# The HTS-2008 System: Yet Another Evaluation of the Speaker-Adaptive HMM-based Speech Synthesis System in The 2008 Blizzard Challenge

*Junichi Yamagishi*[1], *Heiga Zen*[2], *Yi-Jian Wu*[2] *Tomoki Toda*[3], *Keiichi Tokuda*[2]

[1]The Centre for Speech Technology Research, University of Edinburgh, Edinburgh, United Kingdom

[2]Department of Computer Science, Nagoya Institute of Technology, Nagoya, Japan

[3] Graduate School of Information Science, Nara Institute of Science and Technology, Nara, Japan

jyamagis@inf.ed.ac.uk

## Abstract

For the 2008 Blizzard Challenge, we used the same speaker-adaptive approach to HMM-based speech synthesis that was used in the HTS entry to the 2007 challenge, but an improved system was built in which the multi-accented English average voice model was trained on 41 hours of speech data with high-order mel-cepstral analysis using an efficient forward-backward algorithm for the HSMM. The listener evaluation scores for the synthetic speech generated from this system was much better than in 2007: the system had the equal best naturalness on the small English data set and the equal best intelligibility on both small and large data sets for English, and had the equal best naturalness on the Mandarin data. In fact, the English system was found to be as intelligible as human speech.

**Index Terms**: speech synthesis, HMM, HTS, speaker adaptation

## 1. Introduction

Statistical parametric speech synthesis based on hidden Markov models (HMMs) [1] has become a mainstream method of speech synthesis because of its natural-sounding synthetic speech and its flexibility. It has the potential to go far beyond conventional unit-selection type methods because the speech is generated from a parametric model, which can be modified in various ways. Since HMM-based speech synthesis now has history of more than 10 years, it is worth briefly summarising the progress to date. Research on HMM-based speech synthesis started with the development of algorithms for generating smooth and natural parameter trajectories from HMMs [2]. Next, to simultaneously model the excitation parameters of speech as well as the spectral parameters, the multi-space probability distribution (MSD) HMM [3] was developed. To simultaneously model the duration for the spectral and excitation components of the model, the MSD hidden semi-Markov model (MSD-HSMM) [4] was developed. These basic systems employed a mel-cepstral vocoder with simple pulse or noise excitation, resulting in synthetic speech with a "buzzy" quality. To reduce buzziness, a more sophisticated excitation technique, called *mixed excitation* was integrated into the basic system to replace the simple pulse or noise excitation [5]. A high-quality speech vocoding method called STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum) [6] was also used, in conjunction with the mixed excitation [7]. STRAIGHT explicitly uses $F_0$ information for removing the periodic components from the estimated spectrum, i.e., it interpolates missing frequency components considering neighboring harmonic components based on an $F_0$ adaptive smoothing process on a time-frequency region. This enables the generation of better spectral parameters and consequently more natural synthetic speech. Still, all these basic systems had a serious shortcoming: the trajectories generated from the HMMs were excessively smooth due to statistical processing; over-smooth spectral parameters result in synthetic speech with a "muffled" quality which lacks the "sharpness" or "transparency" so easily achieved by concatenative methods. To alleviate this problem, a parameter generation algorithm that considers the global variance (GV) of the trajectory being generated was proposed [8]. In order to reflect within-frame correlations and optimize all the acoustic feature dimensions together, semi-tied covariance (STC) modeling [9] was employed to enable the use of full-covariance Gaussians in the HSMMs [10]. Taken together, these modest incremental improvements have an accumulative effect [7, 10, 11]. Compared with early buzzy and muffled HMM-based speech synthesis, the latest systems have a dramatically improved quality. They have exhibited good performance in the Blizzard Challenges, which are open evaluations of corpus-based text-to-speech (TTS) synthesis systems in which both HMM-based and concatenative systems from many different research groups have been compared [12–14].

