenge organizers were used for training. Speech signals were sampled at a rate of 16 kHz and windowed with a 5-ms shift. For each frame, a STRAIGHT spectrum, a F_0 value, and aperiodicity measures in five frequency sub-bands were extracted, and then 40-th order mel-cepstral coefficients ($\alpha = 0.42, \gamma = 0$ in Eq. (1)), MGC-LSP parameters ($\alpha = 0.42, \gamma = -1/3$), or mel-LSP parameters ($\alpha = 0.42, \gamma = -1$) were extracted from the STRAIGHT spectrum.

Almost the same model structure and training procedure described in [4] were used. The main difference was that an MLLT global transform and diagonal matrices were estimated in the final iteration of the EM algorithm. Here, we applied MLLT only

[1] The GV is defined as an intra-utterance variance.

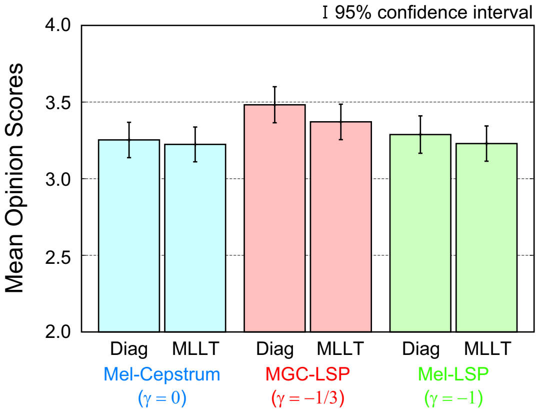

Figure 2: MOSs of mel-cepstrum, MGC-LSP, and mel-LSP based systems trained using diagonal covariance matrices (Diag) or MLLT ***without*** considering GV in speech parameter generation.

to spectral stream. For training the systems, any manual corrections for extracted F_0 values, phoneme boundaries, linguistic and prosodic information files were not performed.

To evaluate the effectiveness of the new features, four subjective listening tests were conducted. The first and second tests evaluated the effect of the spectral parameters *with* and *without* considering GV in speech parameter generation by mean opinion score (MOS) tests, respectively. The third test investigated the type of covariance matrix for the GV pdf by a MOS test. The final test compared Nitech-HTS 2005 and Nitech-NAIST-HTS 2006 trained by this year's dataset in a preference test. In all experiments, 50 conversational and 50 novel sentences from the last year's evaluation were used. Subjects were 17 Japanese graduate students (not native US-English speakers) and all of them completed the above four tests. For each test of each subject, 10 sentences were randomly chosen from the 100 test sentences. For each test sentence, samples were presented in a random order. In the MOS tests, after listening each sample, the subjects were asked to assign a 5-point score (5: natural – 1: poor) to presented speech. In the preference test, after listening each pair of samples, the subjects were asked which sample sounded more natural. All experiments were carried out in a sound-proof room using headphones.

Figures 2 and 3 plot MOSs of mel-cepstrum, MGC-LSP, and mel-LSP based systems trained using diagonal covariance matrices or MLLT *without* and *with* considering GV in speech parameter generation, respectively. It can be seen from the figures that the MGC-LSP based system achieved better scores than the mel-cepstrum and mel-LSP ($\gamma = -1$) based systems, and the use of MLLT degraded MOSs in all systems if the GV was not considered. However, the mel-cepstrum based system with MLLT achieved the best score if the GV was used. Figure 4 shows MOSs of synthesized speech generated without using GV, and with diagonal, FA (10 factors), and full covariance GV pdfs. In this experiment, mel-cepstral coefficients with MLLT was used. Although the full covariance GV pdf achieved the best score, there was no statistically significant difference among GV pdfs using diagonal, FA and full covariance matrices.

Figure 5 plots the preference scores of Nitech-HTS 2005 and Nitech-NAIST-HTS 2006. It shows that the reported naturalness of the speech samples generated by Nitech-NAIST-HTS 2006 was significantly better than that by Nitech-HTS 2005. Some speech samples generated by Nitech-NAIST-HTS 2005 are available at [15].

