PAPER

Details of the Nitech HMM-Based Speech Synthesis System for the Blizzard Challenge 2005

Heiga ZEN[†a)], *Nonmember*, **Tomoki TODA**[††b)], *Member*, **Masaru NAKAMURA**[†c)], *Nonmember*, *and* **Keiichi TOKUDA**[†d)], *Member*

SUMMARY In January 2005, an open evaluation of corpus-based text-to-speech synthesis systems using common speech datasets, named *Blizzard Challenge 2005*, was conducted. Nitech group participated in this challenge, entering an HMM-based speech synthesis system called Nitech-HTS 2005. This paper describes the technical details, building processes, and performance of our system. We first give an overview of the basic HMM-based speech synthesis system, and then describe new features integrated into Nitech-HTS 2005 such as STRAIGHT-based vocoding, HSMM-based acoustic modeling, and a speech parameter generation algorithm considering GV. Constructed Nitech-HTS 2005 voices can generate speech waveforms at 0.3 ×RT (real-time ratio) on a 1.6 GHz Pentium 4 machine, and footprints of these voices are less than 2 Mbytes. Subjective listening tests showed that the naturalness and intelligibility of the Nitech-HTS 2005 voices were much better than expected.

key words: *HMM-based speech synthesis, Blizzard Challenge 2005, STRAIGHT, HSMM, GV*

1. Introduction

The increasing availability of large speech databases makes it possible to construct data-driven speech synthesis systems, referred to as *corpus-based speech synthesis systems* [1], by applying statistical learning algorithms. These systems can both synthesize natural and high quality speech and reproduce the original speaker's voice characteristics.

In recent years, a kind of corpus-based speech synthesis system based on Hidden Markov Models (HMMs) has been developed [2]. In the training part of this system, spectral and excitation parameters are extracted from a speech database and modeled by context-dependent HMMs. In the synthesis part, spectral and excitation parameters are generated from the HMMs themselves [3]. By filtering the generated excitation, a speech synthesis filter controlled by the generated spectral parameters synthesizes a speech waveform. This system has the following features:

1) Smooth and natural sounding speech can be synthesized,
2) voice characteristics can be easily modified,
3) it is *trainable*.

The speech synthesis in 1) can be carried out by taking account of the statistics for both static and dynamic features, which constrains the dynamics of generated speech parameter vector to be realistic. The voice characteristics in 2) can be changed by transforming HMM parameters appropriately because the system generates speech waveforms from the HMMs themselves. In fact, speaker adaptation [4], speaker interpolation [5], and eigenvoice [6] techniques have been applied to this system to modify its voice characteristics. As for 3), this system can be automatically constructed.

In January 2005, Black and Tokuda conducted an open evaluation of corpus-based text-to-speech synthesis systems using common speech datasets, named *Blizzard Challenge 2005* [7], [8]. Nitech group participated in this challenge, entering an HMM-based speech synthesis system called Nitech-HTS 2005. In this paper, we describe technical details, building processes, and performance of the Nitech-HTS 2005 voices.

One of the major limitations of the basic HMM-based speech synthesis system is that synthesized speech is *buzzy* since it uses a simple mel-cepstral vocoder. To solve this problem, Nitech-HTS 2005 uses Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum (STRAIGHT) [9]. Other technologies such as Hidden Semi-Markov Model (HSMM) [10] and the speech parameter generation algorithm considering Global Variance (GV) [11] were also integrated.

The rest of this paper is organized as follows. First, a brief overview of the Blizzard Challenge 2005 is given. Then, the basic HMM-based speech synthesis system is reviewed. After that, the new features integrated into Nitech-HTS 2005 and the details of constructed voices submitted to the Blizzard Challenge 2005 are discussed. Finally concluding remarks are presented.

2. Blizzard Challenge 2005

In January 2005, Black and Tokuda conducted the Blizzard Challenge 2005 [7] to more closely compare the labeling, pruning, target and concatenation costs, signal processing, and others techniques of corpus-based text-to-speech synthesis systems. Organizers asked participants to use the designated databases to synthesize utterances from a small

Manuscript received July 4, 2006.
Manuscript revised October 3, 2006.

[†]The authors are with the Department of Computer Science and Engineering, Nagoya Institute of Technology, Nagoya-shi, 466–8555 Japan.
[††]The author is with the Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma-shi, 630–0192 Japan.
a) E-mail: zen@ics.nitech.ac.jp
b) E-mail: tomoki@is.naist.jp
c) E-mail: masha@ics.nitech.ac.jp
d) E-mail: tokuda@ics.nitech.ac.jp

DOI: 10.1093/ietisy/e90–d.1.325

number of genres. They also did an organized evaluation based on subjective listening tests to try to rank the systems and help identify the effectiveness of the techniques used by each [12].

In 2004, the first two single-speaker datasets (US-English) consisting of 1132 phonetically balanced utterances were released as the CMU ARCTIC databases [13]. Then, groups working in speech synthesis around the world were asked to build their best voices from these databases. In January 2005, the two additional databases (US-English) and a set of 50 utterance texts from each of five genres were released. The participants were asked to build four voices using these datasets and synthesize these utterances in a week. Their resulting synthesized utterances were then presented to three groups of listeners: speech experts, volunteers recruited from the Web, and native US-English speaking undergraduates. The evaluation included five separate tests, one from each genre. Tests 1 through 3 were Mean Opinion Score (MOS) tests, where the listener made a judgment about the naturalness of a particular sample by assigning it a score of 1 to 5. The remaining two tests were type-in tests that required the listener to enter the words they heard into a textbox. Details of the Blizzard Challenge 2005 itself are described by Tokuda and Black [7]. Discussions of its results and evaluation methodology are available in Bennett [12]. In the following section, technical details, building processes, and performance of the Nitech-HTS 2005 voices are described.

3. Basic System

Figure 1 is an overview of the basic HMM-based speech synthesis system [2]. In this system, feature vectors for training HMMs consist of spectrum and excitation parameter vectors: the spectrum parameter vectors are composed of mel-cepstral coefficients [14], their delta, and delta-delta, and excitation parameter vectors consist of logarithmic fundamental frequency ($\log F_0$) values, their delta, and delta-delta.