The systems mentioned above are *speaker-dependent*. In parallel, we have also been developing a *speaker-adaptive* approach in which "average voice models" are created using data from several speakers. The average voice models may then be adapted using speech from a target speaker (e.g. [15]). To adapt spectral, excitation, and duration parameters within the same framework, an extended MLLR adaptation algorithm for the MSD-HSMM has recently been proposed [16]. A more robust and advanced adaptation algorithm called constrained structural maximum a posteriori linear regression (CSMAPLR) has been proposed [15]. We have also developed several techniques for training the average voice model, such as a speaker-adaptive training (SAT) algorithm [17]. To further explore the potential of HMM-based speech synthesis, for the 2007 Blizzard Challenge we combined these advances in the speaker-adaptive approach with our current speaker-dependent system that employs STRAIGHT, mixed excitation, HSMMs, GV, and full-covariance modeling [18]. However, the resulting system had two significant problems regarding the quality of the synthetic speech and the training time of the HMMs: 1) Although it was pleasing that the speaker-adaptive system provided good intelligibility without requiring manual modifications to the database, including speech and label files, the system had a lower naturalness and similarity to the original speaker than we had hoped. 2) Moreover, the training procedures for the new system were

considerably more computationally demanding than previous systems: it took about 40 days (wall-clock time) to train the HMMs on a total of 14 hours of speech training data, despite using 264 cores in of a compute cluster in parallel [19]. We analyzed the reasons for the lower-than-expected quality of the synthetic speech in detail [18]. The answers were simple:

**The order of the STRAIGHT mel-cepstral analysis**
From additional listening evaluations, we confirmed that a higher order of STRAIGHT mel-cepstral analysis can improve the similarity of synthetic speech, when the amount of speech data available is more than one hour [18]. One of the reasons the 2007 system had poor similarity scores was the use of 24-order STRAIGHT mel-cepstral coefficients. Previous HTS systems, used for the 2005 and 2006 Blizzard Challenges, utilized 39-order STRAIGHT mel-cepstral coefficients.

**The amount of training data for the average voice model**
We confirmed that there is a strong correlation between the average scores for the naturalness of synthetic speech generated from adapted models and the number of leaf nodes of the decision trees constructed for the average voice model [15]. Since the number of leaf nodes of the decision tree increases linearly with the amount of training data for the average voice model, we can also say that the naturalness of synthetic speech generated from the adapted models is closely correlated with the amount of training data for the average voice model. Using more training data is a very simple and straightforward but effective and reliable method for improving the quality of synthetic speech obtained using speaker adaptation methods. However, in the 2007 Blizzard Challenge, we used only 6 hours of training data for the average voice model, which resulted in the lower naturalness score.

These results imply that, by using an average voice model trained on a much larger amount of speech data, with a higher order of STRAIGHT mel-cepstral analysis, we can straightforwardly improve both the naturalness and the similarity of the synthetic speech. In addition, the computational cost of model training can be reduced by using an improved version of the forward-backward algorithm for hidden semi-Markov models [20].

In the 2008 Blizzard Challenge we therefore simply used the same speaker-adaptive approach used in 2007, but the model was trained on more data using a more efficient algorithm and employed a higher order cepstral analysis.

## 2. An Efficient Forward-Backward Algorithm for Hidden Semi-Markov Model

Since the original HSMM-based training algorithm was computationally expensive [20, 21] and it was necessary to build the systems within only one month during the Blizzard Challenge 2007, we had to simplify the training procedures for the average voice model used in our Blizzard Challenge 2007 entry.

Subsequently, the computational cost problem has been solved by using an efficient forward-backward algorithm for HSMMs proposed by Yu and Kobayashi [20]. The computational complexity of the efficient algorithm is $\mathcal{O}(N(D+N)T)$, where $N$ is the number of states used; $D$ is the maximum state duration; and $T$ is the number of total frames of the observations, whereas the conventional forward-backward algorithm requires $\mathcal{O}(N^2DT)$ computations [4, 22]. This makes training time for the HSMMs much shorter [23]. Therefore, we were able to use HSMM-based training algorithms, including SAT, in all stages of model training for the 2008 Blizzard Challenge. The new efficient algorithm for HSMMs has been implemented and released in HTS version 2.1 [24].

