PERFORMANCE EVALUATION OF THE SPEAKER-INDEPENDENT HMM-BASED SPEECH SYNTHESIS SYSTEM "HTS-2007" FOR THE BLIZZARD CHALLENGE 2007

Junichi Yamagishi[1], *Takashi Nose*[2], *Heiga Zen*[3], *Tomoki Toda*[4], *Keiichi Tokuda*[3]

[1]University of Edinburgh, [2]Tokyo Institute of Technology,
[3]Nagoya Institute of Technology, [4]Nara Institute of Science and Technology,
jyamagis@inf.ed.ac.uk, takashi.nose@ip.titech.ac.jp,
zen@sp.nitech.ac.jp, tomoki@is.naist.jp, tokuda@nitech.ac.jp

ABSTRACT

This paper describes a speaker-independent/adaptive HMM-based speech synthesis system developed for the Blizzard Challenge 2007. The new system, named "HTS-2007", employs speaker adaptation (CSMAPLR+MAP), feature-space adaptive training, mixed-gender modeling, and full-covariance modeling using CSMAPLR transforms, in addition to several other techniques that have proved effective in our previous systems. Subjective evaluation results show that the new system generates significantly better quality synthetic speech than that of speaker-dependent approaches with realistic amounts of speech data, and that it bears comparison with speaker-dependent approaches even when large amounts of speech data are available.

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

1. INTRODUCTION

Statistical parametric speech synthesis based on hidden Markov models (HMMs) [1] has become established and well-studied, and is able to generate natural-sounding synthetic speech. In this framework, we have pioneered the development of the HMM Speech Synthesis System, HTS (H Triple S) [2]. Several high-quality text-to-speech synthesis systems have been developed with HTS [3, 4, 5], and they have demonstrated good performance in the Blizzard Challenges, which are open evaluations of corpus-based text-to-speech (TTS) synthesis systems. In the Nitech-HTS system [3] used for the 2005 Blizzard Challenge, a high-quality speech vocoding method (STRAIGHT [6] with mixed excitation), hidden semi-Markov models (HSMMs) [7], and a parameter generation algorithm that considered the global variance (GV) [8] were integrated into the basic system [1, 2]. In the Nitech-NAIST-HTS system [4] for the Blizzard Challenge 2006, a semi-tied covariance (STC) [9] was used for full-covariance modeling in the HSMMs, and the structure of the covariance matrices for the GV pdfs was changed from diagonal to full. Furthermore, for the Blizzard challenges 2007, we developed the new HTS-2007 system [5] underpinned by a speaker-adaptive approach: speaker adaptation techniques (CSMAPLR+MAP); adaptive training for the HSMMs; mixed-gender modeling; and full covariance modeling using the CSMAPLR transforms in addition to the above techniques. However, we could not fully evaluate the new system in [5] because of its tight schedule and the time-consuming training procedures of the HTS-2007 system.

Thus, we report several results for the performance evaluation of the HTS-2007 system and the past systems in this paper. Since the effects on the speaker adaptation and adaptive training for the HSMMs were reported in [10, 11] in detail, we focus on the following two interesting and beneficial aspects – analysis of the speaker-dependent and speaker-adaptive approaches from the viewpoint of the amount of the speech data, and comparison of the full-covariance modeling techniques. We have already analyzed several comparative merits and demerits of the speaker-dependent and speaker-adaptive approaches between 1 and 60 minutes of speech data using the conventional systems, and we found that synthetic speech using the speaker-adaptive approach was perceived as being more natural sounding than that of the speaker-dependent approach within the amount of the speech data [12, 13]. Therefore, it would be very interesting for us to extensively investigate the aspect of these approaches in the latest systems using much more larger amount of speech data. Then, we assess the effect on the CSMAPLR-based full-covariance modeling compared to the semi-tied covariance and diagonal covariance. Although CSMAPLR is a speaker adaptation method rather than a full-covariance modeling method, it has the same transforms for the covariance matrices as STC and the additional MAP adaptation estimates diagonal elements of the covariance matrix in a similar way to updating processes for STC. For CSMAPLR, multiple transforms are estimated using the robust SMAP criterion [14], which is expected to alleviate the artificiality and to improve the quality of synthetic speech as well as STC. We show their effectiveness from several subjective evaluation results using English and Japanese speech synthesis systems.

