Overview of NIT HMM-based speech synthesis system for Blizzard Challenge 2009

Keiichiro Oura[†], *Yi-Jian Wu*[‡*] , *Keiichi Tokuda*[†]

[†] Department of Computer Science and Engineering, Nagoya Institute of Technology, Japan
[‡] TTS group, Microsoft Business Division, China

uratec@sp.nitech.ac.jp, yijiwu@microsoft.com, tokuda@sp.nitech.ac.jp

Abstract

We describe a hidden Markov model (HMM)-based speech synthesis system developed at the Nagoya Institute of Technology (NIT) for Blizzard Challenge 2009. We incorporated several state-of-the-art technologies into this system, including the Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum (STRAIGHT) vocoder, minimum generation error (MGE) training, phone duration modeling, parameter generation algorithm considering global variance, and linear spectrum pair (LSP)-based formant enhancement. The runtime of system synthesizes speech around 0.3 xRT (real time ratio), and its footprint is less than 25 MB. The results of listening tests showed that the overall speech quality and intelligibility of our systems are better than most other systems, especially when we have better labeling for a speech corpus.

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

1. Introduction

The hidden markov model (HMM) has been commonly used for speech recognition [1], and there has been significant progress over the decade. Recently an HMM-based speech synthesis method was proposed [2]. In this method, the spectrum, pitch, and duration are modeled simultaneously in a unified framework of HMMs [3], and the parameter sequence is generated by maximizing the likelihood of the HMMs related to the parameter sequence under the constraint of the explicit relationship between static and dynamic features [4]. Compared to other synthesis methods, this method has several advantages, 1) under its statistical training framework, it can learn salient statistical properties of speakers, speaking styles [5], emotions [6], etc., from the speech corpus; 2) many techniques developed for HMM-based speech recognition can be applied to speech synthesis [7, 8]; 3) voice characteristics of synthesized speech can be easily controlled by modifying acoustic statistics of HMMs [9, 10]. Furthermore, it can generate smooth and stable speech under a small footprint. As a result, HMM-based speech synthesis gradually became popular both in research and application [11–13].

Although the performance of the conventional HMM-based speech synthesis framework is quite good, the quality of synthesized speech still needs to be improved. In recent years, several techniques had been proposed to improve the quality of synthesized speech for HMM-based speech synthesis, including a high quality vocoder Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum (STRAIGHT) [14] for spectral analysis, a minimum generation error (MGE) [15] criterion for model training, phone duration modeling [13], parameter generation algorithm considering global variance (GV) [16], and a postfilter for the linear spectrum pair (LSP) to enhance the formant for generated speech [13]. In NIT's system for the Blizzard Challenge, we use the HMM-based speech synthesis method and integrate these state-of-the-art technologies.

The rest of the paper organized as follows. In sections 2, 3, 4, 5, and 6, we briefly review STRAIGHT vocoding, MGE training with log spectral distortion, phone duration modeling, parameter generation algorithm considering GV, and LSP-based formant enhancement, respectively. In section 7, we describe experiments for evaluating our system, and present the results. Finally, our conclusions are given in section 8.

2. STRAIGHT vocoding

We use STRAIGHT a high-quality speech vocoding method proposed by Kawahara *et al.* [14]. It consists of three main components, F_0 extraction, spectral, and aperiodic analysis, and speech synthesis.

The STRAIGHT method is used to automatically extract F_0 with fixed-point analysis [17]. We use a two-stage extraction to alleviate errors of the F_0 extraction. First, we perform the F_0 extraction for all training data for each speaker in which a search range is set to 55-500 Hz. Using of a histogram of the extracted F_0, we roughly estimate the F_0 range of each speaker. Then, F_0 is again extracted in the speaker-specific range.

Using the extracted F_0, we use the STRAIGHT method to perform pitch-adaptive spectral analysis combined with a surface reconstruction method in the time-frequency region to remove signal periodicity. As a spectral parameter, we use the 40-th STRAIGHT mel linear spectrum pair (mel-LSP) coefficients. An aperiodicity measure on the frequency domain based at a ratio between the lower and upper smoothed spectral envelopes to represent the relative energy distribution of aperiodic components [18] is also extracted. As a parameter for constructing a mixed excitation sources in speech synthesis, average values of the aperiodicity measures on five frequency bands, 0-1, 1-2, 2-4, 4-6, and 6-8 kHz are used.