Table 1: The number of leaf nodes of constructed decision trees for each system of each English voice.

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

| System | Mel-cepstrum | $\log F_0$ | Aperiodicity | Duration |
|---|---|---|---|---|
| Benchmark | 826 | 3,666 | 906 | 391 |
| 2008 | 9,107 | 49,269 | 5,138 | 8,593 |

(b) Voice A (all the sentences: 15 hours)

| System | Mel-cepstrum | $\log F_0$ | Aperiodicity | Duration |
|---|---|---|---|---|
| Benchmark | 5,833 | 27,137 | 6,790 | 4,045 |
| 2008 | 9,380 | 57,135 | 5,559 | 8,623 |

## 3. The use of UNILEX: a Multi-Accent English Average Voice Model

As mentioned above, we found that the naturalness of synthetic speech generated from the adapted model is closely correlated with the amount of the training data for the average voice model [15]. Hence, for the 2008 Blizzard Challenge, we increased the amount of speech data for the average voice model as much as possible. We used the Unilex pronunciation lexicon from CSTR [25], which supports all accents of English in a unified way by deriving surface-form pronunciations from and underlying 'meta-lexicon' defined in terms of key symbols. The training data included speech from speakers with various English accents that differed from the target speaker's British English RP (received pronunciation) accent. Specifically, we utilized sets of general American English, Scottish English and RP English speech data and built multi-accented average voice models. Thus, the average voice model could be used as a initial model from which adaptation to any of those accents could be performed.

The amount of speech data used for the English average voice model totalled 41 hours and comprised data from 15 speakers (5 RP, 8 North American and 2 Scottish) uttering various sets of texts from various domains.

The labels for the speech data were automatically generated from word transcriptions using Festival's Multisyn module [26]. We did not modify the labels at all. A 39-order STRAIGHT mel-cepstral analysis, which is higher than in 2007, was used to improve the similarity of the synthetic speech to the speech of the original speaker.

Table 1 shows the number of leaf nodes of the constructed decision trees of each voice. The numbers for the speaker-adaptive system built for this Blizzard Challenge and the HTS Benchmark system [7] are shown together. The number of leaf nodes (which is closely related to the number of model parameters) for HTS 2008 is about twice that of the HTS benchmark system using even 15 hours of speech data. These are the biggest average voice models which we have built to date. Note that the training data for the average voice model includes the appropriate set of data for the target speaker. This is why the number of leaf nodes in the average voice model varies between

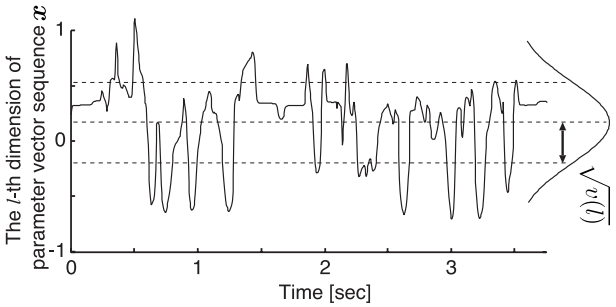


Figure 1: A GV vector contains the variance of each dimension of the parameter trajectory sequence. $v(l)$ is the $l$-th element of the GV vector $\boldsymbol{v}$.

Voice A and Voice B.

# 4. Improved GV Parameter Generation Algorithm

In the GV parameter generation algorithm proposed by Toda *et al.* [8], the objective function is manipulated by adding a penalty term to the likelihood function of the HMM $P(\boldsymbol{O} \mid \boldsymbol{q}, \lambda, T)$ as follows:

$$\log P(\boldsymbol{O} \mid \boldsymbol{q}, \lambda, T) + \omega \log \mathcal{N}(\boldsymbol{v}; \boldsymbol{\theta}, \boldsymbol{\kappa}) \tag{1}$$

where $\boldsymbol{v} \in \mathcal{R}^L$ is a GV vector containing the variance of each dimension of the parameter trajectory sequence as shown in Fig. 1. Then, $\boldsymbol{\theta} \in \mathcal{R}^L$ and $\boldsymbol{\kappa} \in \mathcal{R}^{L \times L}$ are the mean vector and covariance matrix of the GV vectors estimated from the training data. We set a weight for controlling the balance between these terms, $\omega$, to $3T$, based on the number of Gaussian distributions included in the first term. The penalty term for the GV vector is intended to keep the 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 [8].