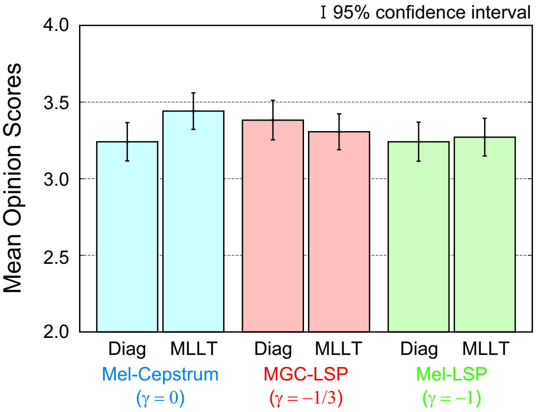

Figure 3: MOSs of mel-cepstrum, MGC-LSP, and mel-LSP based systems trained using diagonal covariance matrices (Diag) or MLLT ***with*** considering GV in speech parameter generation.

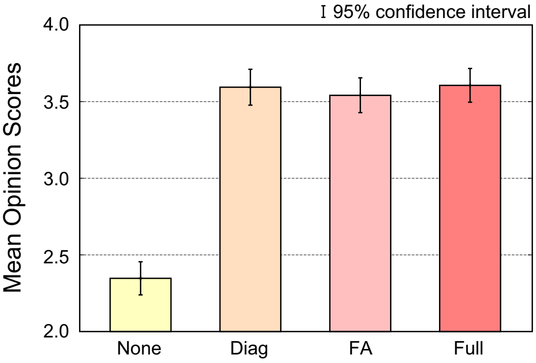

Figure 4: MOSs of speech speech samples generated without GV pdf (None) and with diagonal (Diag), factor analysis with 10 factors (FA), and full covariance (Full) GV pdfs.

4. Results of the Blizzard Challenge 2006

This year, the organizers asked the participants to submit two systems: one was trained using all data (full) and another was trained using only ARCTIC subset of data (arctic). We used mel-cepstral coefficients, MLLT and full covariance GV pdf in both systems. Training our full system took about a day using 9 × 3.2 GHz Pentium D machines (4 GB RAM). For synthesizing speech, utterance files provided by the organizers were used.

An organized evaluation based on subjective listening tests was carried out by the Blizzard Challenge organizers. This year, 14 groups participated in the challenge. Just like the Blizzard Challenge 2005, subjects consisted of three groups (real users from web, speech synthesis experts, and US English native undergradu-

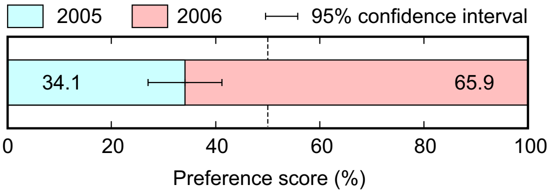

Figure 5: Preference scores of the Nitech-HTS 2005 and Nitech-NAIST-HTS 2006 systems.

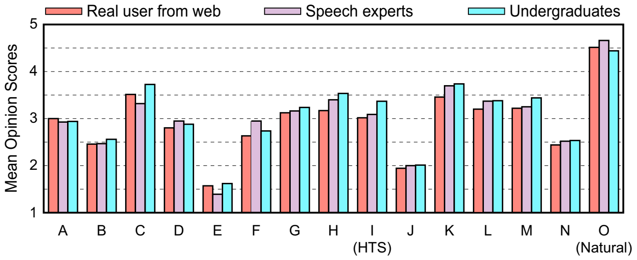

Figure 6: MOSs of all systems which participated in the Blizzard Challenge 2006 (full amount of data were used for training).

ates), and two types of tests (MOS tests for conversational, novels, and news texts, and transcribing tests for DRT/MRT phonetically confusable words within sentences and semantically unpredictable sentences [16]) were carried out.

Figures 6 and 7 shows the results of subjective listening tests carried out in the Blizzard Challenge 2006 (only full systems). In these figures, system “I” corresponds to Nitech-NAIST-HTS 2006. They show that our HMM-based system is still competitive even a relatively large amount of training data is used.