3. MGE training with log spectral distortion

In the conventional HMM-based speech synthesis framework, Maximum Likelihood (ML) criterion was adopted for HMM

*The work in this paper was done when the author was in Nagoya Institute of Technology, Japan.

training. However, there are two issues [15] related to ML-based HMM training for speech synthesis, including the mismatch between training and application of the HMMs and the ignorance of constraint between static and dynamic features. To resolve these two issues, a minimum generation error (MGE) criterion [15] had been proposed for HMM training, where a generation error function using Euclidean distance was defined, and the HMM parameters were optimized to minimize the total generation errors of training data. Furthermore, a log spectral distortion (LSD) was adopted to replace the Euclidean distance to define the generation error between the original and generated LSPs [19] in MGE training, and the quality of synthesized speech was improved [20]. The LSPs extracted from original speech waveforms were used as the reference measuring for spectral distortion in this training.

We use the MGE-LSD training by directly using the original spectrum for measuring spectral distortion [21]. First, we adopt the spectral envelope extracted from the original speech waveforms using the STRAIGHT method [14] as a reference to calculate the LSDs and define the generation error function. However, the speech waveforms are the actual target signals we want to simulate. The STRAIGHT-based spectral analysis can be basically regarded as a process for recovering the spectral envelope from the short-time fast Fourier transform (FFT) spectrum calculated from the speech waveforms. However, some information may be lost in this process. Therefore, we directly use the short-time FFT spectrum calculated from speech waveforms as the original reference spectrum for LSD calculation. Since only the harmonics of the FFT spectrum are coincident with the underlying spectral envelope, the LSD between generated LSPs and original FFT spectrum is calculated by sampling at the harmonic frequencies. The MGE-LSD training with FFT spectrum can be regarded as a unified training framework by incorporating spectral analysis and parameter generation into model training. This is a similar concept to the analysis-by-synthesis in speech coding and the closed-loop training [22] for concatenative speech synthesis.

4. Phone duration modeling

In the conventional framework, a state duration model is trained to predict the duration of every state in the utterance for synthesis. A phone duration model is also constructed in our system, by taking into account a phonetic unit, and combined with the state duration model for predicting the duration of each state [13].

5. Parameter generation algorithm considering global variance

Usually, speech parameter vector sequences generated from HMMs are smoothed excessively. Synthesized speech using over-smoothed parameters sounds muffled. To reduce this effect, we use a parameter generation algorithm considering GV of the generated parameters [16].

We apply this algorithm to both spectral and F_0 parameter generation processes. One GV is calculated from a parameter sequence over the entire of one utterance. It should be noted that only voiced frames are used for calculating GV of F_0 parameters. Probability density on GV is modeled using a Gaussian distribution with a diagonal covariance matrix.

In parameter generation, we first generate a parameter trajectory with the speech parameter generation algorithm. Then, we convert the generated trajectory so that its GV is equal to a mean of Gaussian distribution. Using this converted trajectory as an initial value, we iteratively calculate the parameter trajectory that maximizes the likelihood function consisting of the output probability of the parameter sequence and that of its GV with the Newton-Raphson method.

We changed the GV Gaussian probability density function (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 to tie their parameters. The number of leaf nodes of the decision trees was automatically determined by the MDL criterion [23]. To simplify implementation, only sentence-level contextual features (e.g., number of phonemes in a sentence) were used at this time. Furthermore, we calculated the GV vector from only speech and excluded silence and pause regions from the calculation, based on automatic segmentation, to improve the estimation accuracy of the GV vector.

6. LSP-based formant enhancement

We select mel linear spectral pair (mel-LSP) to present each frame spectral envelop estimated using the STRAIGHT method because LSPs relate more closely to formant positions and have better smoothness among adjacent frames. Because of the averaging effect of statistic modeling, the spectra reconstructed from parameter generation are always over-smoothed and the formants are broaden, which make the synthetic speech sound muffled. The relationship between spectral peaks and LSP, especially the difference between its adjacent orders, is used to enhance the formants of synthesized speech [13].

7. Experiments

7.1. Experimental conditions for all tasks

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. In Blizzard Challenge 2009, an English speech database consisting of about 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 iFlytek Beijing, China.

Speech signals were sampled at a rate of 16kHz and windowed with an F_0-adaptive Gaussian window with a 5-ms shift. The feature vectors consisted of 40 STRAIGHT mel-LSP coefficients, log F_0, aperiodicity measures, and their dynamic and acceleration coefficients. We used 5-state left-to-right context-dependent multi-stream MSD-HMMs without skip paths. Each state had a single Gaussian pdf with a diagonal covariance matrix. The iteration for the GV calculation was 20, and the post-filter rate was 0.8.