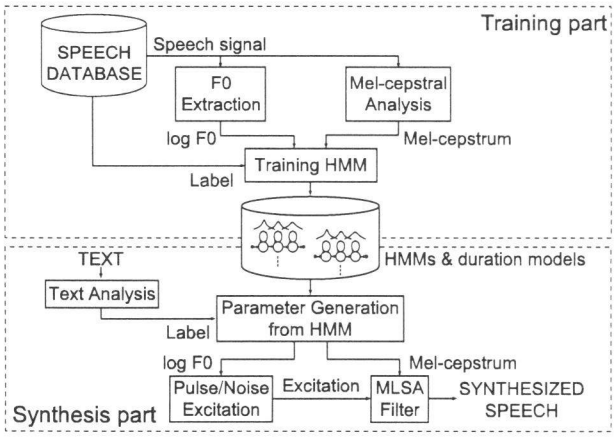


Fig. 1 The overview of the basic HMM-based speech synthesis system.

In the training part, sequences of feature vectors are modeled by context-dependent HMMs. Although sequences of mel-cepstral coefficients can be modeled by continuous HMMs, $\log F_0$ sequences cannot be represented by continuous or discrete HMMs since each observation is composed of a one-dimensional continuous $\log F_0$ value or a discrete symbol, which represents *unvoiced* frame. To model this kind of observation, HMMs based on multi-space probability distributions (MSD-HMMs) have been proposed [15]. An MSD-HMM includes both discrete and continuous HMMs as its special cases and can model $\log F_0$ sequences with no heuristic assumptions. The training procedure of the context-dependent HMMs is very similar to that used in speech recognition. The main differences are as follows:

- It estimates model parameters based on the maximum likelihood criterion rather than the discriminative one [16];†
- it uses a Gaussian distribution rather than mixture of Gaussian distributions for each state output probability density function;††
- it takes into account linguistic contexts as well as phonetic contexts;
- it also models state duration probability density functions by multi-variate Gaussian distributions.†††

In the synthesis part, a text to be synthesized is first converted to a context-dependent label sequence and then the sentence HMM is constructed based on the label sequence. Second, the state durations are determined so as to maximize their probabilities based on the state duration probability density functions. Third, the speech parameter generation algorithm (typically, case 1 as described by Tokuda et al. [3] is used) generates the sequences of mel-cepstral coefficients and $\log F_0$ values that maximize their output probabilities. Finally, a speech waveform is synthesized directly from the generated mel-cepstral coefficients and $\log F_0$ values using the Mel Log Spectrum Approximation (MLSA) filter [20] with binary pulse or noise excitation.

4. New Features Integrated into Nitech-HTS 2005

One of the major limitations of the basic HMM-based speech synthesis system is that synthesized speech is *buzzy* since it uses a mel-cepstral vocoder with binary pulse or noise excitation. To solve this problem, several attempts to incorporate advanced excitation models into the HMM-based speech synthesis have been reported [21], [22]. In Nitech-HTS 2005, a high-quality vocoder called STRAIGHT [9] was introduced to address this problem. The system also integrated other technologies such as

†A parameter optimization criteria so as to minimize mean squared error between generated speech parameter trajectories and training data has also been proposed [17].

††The use of mixture of Gaussian distributions increases computational complexity in the speech parameter generation [18].

†††Gamma distributions have also been applied [19].

HSMM-based acoustic modeling [10] and a speech parameter generation algorithm considering GV [11], (which improved the basic system). Figure 2 is a block diagram of Nitech-HTS 2005. Differences can be seen between the basic system and Nitech-HTS 2005 in the front-end, acoustic modeling, and speech parameter generation parts. The following section details and describes the evaluations of these new features.

4.1 STRAIGHT Vocoding

Nitech-HTS 2005 uses the high-quality speech vocoding method STRAIGHT. This is a vocoder type algorithm proposed by Kawahara et al. [9] and is diagrammed in Fig. 3. It consists of three main components: F_0 extraction, spectral and aperiodicity measurement analysis, and speech synthesis.

First, STRAIGHT automatically extracts F_0 values with fixed-point analysis [23]. Using the extracted F_0 values, STRAIGHT carries out F_0-adaptive spectral analysis combined with a surface reconstruction method in the time-frequency region to remove signal periodicity. It also extracts aperiodicity measurements on the frequency domain. These are based on a ratio between the lower and upper smoothed spectral envelopes and represent the relative energy distribution of aperiodic components [24]. In the synthesis part, a mixed excitation is designed as a weighted sum of a pulse train with phase manipulation and Gaussian noise. The weighting process is carried out in the frequency domain using the aperiodicity measurements. Using the smoothed spectrum and mixed excitation, STRAIGHT synthesizes a speech waveform with FFT-based processing.

To construct its voice, Nitech-HTS 2005 used a two-stage extraction to alleviate errors of F_0 extraction such as halving and doubling. First, the system extracted F_0 values for all training data for each speaker within a search range from 40 to 600 Hz. Then, the F_0 range of each speaker was roughly estimated taking account of a histogram of the extracted F_0 values. Then, F_0 values were re-extracted in the speaker-specific range. In STRAIGHT, a smoothed spectrum is used as a spectral parameter. However, it is high-dimensional representation (e.g., 512 dimensions). Estimating statistically reliable acoustic models using such high-dimensional observations is very difficult. To avoid this problem, in Nitech-HTS 2005 mel-cepstral coefficients, converted from the smoothed spectrum with a recursive algorithm [25], hereafter referred to as *STRAIGHT mel-cepstrum*, were used as its spectral parameters. For the same reason, the aperiodicity measurements were averaged on five frequency sub-bands, i.e., 0–1, 1–2, 2–4, 4–6, and 6–8 kHz. Figure 4 shows examples of spectral envelopes extracted with the FFT and represented by conventional and STRAIGHT mel-cepstrum. It can be seen from the figure that the conventional mel-cepstral analysis suffers from the F_0 effect in the lower frequency range when the order of mel-cepstral analysis is high. In contrast, the STRAIGHT mel-cepstrum avoids this effect and approximates the spectral envelope well. To synthesize a speech waveform, it is necessary to convert the mel-cepstrum to the linear-scaled spectrum, since STRAIGHT uses FFT-based processing to synthesize speech waveforms. However, this increases computational complexity. To reduce computational cost, we used the MLSA filter [20] in Nitech-HTS 2005 in a process illustrated in Fig. 5. Specifically, a one-pitch waveform was synthesized from mel-cepstral coefficients and the mixed-

