Speaker-Independent HMM-based Speech Synthesis System — HTS-2007 System for the Blizzard Challenge 2007

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

[1]University of Edinburgh, [2]Nagoya Institute of Technology,
[4]Nara Institute of Science and Technology,

jyamagis@inf.ed.ac.uk, zen@sp.nitech.ac.jp, tomoki@is.naist.jp, tokuda@nitech.ac.jp

Abstract

This paper describes an HMM-based speech synthesis system developed by the HTS working group for the Blizzard Challenge 2007. To further explore the potential of HMM-based speech synthesis, we incorporate new features in our conventional system which underpin a speaker-independent approach: speaker adaptation techniques; adaptive training for HSMMs; and full covariance modeling using the CSMAPLR transforms.

1. Introduction

A statistical parametric speech synthesis system based on hidden Markov models (HMMs) [1] can easily and flexibly generate natural sounding synthetic speech. We have developed high-quality speaker-dependent speech synthesis systems and shown their performance in the past Blizzard challenges [2, 3].

In addition to the speaker-dependent systems, we have also been developing a speaker-independent HMM-based speech synthesis in which statistical "average voice models" are created from several speakers' speech data and are adapted with a small amount of speech data from a target speaker (e.g. [4]). This research started by directly using several speaker adaptation techniques developed for automatic speech recognition such as maximum likelihood linear regression (MLLR) [5] as transformation techniques for spectral parameters of speech [6]. Then, in order to simultaneously model and adapt excitation parameters of speech as well as spectral parameters, we developed the multi-space probability distribution (MSD) HMM [7] and its MLLR adaptation algorithm [8]. We utilize logarithmic fundamental frequency ($\log F_0$) as the excitation parameter, and the MSD-HMM enables us to treat the $\log F_0$ observation, which is a mixture observation of a one-dimensional real number for voiced regions and a symbol string for unvoiced regions, within a generative model. Furthermore, in order to simultaneously model and adapt duration parameters for the spectral and excitation parameters as well, we developed the MSD hidden semi-Markov model (MSD-HSMM) [9] and its MLLR adaptation algorithm [4]. The HSMM [10] is an extended HMM, having explicit state duration distributions instead of the transition probabilities to directly model and control state durations. More advanced speaker adaptation techniques including constrained structural maximum a posteriori linear regression (CSMAPLR) [11] were also defined within the framework of the MSD-HSMM. In addition to the improvements of the speaker adaptation techniques, several training techniques for an initial model for the above speaker adaptation techniques have also been developed. As the initial model, we utilize an average voice model constructed from training data which consists of several speakers' speech. However, the training data of the average voice model includes a lot of speaker-dependent characteristics, and they crucially affect the adapted models and the quality of synthetic speech generated from them. Therefore, we have incorporated the speaker-adaptive training (SAT) algorithm [12] into our speech synthesis system for normalizing the negative influence of speaker differences [13, 14]. In the SAT algorithm, the model parameters for the average voice model are blindly estimated based on an assumption that the speaker difference is expressed by linear transformations of the average voice model. An application to multilingual/polyglot TTS systems is also proposed [15]. By using this framework, we can obtain synthetic speech for a target speaker from as little as 100 utterances (about 6 minutes). Interestingly, we have shown that synthetic speech using this approach is perceived as being more natural sounding than that of speaker-dependent (SD) systems trained on between 30 and 60 minutes of speech data by many listeners because of the data-rich average voice model [4, 16, 14].

At first sight, this seems strange; we cannot obtain a perfect speaker-depend model even from 60 minutes of speech data. However, the average voice model trained on a much larger amount of speech data can provide a lot of prior information of speech data, since the speech of multiple speakers exhibits similar pattern or tendency to some degree. Hence, when the prior information is appropriate for the target speaker, it can reduce error and improve the estimator for the target speaker's model. Even when the prior is not useful for the target speaker, it can provide a maximum likelihood estimator from the target speaker's data. In addition, there is a famous statistics paradox (called "Stein's paradox" [17]) intimately linked with this phenomenon. Stein proves that when we estimate more than 3 random variables, an empirical Bayes estimator [18] (in which training data for these random variables is used for the estimation of hyperparameters of their prior distributions) is better than a standard maximum likelihood estimator (i.e., when we have data from 3 people and estimate a mean value per person, the mean value estimated by the empirical Bayes with the prior distributions created from 3 persons' data is better than a mean value separately estimated by a likelihood method with a person's data only). In fact, our speaker-independent method utilizes maximum a posteriori (MAP) algorithm of the empirical Bayes method as one of speaker adaptation algorithms. Therefore, the speaker-independent approach has the potential to surpass the common speaker-dependent approach, and it would be very interesting to investigate the aspect via several evaluations for the Blizzard Challenge 2007.