This year we improved three aspects of this algorithm: 1) Since $\boldsymbol{v}$ is a positive vector, we used a logarithmic transform before modelling it with a Gaussian pdf; 2) We changed the GV Gaussian pdf from a single global distribution to a context-dependent one. In a similar way to HMM observation density tying, decision-tree-based clustering was applied to the context-dependent GV pdfs in order to tie their parameters. The number of leaf nodes of the decision trees was automatically determined by the MDL criterion. The number of leaf nodes for each feature is shown in Table 2, where we can see that Voice A (trained on 15 hours of speech data) has 10 to 30 GV pdfs per stream. To simplify implementation, only sentence-level contextual features (e.g. # of phonemes in a sentence) were used at this time. Thus, the objective function for the GV parameter generation used for this Blizzard Challenge can be written as:

$$\log P(\boldsymbol{O} \mid \boldsymbol{q}, \lambda, T) + \omega \log \mathcal{N}_s(\log \boldsymbol{v}; \boldsymbol{\theta}, \boldsymbol{\kappa}) \tag{2}$$

Finally, 3) we calculated the GV vector $\boldsymbol{v}$ only from speech and excluded silence and pause regions from the calculation, based on automatic segmentation, in order to improve the estimation accuracy of the GV vector. This improved GV algorithm has also been implemented and released in HTS version 2.1.

In this parameter generation process, gradient methods are employed for iteratively updating the generated parameter trajectories. An increase of the objective function is often used as a stopping criterion of the iterative update. When setting the stopping criterion to a small value, the generated trajectories with the larger objective function are caused because the number of iterative updates increases. However, we found that a chance to cause unstable sounds in synthetic speech also tends to increase. One possibility causing this problem would be a fact that HMM and GV pdfs are trained independently. In this challenge, we alleviate this problem by carefully adjusting the stopping criterion.

Table 2: The number of leaf nodes of constructed decision trees for the context-dependent GV pdfs of each English voice.

| System | Mel-cepstrum | $\log F_0$ | Aperiodicity |
|---|---|---|---|
| Voice B | 2 | 3 | 3 |
| Voice A | 12 | 33 | 14 |

# 5. The Blizzard Challenge 2008

The Blizzard Challenge is an annual evaluation of corpus-based speech synthesis systems, in which each participating team builds a synthetic voice from common training data, then synthesizes a set of test sentences. Listening tests are adopted to evaluate the systems in term of naturalness, similarity to original speaker and intelligibility. The Blizzard Challenge 2005 used the CMU-ARCTIC speech databases; in 2006, a database consisting of five hours of speech uttered by a male speaker was released by ATR from their ATRECSS corpus. In the Blizzard Challenge 2007, an extended version of the 2006 corpus was released by ATR, containing eight hours of speech data uttered by the same male speaker. In the Blizzard Challenge 2008, an English speech database consisting of 15 hours of speech uttered by a British male speaker and a Mandarin speech database consisting of about 6 hours of speech uttered by a Beijing female speaker were released by the Centre for Speech Technology Research (CSTR), University of Edinburgh, UK, and the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, respectively.

## 5.1. Experimental Conditions for English Systems

We used 41 hours of speech including the released data as the training data for the English average voice model for the full data set. The labels for the data were automatically generated using Unilex [25] and Festival's Multisyn module, with no further modification. The English phonetic, linguistic and prosodic context factors used were similar to those in [27]. To investigate the effect of the corpus size, two systems were submitted by almost all participants: one built using all the speech data included in the released database (Voice A), and a second built using only the ARCTIC subset (Voice B). Note that, when building Voice B, we did not utilise the remaining data in the full data set either to train acoustic models used for segmentation or to train the average voice model. Thus only 27 hours of speech were used as the training data for the English average voice model for Voice B. For Voice B, we adapted the average voice model to the British English target speaker using only the speech data specified for Voice B.