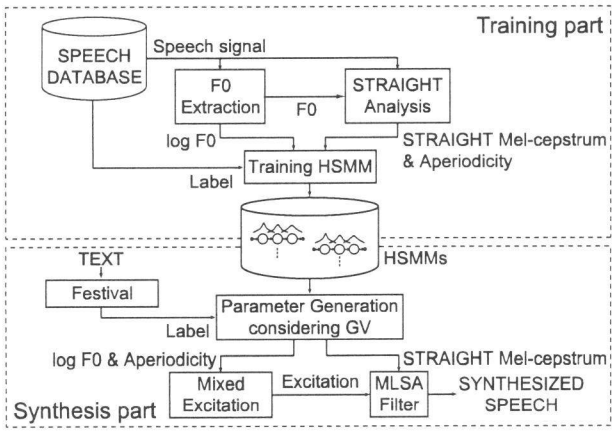


Fig. 2 The overview of the Nitech-HTS 2005.

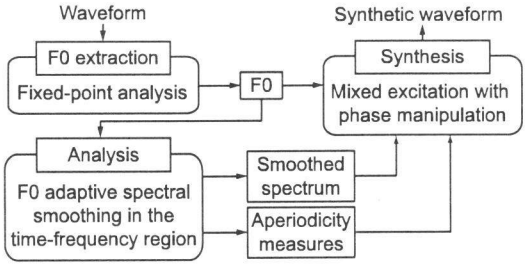


Fig. 3 A block diagram of STRAIGHT vocoding method.

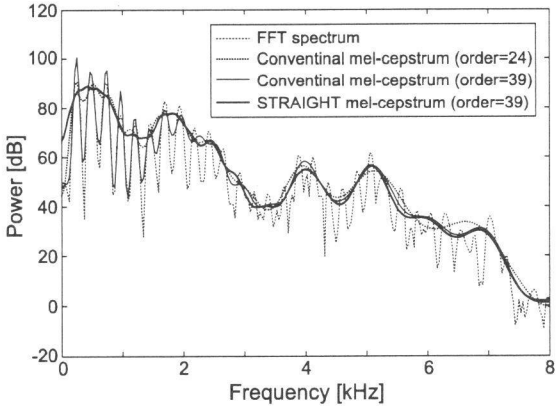


Fig. 4 Examples of spectral extracted with FFT and represented by mel-cepstral coefficients extracted with the mel-cepstral analysis (order of mel-cepstral analysis is 24 and 39) and converted from the smoothed spectrum (order of mel-cepstral analysis is 39).

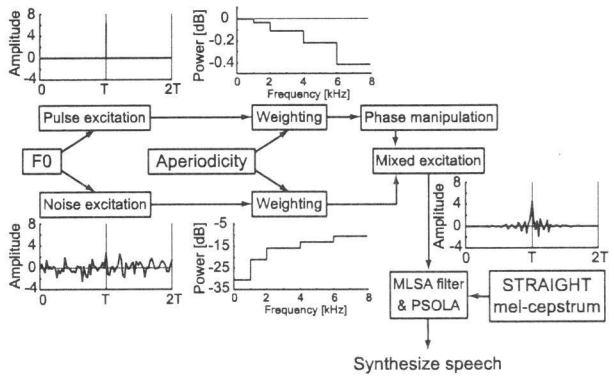


Fig. 5 A block diagram of the synthesis part of STRAIGHT.

excitation with the MLSA filter, and then a synthesized waveform was generated with Pitch-Synchronous OverLap-Add (PSOLA) [26].

4.2 HSMM-Based Acoustic Modeling

In the HMM-based speech synthesis system, rhythm and tempo are controlled by the state duration probabilities modeled by single Gaussian distributions. They are estimated from statistical variables obtained in the last iteration of the forward-backward algorithm, and then clustered by a decision tree-based context-clustering algorithm: they are not re-estimated in the Baum-Welch iteration. In the synthesis stage, we construct a sentence HMM and determine its state durations so as to maximize their probabilities. Then, a speech parameter vector sequence is generated. However, there is an inconsistency in the basic system: although parameters of HMMs are re-estimated without explicit state duration probability density functions, speech parameter vector sequences are generated from the HMMs using the explicit state duration probability density functions. This inconsistency can degrade the quality of synthesized speech.

To resolve the discrepancy, HSMMs [27], which can be viewed as HMMs with explicit state duration probability density functions, were introduced into the training part of the system [10]. The use of HSMMs makes it possible to simultaneously re-estimate state output and duration probability density functions. Improvements in durations, spectrum, and F_0 have been reported [10].

4.3 Speech Parameter Generation Algorithm Considering GV

In the basic system, the speech parameter generation algorithm (typically case 1 described by Tokuda et al. [18]) is used to generate spectral and excitation parameters from the HMMs. By taking account of constraints between the static and dynamic features, it can generate smooth speech parameter trajectories. However, the generated spectral and excitation parameters are often over-smoothed. Synthesized speech using over-smoothed spectral parameters sounds muffled. There were several attempts to reduce this

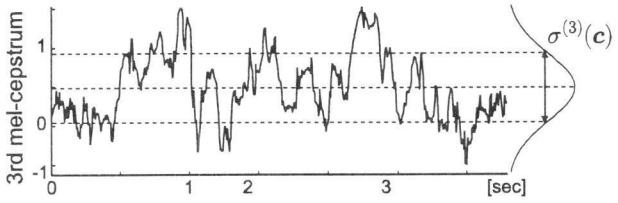


Fig. 6 Examples of the third mel-cepstrum's trajectory $\boldsymbol{c}$ for an utterance and its global variance $\sigma^{(3)}(\boldsymbol{c})$ (the left side of the figure plots the trajectory and the right side shows a Gaussian distribution calculated from $\boldsymbol{c}$).