2. Overview of the HTS 2007 System

In the Nitech-HTS2005 system [2], high-quality speech vocoding methods (*STRAIGHT* with mixed excitation) [19], HSMMs,

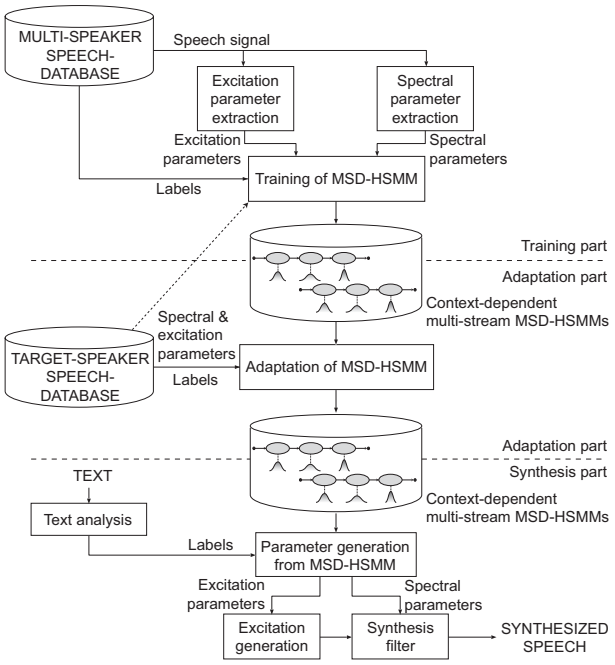

Figure 1: Overview of the HTS-2007 speech synthesis system.

and a parameter generation algorithm considering global variance (GV) [20] were integrated. The Nitech-HTS2006 system [3] additionally adopted semi-tied covariance [21] (also known to as maximum likelihood linear transform [22]) for full covariance modeling. Our new system "HTS-2007" incorporates the following new features into the above systems as well:

- Using a speaker-independent approach instead of the speaker-dependent approaches
- Applying adaptation and adaptive training techniques to HSMMs instead of the standard ML training
- Using the CSMAPLR transforms for full covariance modeling instead of the semi-tied covariance transform

This new system consists of a speech analysis part, training part for the average voice model, speaker adaptation part, and speech synthesis part as shown in Fig. 1.

2.1. Speech Analysis

In this system, we use three kinds of parameters for the STRAIGHT mel-cepstral vocoder with mixed excitation, that is, the STRAIGHT mel-cepstrum [2], $\log F_0$, and aperiodicity measures. This is the same as the speaker dependent system Nitech-HTS 2005. The mel-cepstral coefficients are obtained from STRAIGHT spectral analysis [19] in which F_0-adaptive spectral smoothing is carried out in the time-frequency region. The F_0 values are estimated using the following three-stage extraction to reduce error of F_0 extraction such as halving and doubling and to suppress voiced/unvoiced error. First, using the instantaneous frequency amplitude spectrum (IFAS) based method [23], the system extracts F_0 values for all speech data of each speaker within a common search range. Second, the F_0 range of each speaker is roughly determined based on a histogram of the extracted F_0 values. Third, F_0 values are re-extracted in the speaker-specific range using the IFAS algorithm, fixed-point analysis [24], and ESPS get-F_0 [25]. Finally,

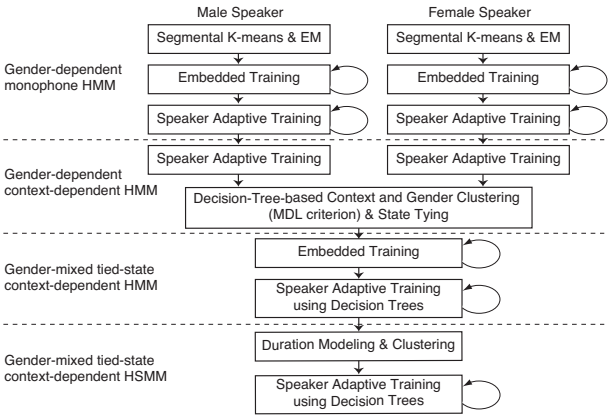

Figure 2: Details of training procedures.

a median value of the extracted F_0 values at each frame is utilized as an eventual F_0 value. The aperiodicity measures for mixed excitation are based on a ratio between the lower and upper smoothed spectral envelopes, and averaged on five frequency sub-bands: 0-1, 1-2, 2-4, 4-6, and 6-8 kHz. In addition to these static features, their dynamic and acceleration features are used.