2. OVERVIEW OF THE HTS 2007 SYSTEM

2.1. Speaker-Adaptive Approach

To simultaneously model the STRAIGHT mel-cepstral coefficients, $\log F_0$, and aperiodicity measures (which are parameters for the STRAIGHT mel-cepstral vocoder with mixed excitation) together with duration in a unified modeling framework, we utilize multi-stream left-to-right MSD-HSMMs as acoustic units for speech synthesis. Using the MSD-HSMMs, we train the the mixed-gender average voice model as the initial model of the adaptation from training data which consists of several speakers' speech. Note that we include adaptation data for the target speakers in the training data for the average voice model since the aim of this approach is not rapid speaker adaptation but generating high-quality synthetic speech. To construct an appropriate average voice model, we utilize a feature-space SAT algorithm and a decision-tree-based context and gender clustering for the estimation and tying of the model parameters of the average voice model, respectively.

At the speaker adaptation stage, we adapt the mixed-gender average voice model to the target speaker using a combination of the CSMAPLR adaptation and the MAP adaptation techniques. The CSMAPLR adaptation simultaneously transforms the mean vector

$\boldsymbol{\mu}_i$ and diagonal covariance matrix $\boldsymbol{\Sigma}_i$ of a Gaussian pdf i using the same transforms as follows:

$$\overline{\boldsymbol{\mu}}_i = \boldsymbol{\zeta}_k \boldsymbol{\mu}_i + \boldsymbol{\epsilon}_k, \quad (1)$$

$$\overline{\boldsymbol{\Sigma}}_i = \boldsymbol{\zeta}_k \boldsymbol{\Sigma}_i \boldsymbol{\zeta}_k^{\top}. \quad (2)$$

Then, the SMAP criterion is used to robustly estimate $\boldsymbol{\zeta}_k$ and $\boldsymbol{\epsilon}_k$. In the SMAP estimation, tree structures of the distributions effectively cope with control of hyperparameters. Specifically, we first estimate a global transform at the root node of the tree structure using all adaptation data, and then propagate the transform to its child nodes as their hyperparameters. In the child nodes, transforms are estimated again using their adaptation data, based on the MAP estimation with the propagated hyperparameters. Then, the recursive MAP-based estimation of the transforms from the root node to lower nodes is conducted. For the tree structures of the distributions, we utilize the decision trees for context clustering because the decision trees have phonetic and linguistic contextual questions related to the suprasegmental features by which prosodic features, especially F_0, are characterized. Then since the CSMAPLR adaptation algorithm is a rough piecewise linear regression, we update the model using the MAP adaptation to modify the adapted parameters which have a relatively large amount of speech data from the target speaker.

This speaker-adaptive approach can surprisingly surpass the speaker-dependent approach under certain circumstances. This is mainly due to the relation between the amount of training data and the size of decision trees for the clustering of Gaussian distributions of the HSMMs. To cope with problems of data sparsity and unseen context-dependent HMMs, we utilize the MDL criterion and build the decision trees for clustering of distributions. As a result, the decision trees for the average voice model, which can easily collect a lot of speech data, becomes larger and more precise than those for the speaker-dependent model in general. Although topology of the large speaker-independent trees is not optimal for the target speakers, we confirmed that the naturalness of synthetic speech generated from the adapted models is correlate closely with the the size of the decision trees and better than that of speaker-dependent models [12].

2.2. Speaker Adaptation and Full-Covariance Modeling

In [4], it is reported that full covariance modeling using STC [9] has effect on the parameter generation algorithm considering global variance [8]. As we can see from Eq. (2), we may use the CSMAPLR transforms for the purpose of the full covariance modeling, since $\boldsymbol{\Sigma}_i$ is a diagonal covariance matrix and $\boldsymbol{\zeta}_k$ is a square matrix. In order to precisely model the full covariance, the following updating procedures are used.