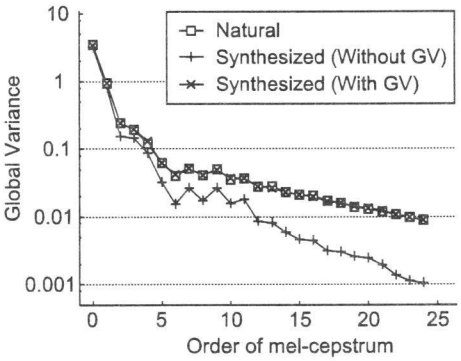


Fig. 7 Averages for GVs of STRAIGHT mel-cepstral coefficients of natural speech and synthesized speech generated by the speech parameter generation algorithm with and without considering GV. They are calculated from 212 utterences by two female and two male speakers (53 utterances for each speaker).

effect [28]. In Nitech-HTS 2005, a speech parameter generation algorithm considering GV [11] was used.

Figure 6 shows how GV, which is defined as an intra-utterance variance, is calculated. The probability density function of GV is modeled by a Gaussian distribution with a diagonal covariance matrix. To generate speech parameters, a trajectory is first generated with the generation algorithm (case 1 in Tokuda et al. [18]). Then, the generated trajectory is converted so that its GV is equal to the mean of the Gaussian distribution. Using the converted trajectory as the initial value, the speech parameter trajectory that maximizes an objective function (defined as a weighted sum of logarithmic output probability of the speech parameter sequence and that of its GV) is iteratively optimized using the Newton-Raphson method. Figure 7 shows plots of averages for GVs of natural and synthesized speech. It can be seen that by considering GV in the speech parameter generation process, the average for GVs of synthesized speech almost reaches the level of natural speech.

The performance of GV in speech synthesis is highly dependent on spectral parameter representations. Figure 8 shows an example of Gaussian distributions for GVs of mel-cepstral coefficients extracted with conventional mel-cepstral analysis and STRAIGHT mel-cepstral analysis. It can be seen that the Gaussian distribution modeling GVs of conventional mel-cepstral coefficients have relatively large means and wide variances in higher orders because of the

Table 1 Conditions of constructed systems for evaluating the effectiveness of STRAIGHT, HSMM and GV.

System setting	Spectrum	Order of mel-cepstrum	Excitation	Model	GV
A	Mel-cepstrum + Δ + Δ^2	24	log F_0 + Δ + Δ^2	HMM	no
B	STRAIGHT mel-cepstrum + Δ + Δ^2	24	log F_0 + Δ + Δ^2, aperiodicity + Δ + Δ^2	HMM	no
C	Mel-cepstrum + Δ + Δ^2	24	log F_0 + Δ + Δ^2	HSMM	no
D	STRAIGHT mel-cepstrum + Δ + Δ^2	24	log F_0 + Δ + Δ^2, aperiodicity + Δ + Δ^2	HSMM	no
E	STRAIGHT mel-cepstrum + Δ + Δ^2	39	log F_0 + Δ + Δ^2, aperiodicity + Δ + Δ^2	HSMM	no
F	Mel-cepstrum + Δ + Δ^2	24	log F_0 + Δ + Δ^2	HSMM	yes
G	STRAIGHT mel-cepstrum + Δ + Δ^2	24	log F_0 + Δ + Δ^2, aperiodicity + Δ + Δ^2	HSMM	yes
H	STRAIGHT mel-cepstrum + Δ + Δ^2	39	log F_0 + Δ + Δ^2, aperiodicity + Δ + Δ^2	HSMM	yes

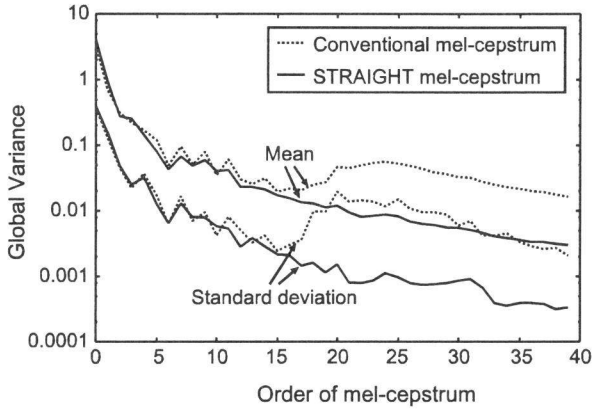


Fig. 8 Gaussian distributions for GVs of mel-cepstral coefficients extracted with conventional mel-cepstral analysis and STRAIGHT mel-cepstral analysis. Means and standard deviations are calculated from 50 sentences uttered by a female speaker whose average F_0 is 235.2 Hz, which are recorded at 16 kHz sampling.

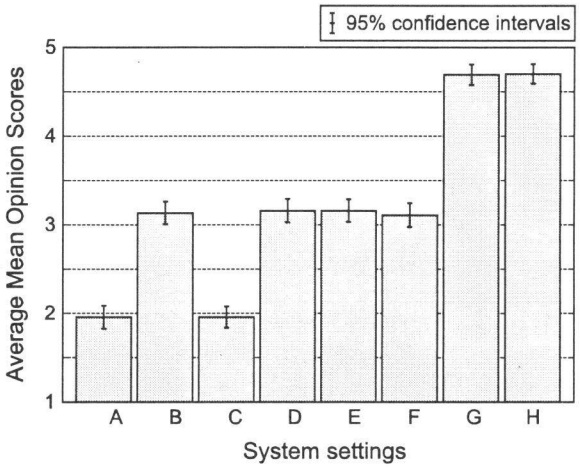


Fig. 9 Results of subjective evaluation (speaker MYI).

F_0 effect shown in Fig. 4. This causes the frequency components around F_0 values in impulse responses of generated mel-cepstral coefficients to be overemphasized. By contrast, as discussed in Sect. 4.1 STRAIGHT mel-cepstrum is not subject to this effect. Therefore, we can use higher-order mel-cepstral analysis and synthesize natural sounding speech waveforms. This effect is more apparent in female speakers who have high average F_0 values.