2.2. Training

To simultaneously model the acoustic features together with duration in a unified modeling framework, we utilize context-dependent multi-stream left-to-right MSD-HSMMs as acoustic units for speech synthesis. The multi-stream model structure is used to simultaneously model the mel-cepstral coefficients, $\log F_0$, and aperiodicity measures. 53 English phonetic, prosodic, and linguistic features used for context-dependent labels contain phonetic, segment-level, syllable-level, word-level, phrase-level, and utterance-level features. Details of these features are given in [26]. In addition to this phonetic and linguistic information, we added the gender of speakers to the context labels for conducting mixed-gender modeling. Note that phoneme boundary labels are used only to obtain the initial model parameters of the average voice model and we do not require the phoneme boundary labels at the adaptation or synthesis stage.

Using the above MSD-HSMMs, we train the average voice model as the initial model of the adaptation from training data which consists of several speakers' speech. In this Blizzard Challenge, we included adaptation data for the target speaker EM001 in the training data for the average voice model, since the amount of speech data for the target speaker exceeded that for the average voice model. To construct an appropriate average voice model, we utilize a feature-space SAT algorithm and a decision-tree-based context and gender clustering [14] for the estimation and tying of the model parameters of the average voice model, respectively. The feature-space SAT algorithm can be viewed as a generalized version of various normalization techniques such as cepstral mean normalization (CMN), cepstral variance normalization (CVN), vocal tract length normalization (VTLN), and bias removal of F_0 and duration.

The actual training procedures for the training of the average voice model are shown in Fig. 2. In order to conduct both normalization of the speaker-dependent characteristics and conservation of the gender-dependent characteristics, we first train gender-dependent HMMs using the SAT algorithm. Then, the

decision-tree-based context and gender clustering technique using minimum description length (MDL) criterion [27] is applied to the HMMs, and the model parameters of the HMMs at each leaf node of the decision trees are tied. We assume that the CMLLR transforms for the SAT algorithm remain unchanged during the clustering, and calculate the description length in a similar way to [27]. Then we re-estimate the clustered HMMs using a SAT algorithm with piecewise linear regression functions. To determine regression classes for the piecewise linear regression, the decision trees constructed for the gender-mixed model are used, since use of the decision tree automatically reflects both differences of gender information and phonetic and linguistic information, and it is expected that more appropriate normalization for the average voice model is achieved. We then calculate initial duration pdfs from trellises which are obtained from the SAT algorithm, and conduct decision-tree-based context and gender clustering for the duration pdfs. Using the tied duration pdfs, we perform the SAT algorithm for the HSMMs with piecewise linear regression functions in order to normalize speaker characteristics included in the duration pdfs as well as other acoustic features. At the SAT stages, we first estimated CMLLR transforms three times, and then updated mean vectors of both output and duration pdfs, their covariance matrices, space weights for MSD, and transition probabilities five times.

2.3. Speaker Adaptation

At the speaker adaptation stage, we adapt the average voice model to the target speaker using speaker adaptation techniques for the multi-stream MSD-HSMM. Here we use a combination of the CSMAPLR adaptation and the MAP adaptation techniques [28]. The CSMAPLR adaptation simultaneously transforms the mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ of a Gaussian pdf i using the same transforms as follows:

$$\overline{\boldsymbol{\mu}}_i = \boldsymbol{\zeta}\boldsymbol{\mu}_i + \boldsymbol{\epsilon}, \tag{1}$$

$$\overline{\boldsymbol{\Sigma}}_i = \boldsymbol{\zeta}\,\boldsymbol{\Sigma}_i\,\boldsymbol{\zeta}^{\top}. \tag{2}$$

Then, structural maximum a posteriori (SMAP) estimation [29] is used to robustly estimate $\boldsymbol{\zeta}$ and $\boldsymbol{\epsilon}$. 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 (Fig. 3). 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. Hence, the propagated prior information would automatically reflect the connection and similarity of the distributions in keeping with the linguistic information.