7.2. Experimental conditions for English hub task 1 (EH1)

Database An approximately 15-hour speech database (roger) with no modification.

Phoneset All labels were generated using Unilex-RPX and Festival's Multisyn module.

Context-clustering Thresholds of MDL criterion α were 1.2

for the spectrum, log F_0, aperiodicity measures, and duration.

Global variance GV weights were 0.7 for the spectrum and log F_0.

7.3. Experimental conditions for English hub task 2 (EH2)

Database An approximately 1-hour speech database (roger) with no modification.

Phoneset All labels were generated using Unilex-RPX and Festival's Multisyn module.

Context-clustering Thresholds of MDL criterion α were 1.2 for the spectrum, log F_0, aperiodicity measures, and duration.

Global variance GV weights were 0.7 for the spectrum and log F_0.

7.4. Experimental conditions for English spoke task 1 (ES1)

Database The CMU-ARCTIC speech database was used for the average voice model. This database contains a set of approximately one thousand phonetically balanced sentences uttered by three male speakers (AWB, BDL, RMS) with a total duration of about 3.5 hours.

Phoneset All labels were generated using Unilex-RPX and Festival's Multisyn module.

Context-clustering Thresholds of MDL criterion α were 0.9, 1.3, 1.3, and 1.3 for the spectrum, log F_0, aperiodicity measures, and duration, respectively.

Global variance GV weights are 0.7 for the spectrum and log F_0.

7.5. Experimental conditions for Mandarin hub task (MH)

Database An approximately 10-hour speech database by iFlytek with no modification.

Phoneset All labels were released by iFlytek with no modification.

Context-clustering Thresholds of MDL criterion α were 0.9, 1.3, 1.3, and 1.3 for the spectrum, log F_0, aperiodicity measures, and duration, respectively.

Global variance GV weights were 0.4 and 1.0 for the spectrum and log F_0, respectively.

7.6. Experimental conditions for Mandarin spoke task 1 (MS1)

Database The iFlytek speech database was used for the average voice model. This database contains one thousand phonetically balanced sentences uttered by one female speaker (f3) with a total duration of about 2.5 hours.

Phoneset All labels were released by iFlytek with no further modification.

Context-clustering Thresholds of MDL criterion α were 0.9, 1.3, 1.3, and 1.3 for the spectrum, log F_0, aperiodicity measures, and duration, respectively.

Global variance GV weights were 0.4 and 1.0 for the spectrum and log F_0, respectively.

7.7. Listening tests

About 1500 and 1000 test sentences were generated for English and Mandarin, respectively. To evaluate naturalness and similarity, 5-point mean opinion score (MOS) and differential mean opinion score (DMOS) tests were conducted. The scale for the MOS test was from 5 for "completely natural" to 1 for "completely unnatural". The scale for the DMOS tests was from 5 for "sounds like exactly the same person" to 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. The evaluations were conducted over a six-week period via the Internet.

7.8. Experimental results of the English systems

Figures 1-9 show the results of the English systems. "A", "B", "C", and "D" correspond to real speech, the Festival "Multisyn" benchmark speech synthesis system [24], the HTS benchmark system 2005 [25], and 2007 [26], respectively. The Festival system uses a conventional unit-selection method. The HTS Benchmark systems are a standard statistical parametric system using HTS toolkit version 2.1 and STRAIGHT.

Our system was equal to the Festival one in naturalness for EH1 task (Figure 1). On the other hands, our system achieved a higher score than the Festival one in naturalness for EH2 task (Figure 4). It seems that our labels for the large database were not accurate. There are significant differences between real speech and all other systems from the point of view of naturalness and similarity.

Intelligibility of our system was best with the smaller dataset (Figure 6). 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.

7.9. Experimental results of the Mandarin systems

Figures 10-15 show the results of the Mandarin systems. As with the English systems, "A", "C", and "D" correspond to real speech, the HTS benchmark system 2005, and 2007, respectively. There is no Festival benchmark system for Mandarin.

Our system scored best in naturalness for MH task (Figures 10), character error rate for MH task (Figure 12), in naturalness for MS1 task (Figure 13), and character error rate for MS1 task (Figure 15). It seems that the labels given by iFlytek were fortunately accurate. However, there are significant differences between real speech and all other systems from the point of view of similarity.