The algorithm described in this section was applied in Nitech-HTS 2005 to both spectral and F_0 generation processes.[†] Note that only voiced frames were used to calculate GV of F_0 values.

4.4 Evaluation of the New Features

To evaluate the effectiveness of the three new features, we conducted a subjective listening test. The first 450 of 503 phonetically balanced sentences from the ATR Japanese speech database B-set uttered by speakers FYM (female) and MYI (male) were used to train the system. The remaining 53 sentences were used for evaluation. Eight systems (A–H) were constructed to evaluate each of the new features. The specifications of these systems are shown in Table 1. In this experiment, we also evaluated the effect of increasing the order of mel-cepstral analysis. The setting for system H was adopted in Nitech-HTS 2005. Speech signals were sampled at a rate of 16 kHz and windowed by a 25-ms Blackman window in the mel-cepstral analysis used in the baseline system with a 5-ms shift or a Gaussian window with F_0 adaptive window size for STRAIGHT mel-cepstral analysis with a 5-ms shift. Then spectral and excitation parameters were extracted. A five-state left-to-right HMM/HSMM structure with no skip was used.

The naturalness of synthesized speech was evaluated by Mean Opinion Score (MOS). For each test sentence, eight samples[††] were presented in random order. After listening to each sample, subjects were asked to assign a 5-point score (5: natural–1: poor) to each sample. The 12 subjects were all students. For each subject, 10 sentences were randomly chosen from 53 test sentences. Figures 9 and 10 show experimental results. It can be seen from the figures that introducing STRAIGHT and the speech parameter generation considering GV dramatically enhanced the quality of synthesized speech. Using the HSMM and increasing the order of mel-cepstral analysis, on the other hand, did not significantly improve the quality. For both speakers, the H

[†]This algorithm was not applied to aperiodicity measures generation because its effect was relatively small and it increased the computational complexity.

[††]Natural speech examples were not included in this experiment.

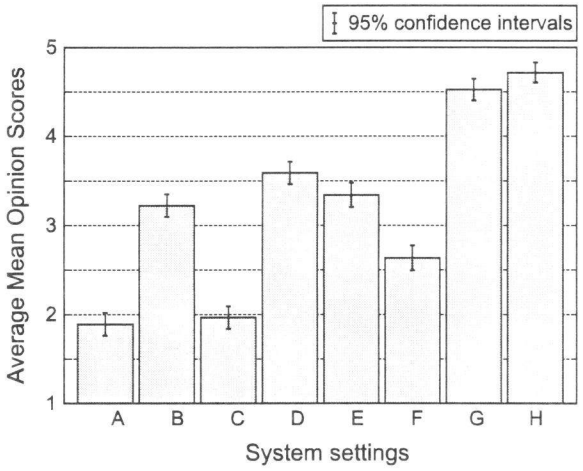


Fig. 10 Results of subjective evaluation (speaker FYM).

system settings performed best.

5. Constructing the Nitech-HTS 2005 Voices

5.1 Preparing Training Data

In the Blizzard Challenge 2005, the participants were asked to build four US-English voices (speakers BDL, CLB, RMS, and SLT) using the CMU ARCTIC databases [13]. Each of them consisted of 1132 phonetically balanced utterances, speech waveforms recorded at 16 kHz, phoneme segmentations, utterance information files, and pitch marks in the Festvox style [29]. Data for two of the four speakers (SLT and BDL) was released in advance and was used to optimize the system settings, e.g., training procedure. The remaining two speakers' data was released in January 2005. The system settings developed with the first two speakers except F_0 search range were used to construct the remaining two voices. Therefore, the building process for the later two voices was completely automatic.

To prepare training data, STRAIGHT mel-cepstral coefficients, $\log F_0$, and average values of aperiodicity measurements were extracted from the databases as described in Sect. 4.1. Feature vectors consisted of 40 mel-cepstral coefficients including the zeroth coefficient, $\log F_0$, average values of aperiodicity measures in five frequency sub-bands, their delta, and delta-delta.

5.2 Acoustic Modeling

The five-state left-to-right structure with no-skip was used. Each state output probability density function was consisted of five streams: STRAIGHT mel-cepstral coefficients with their delta and delta-delta, $\log F_0$, $\Delta \log F_0$, $\Delta^2 \log F_0$, and average values of aperiodicity measurements with their delta and delta-delta. The streams for STRAIGHT mel-cepstral coefficients and aperiodicity measurements were modeled by Gaussian distributions with diagonal covariance matrices. Each of the $\log F_0$, $\Delta \log F_0$, and $\Delta^2 \log F_0$ streams was modeled by an MSD consisting of a Gaussian distribution with a diagonal covariance matrix (voiced space) and a discrete distribution (unvoiced space). The state duration probability density functions of each HSMM were modeled by a multi-variate Gaussian distribution whose dimensionality was equal to the number of states.

Table 2 The numbers of leaf nodes of constructed decision trees for spectrum, F_0, aperiodicity measures, and durations.

	BDL	CLB	RMS	SLT
Spectrum	882	1,013	1,021	859
F_0	2,046	1,851	2,090	1,691
Aperiodicity	676	800	924	720
Duration	570	511	521	571

Table 3 System building times (Hours:Minutes:Seconds).