Another advantage of the CSMAPLR adaptation is that we can efficiently construct covariance models. In [3], it is reported that full covariance modeling based on semi-tied covariance [21] has effect on the speech parameter generation algorithm considering GV. In this system, we use the CSMAPLR transform for the purpose of the full covariance modeling instead of the semi-tied covariance.

Then we additionally adopt the MAP adaptation [28] to modify the adapted model parameters which have a relatively

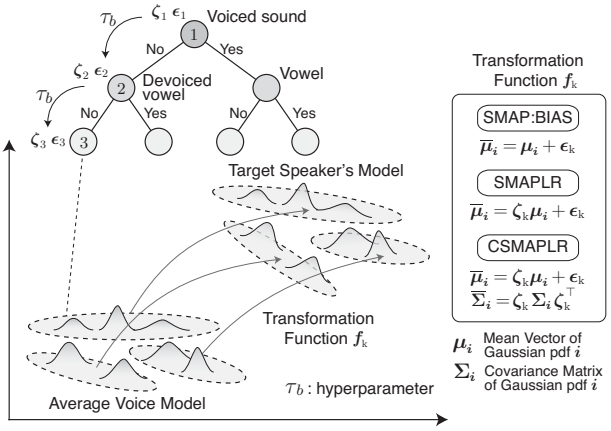

Figure 3: CSMAPLR adaptation

large amount of speech data from the target speaker, since the CSMAPLR adaptation algorithm has a rough assumption that the target speaker model would be expressed by the piecewise linear regression of the average voice model.

2.4. Speech Synthesis

At the synthesis stage, input text is transformed into a sequence of context-dependent phoneme labels with the gender information of the target speaker. Based on the label sequence, a sentence MSD-HSMM is constructed by concatenating context-dependent MSD-HSMMs. After the duration pdfs automatically determine state durations of the sentence MSD-HSMM, the mel-cepstrum, $\log F_0$, and aperiodicity-measure sequences are generated using the speech parameter generation algorithm considering GV. This is a penalized maximum likelihood method in which the GV pdf (a Gaussian pdf for the variance of the trajectory at utterance level) acts as a penalty for the likelihood function. The algorithm tries to keep the global variance of the generated trajectory as wide as that of the target speaker, while maintaining an appropriate parameter sequence in the sense of maximum likelihood. It is possible to adapt the GV pdf from a speaker-independent model to that of a target speaker using MAP adaptation. However, the number of parameters of a GV pdf is very small. Specifically, it is equal to the dimensionality of the static features. Hence we directly estimate the GV pdf from the adaptation data. The generation method for speech waveforms is identical to that of Nitech-HTS 2005. An excitation signal is generated from one-pitch waveforms, which are synthesized with the mixed excitation, using PSOLA, and then a synthesized waveform is generated using the MLSA filter corresponding to the STRAIGHT mel-cepstral coefficients.

3. Experiments

3.1. Experimental Conditions

To construct the HTS-2007 system, we used the CMU-ARCTIC speech database, which contains a set of approximately one thousand phonetically balanced sentences uttered by four male speakers (AWB, BDL, JMK, RMS) and two female speakers (CLB, SLT), and a speech database, which was released from ATR for the purpose of the Blizzard Challenge 2007 and contains the same sentences as that of CMU-ARCTIC speech database and additional sentences uttered by a male speaker EM001. The sizes of these speech corpora were six and eight

Table 1: CPU time (day) for construction of each system of each voice.

(a) Voice B (ARCTIC sentences: 1 hour)

	Order of mel-cepstral analysis	
System	24	39
Nitech-HTS 2005	0.5 day	0.8 days
Nitech-NAIST-HTS 2006	0.7 day	1 days
HTS-2007	8 days	15 days

(b) Voice A (6579 sentences: 8 hours)

	Order of mel-cepstral analysis	
System	24	39
Nitech-HTS 2005	1.5 day	2.5 days
Nitech-NAIST-HTS 2006	1.7 day	2.8 days
HTS-2007	20 days	36 days*

* In progress

hours for the CMU-ARCTIC and the Blizzard Challenge 2007 corpora, respectively. To model the synthesis units, we used the U.S. English phone set "radio" of the Festival speech synthesis system, and took the phonetic and linguistic contexts (obtained from utterance files included in these corpora for the Festival speech synthesis system) into account without any modifications.