5. Conclusions

The present paper described an HMM-based speech synthesis system developed by the Nitech-NAIST group for the Blizzard Challenge 2006 (Nitech-NAIST-HTS 2006). To achieve improvements over the 2005 system (Nitech-HTS 2005), new features such as MGC-LSP, MLLT, and full covariance GV pdf were investigated. Subjective listening test results showed that combining mel-cepstrum, MLLT and full covariance achieved the best MOS score and Nitech-NAIST-HTS 2006 was significantly better than Nitech-HTS 2005. Results of the Blizzard Challenge evaluations revealed that Nitech-NAIST-HTS 2006 was still competitive even a relatively large amount of training data was used.

6. Acknowledgments

The authors would like to thank for all people who supported the Blizzard Challenge 2006. This work was partly supported by MEXT e-Society project.

7. References

[1] K. Tokuda and A.W. Black. The Blizzard Challenge 2005: Evaluating corpus-based speech synthesis on common datasets. In *Proc. of Interspeech*, pages 77–80, 2005.

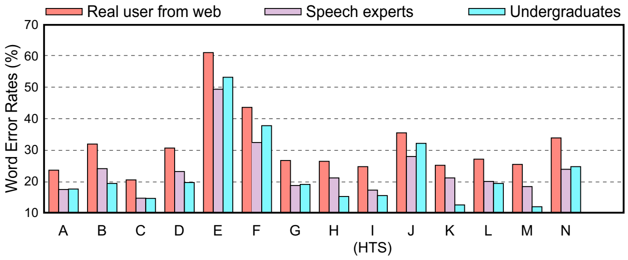

Figure 7: WERs of all systems which participated in the Blizzard Challenge 2006 (full amount of data were used for training).

[2] J. Kominek and A.W. Black. CMU ARCTIC databases for speech synthesis. Technical Report CMU-LTI-03-177, Carnegie Mellon University, 2003.

[3] C. Bennett. Large scale evaluation of corpus-based synthesizers: results and lessons from the 2005 Blizzard Challenge. In *Proc. of Interspeech (Eurospeech)*, pages 105–108, 2005.

[4] H. Zen and T. Toda. An overview of Nitech HMM-based speech synthesis system for Blizzard Challenge 2005. In *Proc. of Interspeech*, pages 93–96, 2005.

[5] K. Koishida, K. Tokuda, T. Kobayashi, and S. Imai. Spectral representation of speech using mel-generalized cepstral coefficients. In *Proc. of 3rd Joint meeting of ASA & ASJ*, pages 963–968, 1996.

[6] R. Gopinath. Maximum likelihood modeling with Gaussian distributions for classification. In *Proc. of ICASSP*, pages 661–664, 1998.

[7] M.J.F. Gales. Semi-tied covariance matrices for hidden Markov models. *IEEE Trans. on Speech & Audio Process.*, 7(3):272–281, 1999.

[8] H. Kawahara, I. Masuda-Katsuse, and A.de Cheveigné. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f_0 extraction: possible role of a repetitive structure in sounds. *Speech Communication*, 27:187–207, 1999.

[9] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. Hidden semi-Markov model based speech synthesis. In *Proc. of ICSLP*, pages 1185–1180, 2004.

[10] T. Toda and K. Tokuda. Speech parameter generation algorithm considering global variance for HMM-based speech synthesis. In *Proc. of Interspeech (Eurospeech)*, 2005.

[11] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai. Mel-generalized cepstral analysis – a unified approach to speech spectral estimation. In *Proc. of ICASSP*, pages 1043–1046, 1994.

[12] K. Koishida, K. Tokuda, T. Kobayashi, and S. Imai. Efficient encoding of mel-generalized cepstrum for CELP coders. In *Proc. of ICASSP*, pages 1355–1358, 1997.

[13] M. Marume, H. Zen, Y. Nankaku, K. Tokuda, and T. Kitamura. An investigation of spectral parameters for HMM-based speech synthesis. In *Proc. of Autumn Meeting of ASJ*, 2006. (In Japanese).

[14] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In *Proc. of Eurospeech*, pages 2347–2350, 1999.

[15] http://hts.ics.nitech.ac.jp/nitech-naist-hts_blizzard2006/.

[16] C. Benot, M. Grice, and V. Hazan. The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences. *Speech Communication*, 18:381–392, 1996.