1. Train all the parameters for the average voice model.
2. Using the current transforms $(\boldsymbol{\zeta}_k, \boldsymbol{\epsilon}_k)$ and the average voice model, estimate the new transforms $(\hat{\boldsymbol{\zeta}}_k, \hat{\boldsymbol{\epsilon}}_k)$ based on the SMAP criterion.
3. Using the estimated transforms $(\hat{\boldsymbol{\zeta}}_k, \hat{\boldsymbol{\epsilon}}_k)$ and the current average voice model, estimate $\boldsymbol{\mu}_i$, $\boldsymbol{\Sigma}_i$ and weight for the average voice model based on the MAP criterion.
4. Go to step 2 until convergence, or appropriate criterion satisfied.
5. Transform the covariance matrices to full covariance using the updated parameters. Transform the mean vectors as well.

Then, in order to assess the effect on only the SMAP criterion and multiple transforms, a combination algorithm with a single STC transform is also investigated. The following updating procedures are used.

6. Diagonalize covariance matrices of the transformed model in the above step 5.
7. Update the mean, diagonalized covariance, and weight of the transformed model based on the MAP criterion. Repeat the update.
8. Using the current semi-tied transform, estimate diagonal elements of the covariance matrices based on the MAP criterion.
9. Using the estimated diagonal elements of the covariance matrices, estimate the current semi-tied transform, which is equivalent to the transform of only the covariance matrices of Eq. (2), based on the ML criterion.
10. Go to step 8 until convergence, or appropriate criterion satisfied.
11. Transform the covariance matrices to full covariance using the estimated semi-tied transform.

We compare diagonal covariance in the step 7 with the CSMAPLR-based full-covariance in the step 5 and semi-tied covariance in the step 11.

3. EXPERIMENTS

3.1. Experimental Conditions

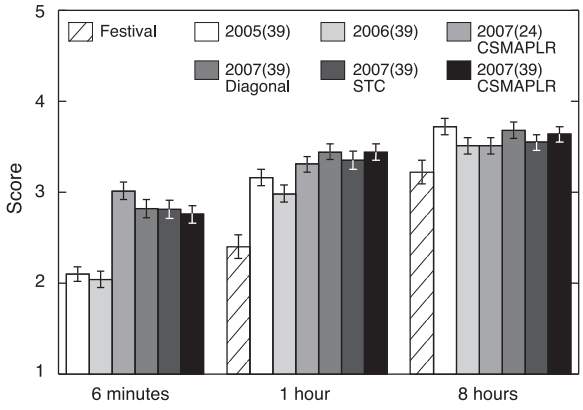
We conducted experiments for U.S. English speech synthesis using Nitech-HTS 2005, Nitech-NAIST-HTS 2006, and HTS-2007 systems. In this section, we report on results using the CMU-ARCTIC and ATRECSS speech databases [15]. The CMU-ARCTIC speech database contains a set of approximately one thousand phonetically balanced sentences uttered by four male speakers (AWB, BDL, JMK, and RMS) and two female speakers (CLB and SLT), with a total duration of about six hours. The ATRECSS speech database was released from ATR to be used in the 2007 Blizzard Challenge and contains the same sentences as CMU-ARCTIC, together with additional sentences, all uttered by a male speaker (EM001), with a duration of about eight hours.

Speech signals were sampled at a rate of 16 kHz and windowed by an F_0-adaptive Gaussian window with a 5 ms shift. The feature vectors consisted of 25 or 39 STRAIGHT mel-cepstral coefficients (including the zeroth coefficient), $\log F_0$, aperiodicity measures, and their dynamic and acceleration coefficients. We used 5-state left-to-right context-dependent multi-stream MSD-HSMMs without skip paths. Each state had a single Gaussian pdf with a diagonal covariance matrix. For the further details such as training time, footprints, the number of leaf nodes of the decision trees, please refer to [5].