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 24 or 39 STRAIGHT mel-cepstral coefficients (including the zeroth coefficient), $\log F_0$, aperiodicity measures, and their dynamic and acceleration coefficients. We used five-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. In the speaker adaptation and adaptive training, each CMLLR or CSMAPLR transform was a triblock diagonal matrix corresponding to the static, dynamic, and acceleration coefficients.

3.2. Details of Voice A and Voice B Systems

To investigate the effect of the corpus size, the organizers requested the participants to submit three systems: the first system was trained using all the speech data included in the released database (Voice A), the second system was trained using only the ARCTIC subset (Voice B), and the third system was trained using a freely selected subset having the same amount of speech data as that of the ARCTIC subset (Voice C). Because of the time-consuming training procedures of our system, we constructed the Nitech-HTS 2005, Nitech-NAIST-HTS 2006, and HTS-2007 systems for the Voices A and B only. We undertook to construct two kinds of systems using 24 or 39 STRAIGHT mel-cepstral coefficients for each voice. Unfortunately, the training for the HTS-2007 system using 39 STRAIGHT mel-cepstral coefficients did not finish by deadline for test utterance submission.

We utilized four grid computing clusters to construct our systems since we can concurrently conduct almost all the training procedures of the MSD-HSMMs per state, per speaker, and/or per subset. These grid computing clusters have from 16 to 136 cores with from 1 GB RAM to 4GB RAM. The total number of these cores is 264. Table 1 shows CPU time for the construction of each system of each voice. Compared with the speaker-dependent systems, the training for constructing the

Table 2: CPU time (hour) for each procedures of the HTS-2007 systems using 39 mel-cepstral coefficients

Procedures	Voice B	Voice A
Segmental K-means & EM	10h	19h
Embedded Training (Monophone HMM) (Mean&Variance×5)	2h	3h
SAT (Monophone HMM) (CMLLR×3, Mean&Variance×5)	7h	9h
SAT (Context-dependent HMM) (Mean&Variance×1)	2h	3h
Decision-Tree-based Clustering	3h	13h
Embedded Training (Tied-state HMM) (Mean&Variance×5)	2h	3h
SAT (Tied-state HMM) 3×(CMLLR×3, Mean&Variance×5)	30h	222h
Duration Modeling & Clustering	11h	34h
SAT (Tied-state HSMM) (CMLLR×3, Mean&Variance×5)	230h	553h
CSMAPLR+MAP Adaptation (HSMM)	28h	*
Total	325h	—

* In progress

HTS-2007 was an extremely time-consuming process. The reasons include that even the system for Voice B utilized a total of 7 hours of speech data and then that for Voice A utilized a total of 14 hours of speech data. In addition to the use of the large multi-speaker corpora, the SAT algorithms in particular required a lot of computation. Table 2 shows CPU time for each procedure of the HTS-2007 systems using 39 mel-cepstral coefficients. In these procedures, the SAT algorithm using the decision trees was the dominant cost, since this algorithm basically requires matrix operations for each state. Although we believe we can make these procedures much faster, we reluctantly submitted the systems using 24 STRAIGHT mel-cepstral coefficients this time because of our tight schedules. Hopefully, we would like to report the results of both systems in the near future.

Tables 3 and 4 show the number of leaf nodes of the constructed decision trees and footprints for each system of each voice, respectively. The numbers of leaf nodes for the Nitech-NAIST-HTS 2006 system are the same as those of the Nitech-HTS 2005. The footprints for the HTS-2007 systems completely depend on the condition of the speaker adaptation algorithms. For example, when we use a global transformation of the CSMAPLR adaptation only, the speaker-specific part of the footprints for the HTS-2007 system is just 40 to 55 kbytes. Then, the other parts of the footprint are common to all the speakers. However, in this Blizzard challenge, we focus not on the footprint size but on the quality of synthetic speech. Therefore, we utilized the combined algorithm of the piecewise CSMAPLR adaptation and MAP adaptation, and thereby it increased the footprint as shown in the Table 4.

4. Results of the Blizzard Challenge 2007

Figures 4 - 6 shows the evaluation results of the mean opinion score (MOS), average word error rate (WER), and similarity to original speaker on a 5-point scale of the Voice A and B of all systems participated in the Blizzard Challenge 2007. This year 16 groups participated in the challenge. In these figures, system "N" corresponds to the HTS-2007 system.