	Data preparation	Acoustic model training	Total
BDL	03:35:06	18:12:24	21:47:30
CLB	04:10:13	23:31:31	27:41:44
RMS	04:18:29	24:55:53	29:14:22
SLT	04:02:10	20:23:42	24:25:52

A modified version of the HMM-based speech synthesis software toolkit [30] was used for training acoustic models. After monophone HSMMs were initialized using the segmental k-means algorithm, they were re-estimated using the EM algorithm. Then their statistics were copied to context-dependent HSMMs. In the Nitech-HTS 2005 voices, contextual factors described in Zen et al. [31] were taken into account. They were extracted from utterance information files included in the databases using the feature extraction functions included in the Festival speech synthesis system. The context-dependent HSMMs were re-estimated (one iteration) and then the decision-tree-based context clustering technique [32] was used to construct the parameter sharing structure. The minimum description length (MDL) criterion [33] was used to stop tree growth [2], [34]. A decision tree was separately constructed for each state position of spectrum, F_0, aperiodicity measurements, and state duration. After re-estimating clustered HSMMs (four iterations), the constructed parameter sharing structure was untied. Then the untied context-dependent HSMMs were re-estimated (one iteration), and the decision-tree-based context clustering was applied once again. Table 2 shows the numbers of leaf nodes of constructed decision trees for spectrum, F_0, aperiodicity measures, and duration. Then the re-clustered HSMMs were re-estimated (five iterations). Finally, they were converted into the input file format of our HMM-based speech synthesis engine.[†]

Table 3 shows the total system building times of the Nitech-HTS 2005 voices on a 3.2 GHz Pentium 4 machine. It shows that these voices could be trained in about a day. Table 4 shows the footprints of the constructed Nitech-HTS 2005 voices. In this table, "Pdfs" means the files including the parameter values of state output and duration prob-

[†]This engine did not include text analyzer.

Table 4 Footprints of constructed voices in Kbytes.

	BDL	CLB	RMS	SLT
Pdfs	1,024	1,161	1,221	1,004
Trees	270	266	289	243
Engine	252			
Others	2			
Total	1,548	1,681	1,764	1,501

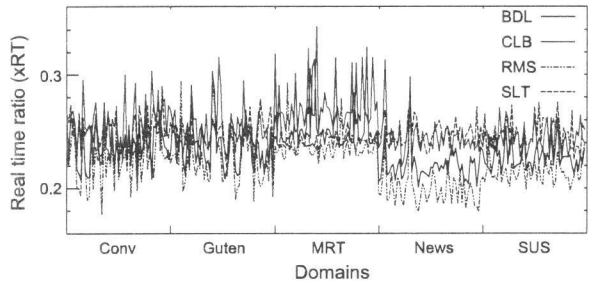


Fig. 11 Real-time ratios of the Nitech-HTS 2005 voices to synthesize speech waveforms.

ability distributions saved in 32-bit single precision floating point number, "Trees" denotes the tree files containing the decision trees of spectrum, F_0, aperiodicity measures, and durations saved in ASCII (HTK format). It can be seen from the figure that the footprint of each voice was less than 2 Mbytes. Their footprints can be reduced further without any quality degradation by eliminating redundant information [35]. Further reduction is also possible with little degradation of quality by vector quantization, saving pdf files in fixed point numbers instead of floating point ones, and pruning of the decision trees.

5.3 Synthesizing Speech

The test texts released from the organizers consisted of five different domains:

- Gutenberg novels (Guten),
- Standard news text (News),
- Conversational/dialog sentences (Conv),
- DRT/MRT phonetically confusable words, within sentences (MRT) [36],
- Semantically unpredictable sentences (SUS) [37].

Each text was converted into the corresponding context-dependent label sequence using the Festival speech synthesis system [38]. While analyzing a given text, any tags that specified accents, stresses or pronunciations to help the text analyzer were omitted, and none of the outputs of the text analyzer were manually corrected. Then the speech synthesis engine generated a speech waveform for the given context-dependent label sequence.

Figure 11 shows plots of the real-time ratios (×RT) of the Nitech-HTS 2005 voices to synthesize a speech waveform for a given label sequence on a 1.6 GHz Pentium 4 machine. An ×RT of 1 means the time to synthesize is the same as the speech signal time. A lower value of ×RT means better performance. It can be seen from the figure that the constructed Nitech-HTS 2005 voices could synthesize speech waveforms at 0.3 ×RT. It should be noted that this engine loaded both acoustic models and decision trees for every sentence. In fact, loading time took up about 60% of the synthesizing time. This means it can be reduced further.

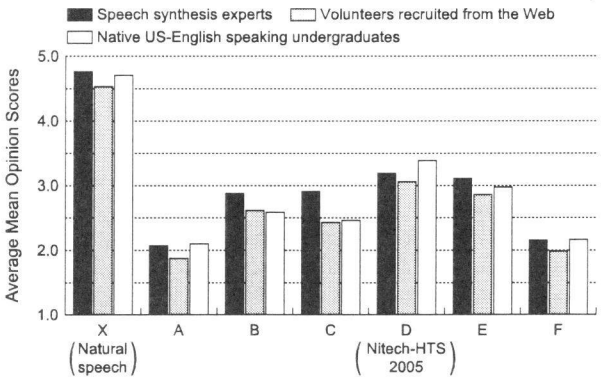


Fig. 12 Average MOSs of natural speech and all submitted systems.

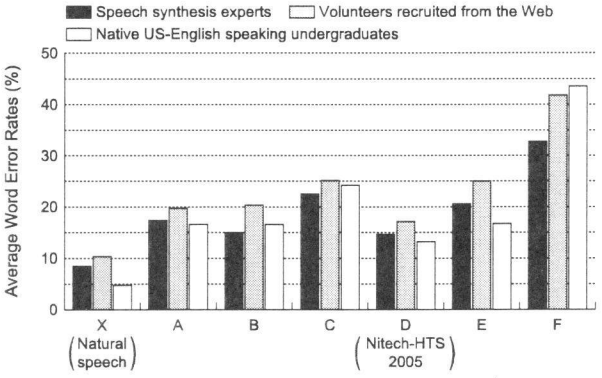


Fig. 13 Average WERs of natural speech and all submitted systems.

This year, six groups including five universities and one company participated in the challenge. Figures 12 and 13 show the average MOSs for the three MOS tests about naturalness and Word Error Rates (WER, the percentage of words that had errors) for the two intelligibility (type-in) tests carried out in the Blizzard Challenge 2005, respectively. To preserve participants' anonymity, the letters A through F are used to denote the systems, and X denotes a real speech reference condition of examples recorded by the voice talent. The Nitech-HTS 2005 is denoted by the letter D. In the Blizzard Challenge 2005, 50 speech synthesis experts, 60 volunteers recruited from the Web, and 58 native US-English speaking undergraduates participated in the evaluation. It can be seen from these figures that the performance of the Nitech-HTS 2005 voices was much better than expected, though it was still far from natural speech. They show that the Nitech-HTS 2005 voices achieved the highest MOSs and the lowest WERs with all types of listeners. Please refer to Bennett [12] for the details.