8. Conclusions

We described HMM-based speech synthesis system developed at the Nagoya Institute of Technology (NIT) for Blizzard Challenge 2009. We incorporated several state-of-the-art technologies into this system, including the STRAIGHT vocoder, minimum generation error training, phone duration modeling, parameter generation algorithm considering GV, and the LSP-based formant enhancement. The runtime of system synthesizes speech around 0.3 xRT (real time ratio) and its footprint is less than 25MB. The results of listening tests showed that the overall speech quality and intelligibility of our systems are better than most other systems, especially when we have better labeling for the speech corpus.

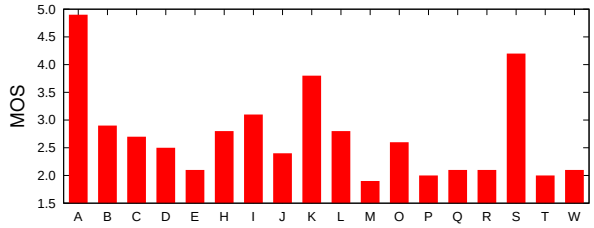

Figure 1: Experimental results: naturalness for EH1 task. (L: NIT system.)

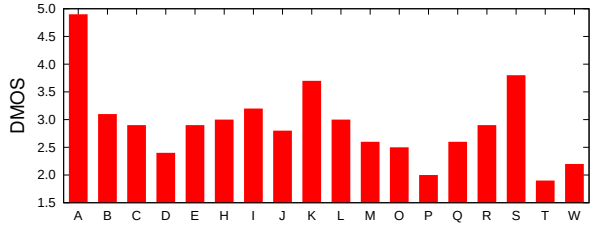

Figure 2: Experimental results: similarity scores comparing to original speaker for EH1 task. (L: NIT system.)

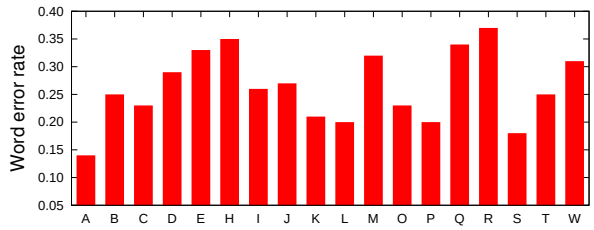

Figure 3: Experimental results: word error rate for EH1 task. (L: NIT system.)

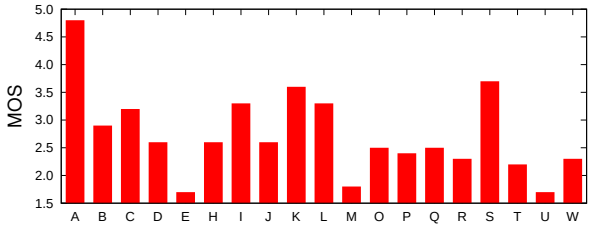

Figure 4: Experimental results: naturalness for EH2 task. (L: NIT system.)

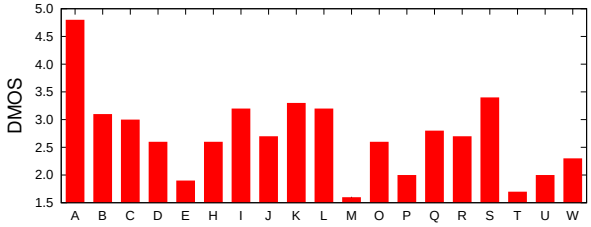

Figure 5: Experimental results: similarity scores comparing to original speaker for EH2 task. (L: NIT system.)

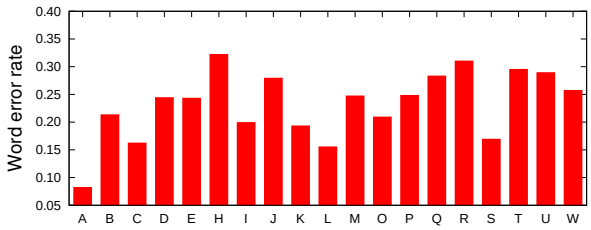

Figure 6: Experimental results: word error rate for EH2 task. (L: NIT system.)

9. Acknowledgements

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 http://www.emime.org), and the Strategic Information and Communications R&D Promotion Programme (SCOPE), Ministry of Internal Affairs and Communication, Japan.

10. References

[1] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proc. IEEE, vol. 77, pp. 257–286, 1989.

[2] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Speech synthesis from HMMs using dynamic features," Proc. of ICASSP, pp. 389–392, 1996.

[3] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," Proc. of Eurospeech, vol. 5, pp. 2347–2350, 1999.