Table 3: The number of leaf nodes of constructed decision trees for each system of each voice.

(a) Voice B (ARCTIC sentences: 1 hour)

System	Mel-cepstrum	log F_0	Aperiodicity	Duration
2005 (24)	1,371	2,101	911	435
2005 (39)	961	2,096	850	459
2007 (24)	3,530	7,136	1,859	3,746
2007 (39)	2,508	13,034	1,735	3,557

(b) Voice A (6579 sentences: 8 hours)

System	Mel-cepstrum	log F_0	Aperiodicity	Duration
2005 (24)	6,959	11,174	4,590	5,702
2005 (39)	4,598	21,189	3,994	5,110
2007 (24)	7,273	14,245	3,740	8,580
2007 (39)	5,285	31,411	3,747	8,438

Table 4: Footprint (MBytes) for each system of each voice.

(a) Voice B (ARCTIC sentences: 1 hour)

	Mel-cepstral analysis	
System	24	39
Nitech-HTS 2005	1.64	1.70
Nitech-NAIST-HTS 2006	1.67	1.76
HTS-2007 (Diagonalized covariance)	4.43	5.38
HTS-2007 (Diagonal covariance + CSMAPLR transforms)	15.06	22.75

(b) Voice A (6579 sentences: 8 hours)

	Mel-cepstral analysis	
System	24	39
Nitech-HTS 2005	8.12	9.29
Nitech-NAIST-HTS 2006	8.15	9.35
HTS-2007 (Diagonalized covariance)	8.91	11.73
HTS-2007 (Diagonal covariance + CSMAPLR transforms)	39.24	49.10

Compared with the MOS results for the Nitech-NAIST-HTS 2006 system used for the Blizzard Challenge 2006 [3], we observe that the MOS results have become slightly worse. From preliminary experiments, we have found that differences in the order of the STRAIGHT mel-cepstral analysis affect MOS results. On the other hand, compared with the WER results for all the other systems, we can see that systems which can reach less than 30% of the WER in both the Voice A and B are J, M, and our system only. It is worth noting that, although we do not conduct any modifications of the released database, including speech and label files, the HTS-2007 system can provide good performance. However, the evaluation results of similarity to the original speaker indicate that the synthetic speech of the HTS-2007 system has relatively low similarity. We believe we can improve the similarity of synthetic speech by using systems with 39 or more STRAIGHT mel-cepstral coefficients.

5. Online Demonstration System and New HTS version 2.1

We plan to create new voices using this HTS-2007 system and release them for the purpose of an online demonstration system.[1] In addition to this, we also plan to integrate these new features and the following methods into future HTS releases:

- SMAPLR adaptation [30],
- Parameter generation using the GV pdf with covariance adaptation such as CMLLR [31] / CSMAPLR.

This new HTS (version 2.1, released March 2008 or later), with the STRAIGHT analysis/synthesis technique and F_0 extraction algorithms, will provide the ability to construct the above HTS-2007 system as well as our speaker-dependent systems developed for the past Blizzard Challenge events.

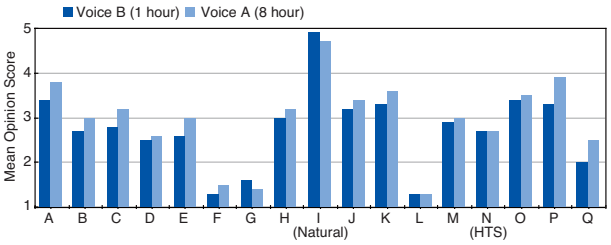

Figure 4: Mean opinion scores of all systems.

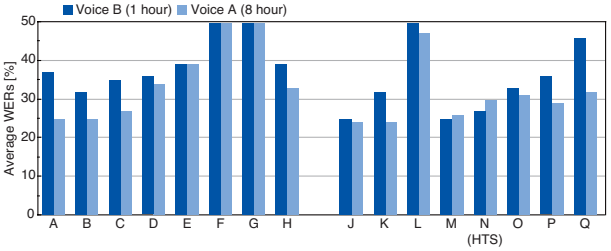

Figure 5: Average word error rate (%) of all systems.