### 5.2. Experimental Conditions for Mandarin Systems

We used a Mandarin speech database consisting of six hours of speech data uttered by six speakers, which was kindly provided by iFlytek, as the training data for the Mandarin average voice model. The labels for the data were automatically generated using the iFlytek text-processing front-end modules. Mandarin phonetic, linguistic and prosodic contexts used were the same as those in [28]. We did not manually modify these labels. We adapted the trained average voice model to the target speaker using all the released speech data.

### 5.3. Listening Tests

English synthetic speech was generated for a set of 600 test sentences, including 400 sentences from conversational, news and novel genres (used to evaluate naturalness and similarity) and 200 semantically unpredictable sentences (used to evaluate intelligibility). Mandarin synthetic speech was generated for a set of 697 test sentences, including 647 sentences from a news genre (used to evaluate naturalness and similarity) and 50 semantically unpredictable sentences (used to evaluate intelligibility). To evaluate naturalness and similarity, 5-point mean opinion score (MOS) and comparison category rating (CCR) tests were conducted. The scale for the MOS test was 5 for "completely natural" and 1 for "completely unnatural". The scale for the CCR tests was 5 for "sounds like exactly the same person" and 1 for "sounds like a totally different person" compared to a few natural example sentences from the reference speaker. To evaluate intelligibility, the subjects were asked to transcribe semantically unpredictable sentences; in English tests average word error rates (WER) were calculated from these transcripts: In Mandarin tests average character error rate and average pinyin and tone error rate were calculated. The evaluations were conducted over a six week period via the internet.

### 5.4. Experimental Results of the English Systems

Figures 2–4 show the evaluation results for English Voice A (15 hours) and Voice B (1 hour) of the 20 participating systems. One participant did not submit speech for either English voice and one further participant did not submit synthetic speech for Voice B. In these figures, systems "V" corresponds to the HTS-2008 system. "A", "B" and "C" correspond to real speech, the Festival "Multisyn" benchmark speech synthesis system [29] and the HTS benchmark system [7], respectively. The Festival system uses a conventional unit-selection method. The HTS Benchmark system is a standard statistical parametric system using speaker-dependent HMMs, which can be trained from scratch by using HTS toolkit version 2.1 and STRAIGHT. This system was highly rated in terms of naturalness and intelligibility in the 2005 Blizzard Challenge. One of main differences between the HTS benchmark system and the speaker-adaptive HTS-2008 system is the use of the average voice models. We can see several interesting findings in the results:

**Naturalness (Figure 2)**

Our HTS-2008 system, "J", and system "S" are equal best on the smaller dataset. Even on the larger dataset, the HTS-2008 system is above average: it is statistically worse than only three systems "J", "K", and "S" ($p < 0.01$). There is no significant difference between the Festival benchmark system ("B"), "O", "P" and the HTS-2008 system. There are significant differences between real speech and all systems.

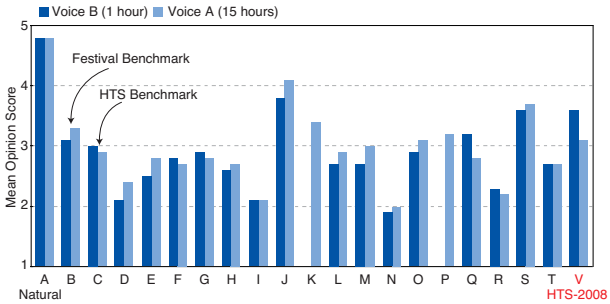


Figure 2: Mean opinion scores of all U.K. English systems.