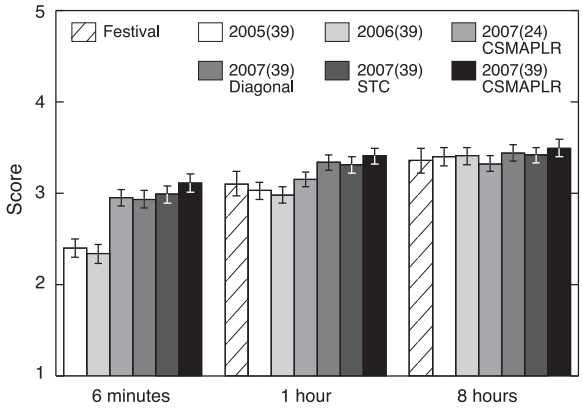
3.2. Evaluation Results of the English Systems

To investigate the effect of the amount of speech data available, we built each system using sets of sentences spoken by target speaker EM001. These consisted of 100 randomly chosen CMU-ARCTIC sentences (about six minutes in duration), all 1032 CMU-ARCTIC sentences (one hour duration), and all 6579 Blizzard sentences (eight hours duration). At the same time, the HTS-2007 systems using the diagonal covariance and semi-tied covariance were also built to evaluate the full-covariance modeling techniques. In addition, we built the HTS-2007 systems using either 24 or 39 order STRAIGHT mel-cepstral coefficients for each voice, in order to investigate the effect of the model order of the STRAIGHT mel-cepstra. In all HTS-2007 systems, all the speech data included in the CMU-ARCTIC database was used as part of the training data for the average voice model. For reference, the Festival speech synthesis system [16] using the same speech data of the speaker EM001 was also evaluated as a baseline unit-selection speech synthesis system.

We evaluated naturalness using mean opinion score (MOS) tests and similarity to a reference using CCR tests. The reference speech



(a) MOS test : Naturalness



(b) CCR test : Similarity

Fig. 1. Subjective evaluation of the English HTS-2007 and past systems. Target speaker is the English male speaker EM001.

included two recorded sentences spoken by target speaker EM001. In those tests, 33 subjects were presented with a set of synthetic speech utterances generated from the systems in random order. For each subject, 14 semantically unpredictable test sentences (as used in Blizzard 2007 [17]) were randomly chosen from a set of 50 test sentences. Subjects were asked to rate them using a 5-point scale, where 5 corresponded to natural (MOS test) or very similar (CCR test), and 1 corresponded to poor (MOS test) or very dissimilar (CCR test).

Figure 1 shows the mean scores, with 95% confidence interval, of the MOS and CCR tests. For both tests, there are significant differences between the HTS-2007 systems and the speaker-dependent systems when six minutes or one hour of target speech data is used. As the amount of training data available decreases, the differences become more significant. However, even in the case of eight hours of target speech data, they were still comparable. In order to make this speaker-adaptive approach beneficial even for large amounts of target speech data, we should train the average voice model from much larger amounts of speech data.

Further results from these experiments concern feature dimensionality and covariance modelling. It is apparent the HTS-2007 system using 39 dimension mel-cepstra was shown to be less natural than that using 24 dimension mel-cepstra only in the case of six minutes of target speech data, due to the number of additional parameter that needs to be estimated for the linear transform in the case of higher feature dimensionality. Although CSMAPLR-based full-covariance modeling had the highest values in the CCR test, the differences were not significant. We discuss the effect of the full-covariance modeling more fully in the next subsection. Finally, we can see that naturalness of synthetic speech generated from the Festival unit-selection speech synthesis system becomes much worse as the amount of target speech data becomes smaller. Moreover it can be also seen that synthetic speech generated from the HTS-2007 system using about six minutes of speech data was rated to be more natural than that of the unit-selection approach using about one hour of speech data.

In summary, the speaker-independent/adaptive HTS-2007 system is rated, in subjective listening tests, to be similar to the speaker-dependent approach even in the case of several hours of target speech data, and to be significantly better than the speaker-dependent approach in the case of less target speech data. This improvement is at the cost of an increased number of model parameters, compared with speaker-dependent systems, resulting in a larger memory footprints. Moreover, the training procedures for the speaker-independent/adaptive system are considerably more computationally demanding.