5.4 Discussion

In the Blizzard Challenge 2005, relatively small databases (around an hour for each speaker) were used. It is generally thought that the HMM-based approach is more appropriate than the unit selection approach for small databases because the HMM-based approach can potentially cover the given training data more effectively [39]. In these days of unit selection-based speech synthesis systems, a much larger speech database (e.g., more than 10 hours) is usually used. It may be worth studying at what size a speech database is more appropriate for the unit selection approach than for the HMM-based approach.

Sometimes it is difficult to assemble a speech database large enough to build a good unit selection system. For example, Black has reported that recording emotional or emphasized speech consistently has been difficult [40]. The HMM-based approach is very useful in these areas because it does not require a large amount of training data and can re-estimate new voices with only a few utterances from existing models that were trained with a large amount of data [41].

6. Conclusion

This paper detailed the HMM-based speech synthesis system (Nitech-HTS 2005) developed for the Blizzard Challenge 2005. We gave an overview of the basic HMM-based speech synthesis system and the new features integrated into Nitech-HTS 2005, such as STRAIGHT-based vocoding, HSMM-based acoustic modeling, and the speech parameter generation algorithm considering global variance. Constructed Nitech-HTS 2005 voices were able to synthesize a speech waveform at 0.3 ×RT (real-time ratio) on a 1.6 GHz Pentium 4 machine, and footprints of these voices were less than 2 Mbytes. Subjective listening test results showed that performance of the Nitech-HTS 2005 voices were much better than expected.

Acknowledgments

The authors thank Dr. Hideki Kawahara for permission to use the STRAIGHT vocoding method and Dr. Yoshihiko Nankaku for helpful discussions. The authors are also grateful to all the people who contributed to the Blizzard Challenge 2005. This work was partly supported by the MEXT e-Society project.

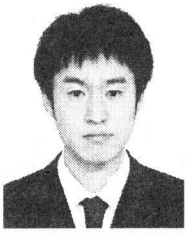
References

[1] R. Sproat, J. Hirschberg, and D. Yarowsky, "A corpus-based synthesizer," Proc. ICSLP, pp.563–566, 1992.
[2] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," Proc. Eurospeech, pp.2347–2350, 1999.
[3] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features," Proc. ICASSP, pp.660–663, 1995.
[4] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR," Proc. ICASSP, pp.805–808, 2001.
[5] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Speaker interpolation in HMM-based speech synthesis system," Proc. Eurospeech, pp.2523–2526, 1997.
[6] K. Shichiri, A. Sawabe, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Eigenvoices for HMM-based speech synthesis," Proc. ICSLP (Interspeech), pp.1269–1272, 2002.
[7] K. Tokuda and A. Black, "The Blizzard Challenge 2005: Evaluating corpus-based speech synthesis on common datasets," Proc. Interspeech (Eurospeech), pp.77–80, 2005.
[8] K. Tokuda and A. Black, "Speech synthesis research in a new age of cooperation and competition – The Blizzard Challenge," J. ASJ, vol.62, no.6, pp.466–470, 2006.
[9] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f_0 extraction: Possible role of a repetitive structure in sounds," Speech Commun., vol.27, pp.187–207, 1999.
[10] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Hidden semi-Markov model based speech synthesis," Proc. Interspeech (ICSLP), pp.1185–1180, 2004.
[11] T. Toda and K. Tokuda, "Speech parameter generation algorithm considering global variance for HMM-based speech synthesis," Proc. Interspeech (Eurospeech), pp.2801–2804, 2005.
[12] C. Bennett, "Large scale evaluation of corpus-based synthesizers: Results and lessons from the 2005 Blizzard Challenge," Proc. Interspeech (Eurospeech), pp.105–108, 2005.
[13] J. Kominek and A. Black, "CMU ARCTIC databases for speech synthesis," Tech. Rep. CMU-LTI-03-177, Carnegie Mellon University, 2003.
[14] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," Proc. ICASSP, pp.137–140, 1992.
[15] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling," Proc. ICASSP, pp.229–232, 1999.
[16] M. Ostendorf and I. Bulyko, "The impact of speech recognition on speech synthesis," Proc. IEEE Workshop on Speech Synthesis, 2002. CD-ROM proceeding.
[17] Y.J. Wu and R.H. Wang, "Minimum generation error training for HMM-based speech synthesis," Proc. ICASSP, pp.89–92, 2006.
[18] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," Proc. ICASSP, pp.1315–1318, 2000.
[19] Y. Ishimatsu, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Investigation of state duration model based on gamma distribution for HMM-based speech synthesis," IEICE Technical Report, SP2001-81, 2001.
[20] S. Imai, "Cepstral analysis synthesis on the mel frequency scale," Proc. ICASSP, pp.93–96, 1983.
[21] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Incorporation of mixed excitation model and postfilter into HMM-based text-to-speech synthesis," IEICE Trans. Inf. & Syst. (Japanese Edition), vol.J87-D-II, no.8, pp.1563–1571, Aug. 2004.
[22] M. Koike, K. Iwano, and S. Furui, "Improving naturalness using residual excitation for HMM-based speech synthesis," Proc. Spring Meeting of ASJ, pp.241–242, 2003.
[23] H. Kawahara, H. Katayose, A. Cheveigné, and R. Patterson, "Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of f_0 and periodicity," Proc. Eurospeech, pp.2781–2784, 1999.
[24] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis

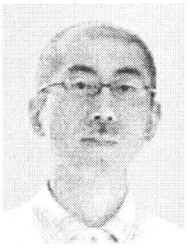
system straight," Proc. MAVEBA, pp.13–15, 2001.
[25] A. Oppenheim and D. Johnson, "Discrete representation of signals," Proc. IEEE, pp.681–691, 1972.
[26] E. Moulines and F. Charpentier, "Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones," Speech Commun., vol.9, pp.453–467, 1990.
[27] S. Levinson, "Continuously variable duration hidden Markov models for automatic speech recognition," Comput. Speech Lang., vol.1, pp.29–45, 1986.
[28] Y. Kishimoto, H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "A postfiltering technique for HMM-based speech synthesis," Proc. Autumn Meeting of ASJ, pp.279–280, 2002.
[29] A. Black and K. Lenzo, "Building synthetic voices," 2003. http://www.festvox.org/bsv/
[30] K. Tokuda, H. Zen, S. Sako, T. Yoshimura, J. Yamagishi, M. Tamura, and T. Masuko, "The HMM-based speech synthesis software toolkit," http://hts.ics.nitech.ac.jp/
[31] H. Zen, K. Tokuda, and T. Kitamura, "An introduction of trajectory model into HMM-based speech synthesis," Proc. ISCA SSW5, pp.191–196, 2004.
[32] J. Odell, The Use of Context in Large Vocabulary Speech Recognition, Ph.D. Thesis, Cambridge University, 1995.
[33] J. Rissanen, Stochastic Complexity in Stochastic Inquiry, World Scientific Publishing Company, 1980.
[34] K. Shinoda and T. Watanabe, "Acoustic modeling based on the MDL criterion for speech recognition," Proc. Eurospeech, pp.99–102, 1997.
[35] Y. Morioka, S. Kataoka, H. Zen, Y. Nankaku, K. Tokuda, and T. Kitamura, "Minutuarization of HMM-based speech synthesis," Proc. Autumn Meeting of ASJ, pp.325–326, 2004.
[36] A.S. House, C.E. Williams, M.H.L. Hecker, and K.D. Kryter, "Psychoacoustic speech tests: A modified rhyme test," Tech. Rep. ESD-TDR-63-403, U.S. Air Force Systems Command, Hanscom Field, Electronics Systems Division, 1963.
[37] C. Benot, M. Grice, and V. Hazan, "The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences," Speech Commun., vol.18, pp.381–392, 1996.
[38] A. Black, P. Taylor, and R. Caley, "The festival speech synthesis system," http://www.festvox.org/festival/
[39] A. Black, "Perfect synthesis for all of the people all of the time," Proc. IEEE Speech Synthesis Workshop, pp.160–163, 2002.
[40] A. Black, "Unit selection and emotional speech," Proc. Eurospeech (Interspeech), pp.1649–1652, 2003.
[41] J. Yamagishi, Average-Voice-Based Speech Synthesis, Ph.D. Thesis, Tokyo Institute of Technology, 2006.

Heiga Zen was born in Osaka, Japan on March 4, 1979. He received the A.E. degree in electronic and information engineering from Suzuka National College of Technology, Suzuka, Japan, in 1999, and the B.E., M.E., and Dr. Eng. degrees in computer science, electrical and computer engineering, and computer science and engineering from Nagoya Institute of Technology, Nagoya, Japan, in 2001, 2003, and 2006, respectively. During 2003, he was an intern researcher at ATR Spoken Language Translation Research Laboratories (ATR-SLT), Kyoto, Japan. From June 2004 to May 2005, he was an intern/co-op researcher in the Human Language Technology group at IBM T.J. Watson Research Center, Yorktown Heights, NY. He is currently a postdoctoral fellow of the MEXT e-Society project at Nagoya Institute of Technology. His research interests include statistical speech recognition and synthesis. He received the Awaya Award in 2006 from the Acoustical Society of Japan (ASJ). He is a member of ASJ.

Tomoki Toda was born in Aichi, Japan on January 18, 1977. He received the B.E. degree in electrical engineering from Nagoya University, Nagoya, Japan, in 1999, and the M.E. and Ph.D. degrees in engineering from the Graduate School of Information Science, Nara Institute of Science and Technology (NAIST), Nara, Japan, in 2001 and 2003, respectively. During 2001–2003, he was an intern researcher and a visiting researcher at ATR Spoken Language Translation Research Laboratories (ATR-SLT), Kyoto, Japan. He was a Research Fellow of the Japan Society for the Promotion of Science (JSPS) in the Graduate School of Engineering, Nagoya Institute of Technology, Nagoya, Japan, during 2003–2005. He was a visiting researcher at the Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, from October 2003 to September 2004. He is currently an Assistant Professor of the Graduate School of Information Science, NAIST and a visiting researcher at ATR-SLT. He received the TELECOM System Technology Award for Student from the Telecommunications Advancement Foundation, Japan, in 2003. He is a member of the Acoustical Society of Japan (ASJ), and International Speech Communication Association (ISCA).



Masaru Nakamura received the B.E. degree in computer science from Nagoya Institute of Technology, Nagoya, Japan, in 2005. He is currently a candidate for the M.E. degree in the Department of Computer Science at Nagoya Institute of Technology. His research interests include text-to-speech synthesis. He is a member of the Acoustical Society of Japan (ASJ).



Keiichi Tokuda received the B.E. degree in electrical and electronic engineering from Nagoya Institute of Technology, Nagoya, Japan, the M.E. and Dr. Eng. degrees in information processing from Tokyo Institute of Technology, Tokyo, Japan, in 1984, 1986, and 1989, respectively. From 1989 to 1996 he was a Research Associate at the Department of Electronic and Electric Engineering, Tokyo Institute of Technology. From 1996 to 2004 he was an Associate Professor at the Department of Computer Science, Nagoya Institute of Technology as an Associate Professor, and now he is a Professor at the same institute. He is also an Invited Researcher at ATR Spoken Language Translation Research Laboratories (ATR-SLT), Kyoto, Japan, and was a Visiting Researcher at the Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, from 2001 to 2002. He is a co-recipient of the Paper Award and the Inose Award both from the Institute of Electronics, Information and Communication Engineers of Japan (IEICE) in 2001, and the TELECOM System Technology Award from the Telecommunications Advancement Foundation, Japan, in 2001. He was a member of the Speech Technical Committee of the IEEE Signal Processing Society. His research interests include speech coding, speech synthesis and recognition, and statistical machine learning.