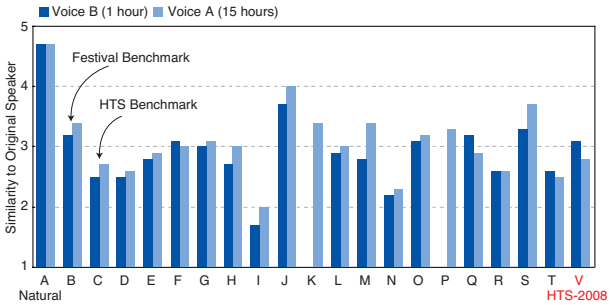


Figure 3: Similarity to original speaker of all U.K. English systems.

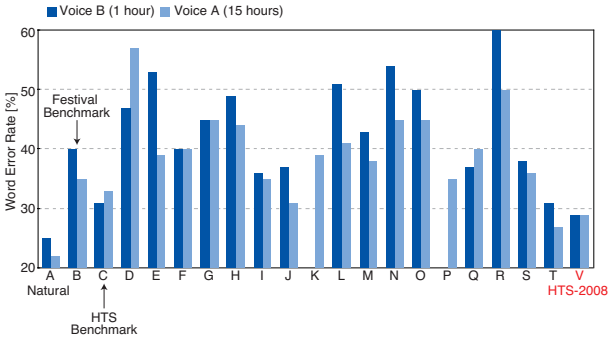


Figure 4: Average word error rate (%) of all U.K. English systems.

**Intelligibility – word error rate, (Figure 4)**

Our HTS-2008 system and "T" are equal best on the smaller dataset. These systems are as intelligible as human speech on smaller data, i.e., there is no significant difference in the word error rate for real speech and the word error rates of these two systems. On the larger dataset, the HTS-2008 system, the HTS benchmark system and "T" are equal best. Likewise, these systems are as intelligible as human speech. 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. In fact 2% of the words in the test set are out-of-vocabulary; it is likely that intelligibility would be further improved if these were added to the lexicon.

**Similarity (Figure 3)**

On the smaller dataset, the HTS-2008 system is above average: it is statistically worse than only one system: "J" ($p$<0.01). All systems are worse than natural speech. However on the larger dataset, the HTS-2008 system is

only average: it is statistically worse than 8 systems "B", "G", "J", "K", "M", "O", "P", and "S". This is one of major shortcomings of current HMM-based speech synthesis.

**Comparison with the HTS benchmark system**
On the smaller dataset, the HTS-2008 system is significantly better than the HTS benchmark system in terms of similarity and naturalness. On the larger dataset, there are no significant differences between them. The HTS-2008 system is speaker-adaptive and we can say that these results are very good: the system has adapted to the target speaker characteristics to the point where it is good as a speaker-dependent system trained on a large amount of target-speaker data.

**Comparison with previous results in the 2007 Challenge**
Although the basic concept of the HTS-2008 system was the same as last year, we obtained much better results this year (see evaluation results for "Voice B" using the ARCTIC dataset in [19]). As mentioned earlier, the system for this year was simply built under better conditions. This result is consistent with the results for speaker-adaptive HMM-based speech synthesis systems reported in [15, 18].

### 5.5. Experimental Results of the Mandarin System

Figures 5–7 show the evaluation results of the Mandarin submissions for 13 participants. Several participants submitted only English synthetic speech and one participant submitted only Mandarin synthetic speech. In these figures, as for English, "A", "C" and "V" correspond to real speech, the HTS benchmark system, and the HTS-2008 system, respectively. There is no Festival benchmark system for Mandarin.

**Naturalness (Figure 5)**
Our HTS-2008 system, the HTS benchmark system, "F", "S", "T", and "U" are equal best. However, there are significant differences between real speech and all systems.

**Intelligibility – pinyin + tone error rate (Figure 7)**
There is no significant difference between the HTS benchmark system and systems "T" or "U", although HTS-2008 is significantly different from natural speech whereas "T" and "U" are not significantly different from natural speech. We expect that the use of larger average voice models would improve intelligibility.