[4] K. Tokuda, T. Kobayashi and S. Imai, "Speech parameter generation from HMM using dynamic features," Proc. of ICASSP, pp. 660–663, 1995.

[5] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis," IEICE Trans. Information and Systems, Vol. E88-D, no. 3, pp. 502–509, 2005.

[6] R. Tsuzuki, H. Zen, K. Tokuda, T. Kitamura, M. Bulut, and S. Narayanan, "Constructing emotional speech synthesizers with limited speech database," Proc. of ICSLP, vol. 2, pp. 1185–1188, 2004.

[7] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Hidden semi-Markov model based speech synthesis," Proc. of ICSLP, vol. 2, pp. 1397–1400, 2004.

[8] J. Yamagishi and T. Kobayashi,"Adaptive training for hidden semi-Markov model," Proc. of ICASSP, vol. 2, pp. 1213–1216, 2004.

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

[10] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Adaptation of pitch and spectrum for HMM-based

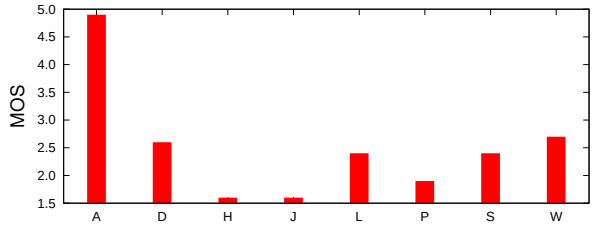

Figure 7: Experimental results: naturalness for ES1 task. (L: NIT system.)

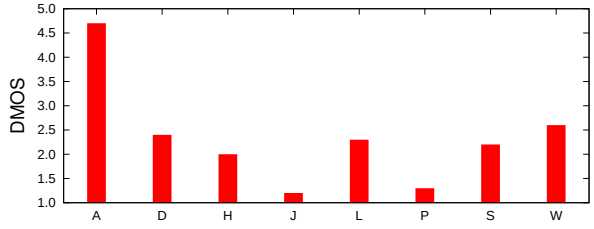

Figure 8: Experimental results: similarity scores comparing to original speaker for ES1 task. (L: NIT system.)

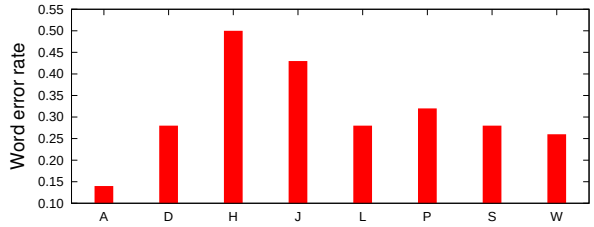

Figure 9: Experimental results: character error rate for ES1 task. (L: NIT system.)

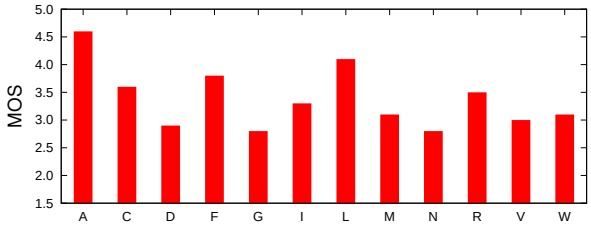

Figure 10: Experimental results: naturalness for MH task. (L: NIT system.)

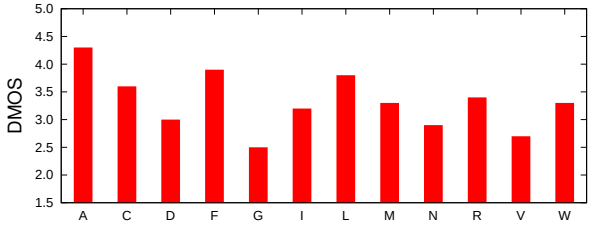

Figure 11: Experimental results: similarity scores comparing to original speaker for MH task. (L: NIT system.)

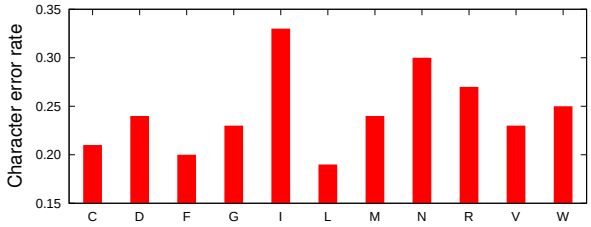

Figure 12: Experimental results: character error rate for MH task. (L: NIT system.)

speech synthesis using MLLR," in Proc. of ICASSP, pp. 805–808, 2001.