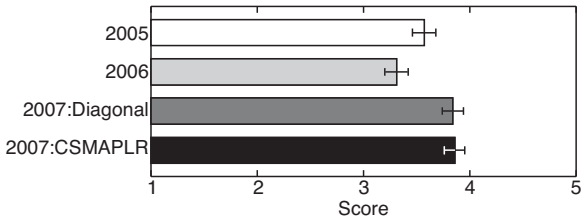
3.3. Evaluation Results of the Japanese Systems

We also conducted experiments for Japanese speech synthesis using Nitech-HTS 2005, Nitech-NAIST-HTS 2006, and HTS-2007 likewise. For the Japanese systems, we used three data sets: The ATR Japanese speech database Set B, containing a set of 503 phonetically balanced sentences uttered by ten speakers (six male: MHO, MHT, MMY, MSH, MTK, and MYI; four female: FKN, FKS, FTK, and FYM), with a duration of about 30 minutes per speaker; The ATR Japanese speech database Set C, containing a set of 100 phonetically balanced sentences each uttered by a female speaker (F109) and a male speaker (M109), about a duration of about six minutes per speaker; A database which contains the same sentences as those of the ATR Japanese speech database (Set B) uttered by a female speaker (FTY) and three male speakers (MJI, MMI, and M001), also with a duration of about 30 minutes per speaker. The sizes of these speech corpora were about five hours, twelve minutes, and two hours, respectively. From these speech databases, we utilized eight males (MHO, MHT, MMY, MSH, MTK, MYI, MJI, and MMI) and five females (FKN, FKS, FYM, FTY, and FTY) for both the training and adaptation, and used the rest of two males (M109 and M001) and female (F109) for only the adaptation.

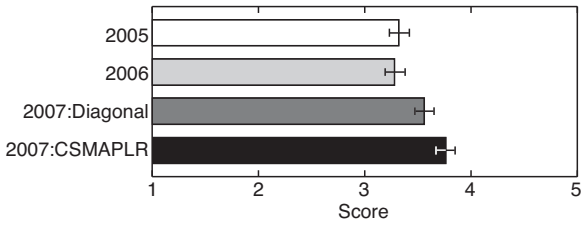
3.4. Evaluation Results of the Japanese Systems

Although the effect of full-covariance modeling in the English experiment above were not statistically significant, we found in preliminary experiments that the effect of full-covariance modeling varies by speaker. Thus, in this experiment, we used seven target speakers (FTY, FYM, MJI, MYI, M001, F109, and M109), with about thirty minutes of adaptation data for each of the first five speakers, and about six minutes of adaptation data obtained for the latter two. The evaluation methods that we employed were the same MOS and CCR tests as in the above English experiments. Ten male subjects were used, each hearing six test sentences randomly chosen from 50 test sentences from ATR Set B.

Figure 2 shows the mean scores with 95% confidence interval for the MOS and CCR tests using the seven target speakers. In both the MOS and CCR tests, there are significant differences between the speaker-independent/adaptive and speaker-dependent systems. Since the amount of speech data used for the target speakers is relatively small, the HTS-2007 system could generate better quality synthetic speech than that of speaker-dependent systems. These results correspond well to those obtained above. Further, it can be seen



(a) MOS test : Naturalness



(b) CCR test : Similarity

Fig. 2. Subjective evaluation of the Japanese HTS-2007 and past systems. Target speakers are seven Japanese speakers.

that CSMAPLR-based full-covariance modeling slightly improves similarity of synthetic speech compared to that using diagonal covariance.

4. CONCLUSIONS

We have described the evaluation of a speaker-independent/adaptive HMM-based speech synthesis system for the Blizzard Challenge 2007. A number of new features were incorporated, which underpin the speaker-adaptive approach: CSMAPLR+MAP speaker adaptation, feature-space adaptive training for HSMMs, mixed-gender modeling and full-covariance modeling using the CSMAPLR transforms. Several subjective and objective evaluation results including the Blizzard Challenge 2007 show that the new system generates high quality speech. In particular, it is able to synthesize speech that is significantly better than the speaker-dependent approaches and unit-selection approaches in the case of realistic amounts of target speaker data, and bears comparison with the speaker-dependent approaches even with the large amount of speech data[1]. We also shown the full-covariance modeling using the CSMAPLR transforms improved similarity of synthetic speech. However, there remain a number of issues related to the efficiency of acoustic modeling and the computational requirements for training. Our future work is to deal with these issues and to further develop the framework to enable unsupervised speaker adaptation for speech synthesis.