**Similarity (Figure 6)**
The HTS-2008 system is once again below average.

**Comparison with the HTS benchmark system**
The HTS-2008 system is significantly worse than the HTS benchmark system in the terms of similarity ($p < 0.01$).

For the Mandarin voices, we could not collect enough speech data for the training of the average voice model. (In addition, we had to omit some training and adaptation procedures because of the differences of computer resources available in each institute.) Both of these factors are thought to have lowered the performance of HTS-2008 for Mandarin. Table 3 shows the number of leaf nodes of the constructed decision trees of the Mandarin systems, where we can see that the trained average voice model was as small as the speaker-dependent model used in the HTS benchmark system. This is not an ideal situation for the speaker-adaptive approach, where we would normally want to use a large average voice model as the basis for adaptation.

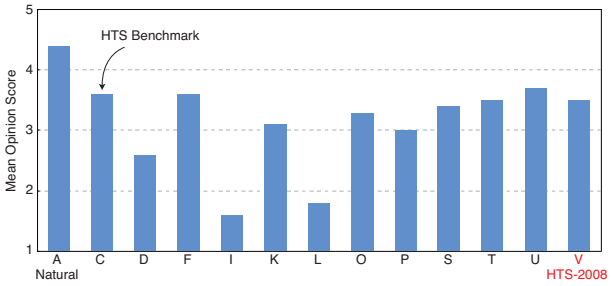


Figure 5: Mean opinion scores of all Mandarin systems.

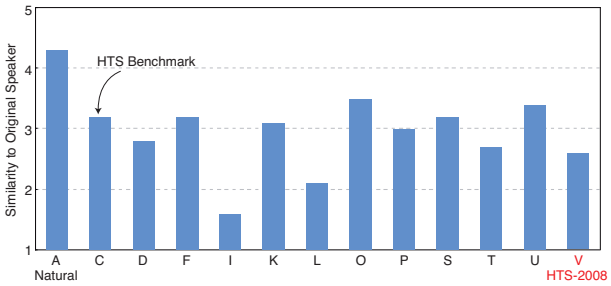


Figure 6: Similarity to original speaker of all Mandarin systems.

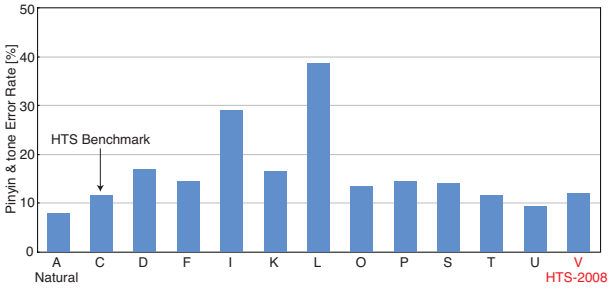


Figure 7: Average pinyin and tone error rate (%) of all Mandarin systems.

## 6. Conclusions

In the 2008 Blizzard Challenge, we tried the same speaker-adaptive approach that we used in 2007, but built the systems under better conditions. Multi-accented English average voice models were trained on 41 hours of speech data using the efficient forward-backward algorithm for HSMMs. The listeners' evaluation scores were much better than those of HTS-2007. For English, HTS-2008 achieved the best naturalness on the smaller data set and the best intelligibility on both data sets. In addition, the two English systems were found to be as intelligible as human speech. Although the training condition for the Mandarin system was far from ideal, the system was found to sound as natural as, or more natural than, all other systems. However, as expected, the imperfect training conditions for the Mandarin systems produced some negative results. The synthetic speech from HTS-2008 submitted for the Blizzard Challenge 2008 can be downloaded from `http://homepages.inf.ed.ac.uk/jyamagis/blizzard08`.