[11] K. Tokuda, H. Zen, and A. W. Black, "An HMM-based speech synthesis system applied to English," IEEE Speech Synthesis Workshop, California, 2002.

[12] H. Zen and T. Toda, "An Overview of Nitech HMM-based Speech Synthesis System for Blizzard Challenge 2005," in Proc. of Eurospeech, pp. 93–96, 2005.

[13] 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," Interspeech 2006 satellite meeting, Blizzard Challenge 2006.

[14] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigne "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.

[15] Y. J. Wu and R. H. Wang, "Minimum generation error training for HMM-based speech synthesis," in Proc. of ICASSP, vol. 1, pp. 889–892, 2006.

[16] T. Toda and K. Tokuda, "Speech parameter generation algorithm considering global variance for HMM-based speech synthesis," in Proc. of Interspeech, pp. 2801–2804, 2005.

[17] H. Kawahara, H. Katayose, A. Cheveigne, and R. Patterson, "Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of f_0 and periodicity," in Proc. of Eurospeech, pp. 2781–2784, 1999.

[18] 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," in Proc. of MAVEBA, pp. 13–15, 2001.

[19] F. Itakura, "Line spectrum representation of linear predictive coefficients of speech signals," in J. Acoust. Soc. Amer., vol. 57, pp. 535, 1975.

[20] Y. J. Wu and K. Tokdua, "Minimum generation error training with direct log spectral distortion on LSPs for HMM-based speech synthesis," in Proc. of Interspeech, pp. 577–580, 2008.

[21] Y. J. Wu and K. Tokuda, "Minimum generation error training by using original spectrum as reference for log spectral distortion measure," in Proc. of ICASSP 2009 (to be appear).

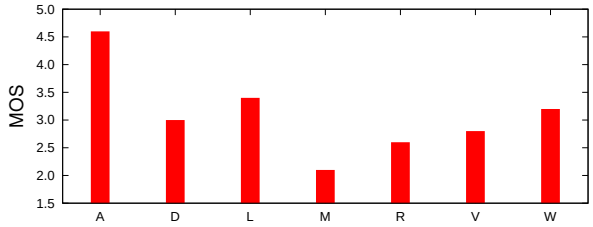

Figure 13: Experimental results: naturalness for MS1 task. (L: NIT system.)

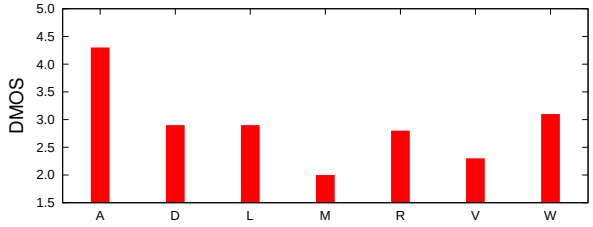

Figure 14: Experimental results: similarity scores comparing to original speaker for MS1 task. (L: NIT system.)

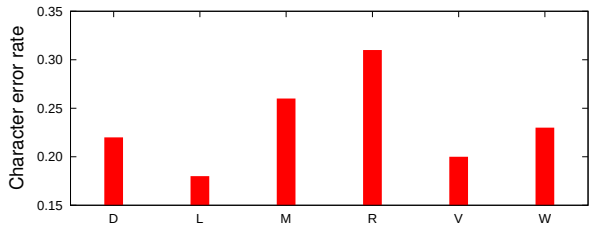

Figure 15: Experimental results: character error rate for MS1 task. (L: NIT system.)

[22] M. Akamine and T. Kagoshima, “Analytic generation of synthesis units by closed loop training for totally speaker driven text to speech system (TOS drive TTS),” in Proc. of ICSLP, 1998.

[23] K. Shinoda and T. Watanabe, “Acoustic modeling based on the MDL criterion for speech recognition,” in Proc. of Eurospeech, vol. 1, pp. 99–102, 1997.

[24] K. Richimond, V. Strom, R. Clark, J. Yamagishi, and S. Fitt, “Festival Multisyn voices for the 2007 Blizzard Challenge,” in Proc. of BLZ3-2007, 2007.

[25] 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, 2007.

[26] J. Yamagishi, T. Nose, H. Zen, T. Toda, and K. Tokuda, “Performance evaluation fo the speaker-independent HMM-based speech synthesis system HTS-2007 for Blizzard Challenge 2007,” in Proc. of ICASSP, pp. 3957–3960, 2008.