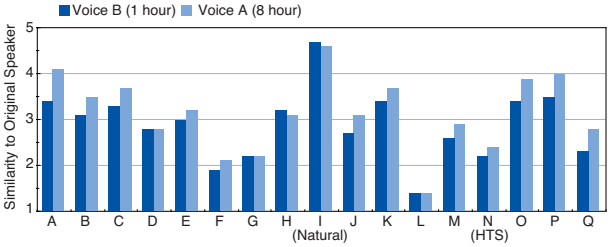

Figure 6: Similarity to original speaker of all systems.

6. Conclusions

This paper described an HMM-based speech synthesis system developed by the HTS working group for the Blizzard Challenge 2007. To further explore the potential of HMM-based speech synthesis, we incorporated new features in our conventional system which underpin a speaker-independent approach: speaker adaptation techniques; adaptive training for HSMMs; and full covariance modeling using the CSMAPLR transforms. Our future work is to analysis this system and compare it with our speaker-dependent systems in detail.

7. 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] http://www.cstr.ed.ac.uk/projects/festival/onlinedemo.html

[2] 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.

[3] 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.

[4] 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.

[5] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, no. 2, pp. 171–185, 1995.

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

[7] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Multi-space probability distribution HMM," *IEICE Trans. Inf. & Syst.*, vol. E85-D, no. 3, pp. 455–464, Mar. 2002.

[8] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR," in *Proc. ICASSP 2001*, May 2001, pp. 805–808.

[9] 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.

[10] S.E. Levinson, "Continuously variable duration hidden Markov models for automatic speech recognition," *Computer Speech and Language*, vol. 1, no. 1, pp. 29–45, 1986.

[11] Y. Nakano, M. Tachibana, J. Yamagishi, and T. Kobayashi, "Constrained structural maximum a posteriori linear regression for average-voice-based speech synthesis," in *Proc. ICSLP 2006*, Sept. 2006, pp. 2286–2289.

[12] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Proc. ICSLP-96*, Oct. 1996, pp. 1137–1140.

[13] J. Yamagishi, M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "A training method of average voice model for HMM-based speech synthesis," *IEICE Trans. Fundamentals*, vol. E86-A, no. 8, pp. 1956–1963, Aug. 2003.

[14] 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, (to be appear).

[15] J. Latorre, K. Iwano, and S. Furui, "New approach to the polyglot speech generation by means of an HMM-based speaker adaptable synthesizer," *Speech Communication*, vol. 48, no. 10, pp. 1227–1242, 2006.

[16] 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.

[17] B. Efron and C. Morris, "Stein's paradox in statistics," *Scientific American*, vol. 20, no. 5, pp. 451–468, 1977.

[18] P. Carlin and Thomas A. Louis, *Bayes and Empirical Bayes Methods for Data Analysis (Second edition)*, Chapman & Hall/CRC, 2000.

[19] 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.

[20] 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.

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

[22] R. Gopinath, "Maximum likelihood modeling with Gaussian distributions for classfication," in *Proc. ICASSP-98*, May 1998, pp. 661–664.

[23] D. Arifianto, T. Tanaka, T. Masuko, and T. Kobayashi, "Robust F0 estimation of speech signal using harmonicity measure based on instantaneous frequency," *IEICE Trans. Inf. & Syst.*, vol. E87-D, no. 12, pp. 2812–2820, Dec. 2004.

[24] H. Kawahara, H. Katayose, A. Cheveigné, and R. Patterson, "Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity," in *Proc. EUROSPEECH 1999*, Sept. 1999, pp. 2781–2784.

[25] Entropic Research Laboratory Inc, *ESPS Programs Version 5.0*, 1993.

[26] K. Tokuda, H. Zen, and A.W. Black, "An HMM-based speech synthesis system applied to English," in *Proc. of IEEE Speech Synthesis Workshop*, Sept. 2002.

[27] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," *J. Acoust. Soc. Japan (E)*, vol. 21, pp. 79–86, Mar. 2000.

[28] V. Digalakis and L. Neumeyer, "Speaker adaptation using combined transformation and Bayesian methods," *IEEE Trans. Speech Audio Processing*, vol. 4, pp. 294–300, July 1996.

[29] 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.

[30] O. Shiohan, T.A. Myrvoll, and C.H. Lee, "Structural maximum a posteriori linear regression for fast HMM adaptation," *Computer Speech and Language*, vol. 16, no. 3, pp. 5–24, 2002.

[31] M.J.F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, no. 2, pp. 75–98, 1998.