## 7. Acknowledgements

The authors would like to thank ANHUI USTC iFLYTEK Co., Ltd. for permission to use their Mandarin speech database and

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

| System | Mel-cepstrum | $\log F_0$ | Aperiodicity | Duration |
|---|---|---|---|---|
| Benchmark | 3,104 | 15,549 | 4,277 | 1,061 |
| 2008 | 3,035 | 20,164 | 3,089 | 1,788 |

thank Dr. Simon King for his useful review and comments. 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). This work has made use of the resources provided by the Edinburgh Compute and Data Facility (ECDF). (`http://www.ecdf.ed.ac.uk/`). The ECDF is partially supported by the eDIKT initiative. (`http://www.edikt.org`.)

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

[2] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorigthms for HMM-based speech synthesis," in *Proc. ICASSP 2000*, Jun. 2000, pp. 1315–1318.

[3] 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.

[4] 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.

[5] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Mixed excitation for HMM-based speech synthesis," in *Proc. EUROSPEECH 2001*, Sep. 2001, p. 22632266.

[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, 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.

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

[10] H. Zen, T. Toda, and K. Tokuda, "The Nitech-NAIST HMM-based speech synthesis system for the Blizzard Challenge 2006," *IEICE Trans. Inf. & Syst.*, vol. E91-D, no. 6, pp. 1764–1773, Jun. 2008.

[11] Z.-H. Ling, Y.-J. Wu, Y.-P. Wang, L. Qin, and R.-H. Wang, "USTC system for Blizzard Challenge 2006 an improved hmm-based speech synthesis method," in *Proc. Blizzard Challenge 2006*, Sep. 2006.

[12] A. Black and K. Tokuda, "The Blizzard Challenge – 2005: Evaluating corpus-based speech synthesis on common datasets," in *Proc. EUROSPEECH 2005*, Sep. 2005, pp. 77–80.

[13] C. Bennett and A. Black, "The blizzard challenge 2006," in *Proc. Blizzard Challenge 2006*, Sep. 2006.

[14] M. Fraser and S. King, "The Blizzard Challenge 2007," in *Proc. BLZ3-2007 (in Proc. SSW6)*, Aug. 2007.

[15] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Trans. Speech, Audio & Language Process.*, 2008, (accept for publication).

[16] 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.

[17] 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.

[18] J. Yamagishi, T. Nose, H. Zen, Z.-H. Ling, T. Toda, K. Tokuda, S. King, and S. Renals, "A speaker-adaptive HMM-based speech synthesis for the Blizzard Challenge 2007," *IEEE Trans. Speech, Audio & Language Process.*, 2008, (under review).

[19] 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.

[20] S.-Z. Yu and H. Kobayashi, "An efficient forward-backward algorithm for an explicit-duration hidden Markov model," *IEEE Signal Processing Letters*, vol. 10, no. 1, pp. 11–14, Jan. 2003.

[21] C. Mitchell, M. Harper, and L. Jamieson, "On the complexity of explicit duration HMM's," *IEEE Trans. Speech Audio Processing*, vol. 3, no. 3, pp. 213–217, May 1995.

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

[23] H. Zen, "Implementing an HSMM-based speech synthesis system using an efficient forward-backward algorithm," in *Technical Report of Nagoya Institute of Technology, TR-SP-0001*, Dec. 2007.

[24] K. Tokuda, H. Zen, J. Yamagishi, A. Black, T. Masuko, S. Sako, T. Toda, T. Nose, and K. Oura, *The HMM-based speech synthesis system (HTS) Version 2.1*, http://hts.sp.nitech.ac.jp/.

[25] S. Fitt and S. Isard, "Synthesis of regional English using a keyword lexicon," in *Proc. Eurospeech 1999*, vol. 2, Sep. 1999, pp. 823–826.

[26] R. A. J. Clark, K. Richmond, and S. King, "Multisyn: Open-domain unit selection for the Festival speech synthesis system," *Speech Communication*, vol. 49, no. 4, pp. 317–330, 2007.

[27] H. Zen, K. Tokuda, and T. Kitamura, "An introduction of trajectory model into HMM-based speech synthesis," in *Proc. ISCA SSW5*, 2004.

[28] Y.-J. Wu, "Research on HMM-based speech synthesis," in *Ph.D. Thesis, University of Science and Technology of China*, 2006.

[29] 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.