5. ACKNOWLEDGMENTS

This research was conducted for the purpose of the Blizzard Challenge 2007.

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," in *Proc. EUROSPEECH-99*, Sept. 1999, pp. 2374–2350.

[1]The Nitech-HTS 2005 and HTS-2007 (excluding full-covariance modeling) systems are available at the CSTR Festival on-line demonstration. http://www.cstr.ed.ac.uk/projects/festival/onlinedemo.html

[2] K. Tokuda, H. Zen, J. Yamagishi, T. Masuko, S. Sako, A.B. Black, and T. Nose, *The HMM-based speech synthesis system (HTS) Version 2.0.1*, http://hts.sp.nitech.ac.jp/.

[3] H. Zen, T. Toda, M. Nakamura, and K. Tokuda, "Details of Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 1, pp. 325–333, Jan. 2007.

[4] H. Zen, T. Toda, and K. Tokuda, "The Nitech-NAIST HMM-based speech synthesis system for the Blizzard Challenge 2006," in *Proc. Blizzard Challenge 2006*, Sept. 2006.

[5] J. Yamagishi, H. Zen, T. Toda, and K. Tokuda, "Speaker-independent HMM-based speech synthesis system — HTS-2007 system for the Blizzard Challenge 2007," in *Proc. BLZ3-2007 (in Proc. SSW6)*, Aug. 2007.

[6] H. Kawahara, I. Masuda-Katsuse, and A. 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, pp. 187–207, 1999.

[7] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "A hidden semi-Markov model-based speech synthesis system," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 5, pp. 825–834, May 2007.

[8] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 5, pp. 816–824, May 2007.

[9] M.J.F. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Trans. Speech Audio Process.*, vol. 7, pp. 272–281, Mar. 1999.

[10] J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 2, pp. 533–543, Feb. 2007.

[11] J. Yamagishi, T. Kobayashi, M. Tachibana, K. Ogata, and Y. Nakano, "Model adaptation approach to speech synthesis with diverse voices and styles," in *Proc. ICASSP 2007*, Apr. 2007, pp. 1233–1236.

[12] K. Ogata, M. Tachibana, J. Yamagishi, and T. Kobayashi, "Acoustic model training based on linear transformation and MAP modification for HSMM-based speech synthesis," in *Proc. ICSLP 2006*, Sept. 2006, pp. 1328–1331.

[13] J. Yamagishi, T. Kobayashi, S. Renals, S. King, H. Zen, T. Toda, and K. Tokuda, "Improved average-voice-based speech synthesis using gender-mixed modeling and a parameter generation algorithm considering GV," in *Proc. of 6th ISCA Workshop on Speech Synthesis*, Aug. 2007.

[14] K. Shinoda and C.H. Lee, "A structural Bayes approach to speaker adaptation," *IEEE Trans. Speech Audio Process.*, vol. 9, pp. 276–287, Mar. 2001.

[15] J. Ni, T. Hirai, H. Kawai, T. Toda, K. Tokuda, M. Tsuzaki, R.M̃aia S.S̃akai, and S. Nakamura, "Atrecss - atr english speech corpus for speech synthesis," in *Proc. BLZ3-2007 (in Proc. SSW6)*, Aug. 2007.

[16] K. Richmond, V. Strom, R. Clark, J. Yamagishi, and S. Fitt, "Festival Multisyn voices for the 2007 Blizzard Challenge," in *Proc. BLZ3-2007 (in Proc. SSW6)*, Aug. 2007.

[17] M. Fraser and S. King, "The blizzard challenge 2007," in *Proc. BLZ3-2007 (in Proc. SSW6)*, Aug. 2007.