The NAIST Text-to-Speech System for the Blizzard Challenge 2015

Shinnosuke Takamichi[1], *Kazuhiro Kobayashi*[1], *Kou Tanaka*[1], *Tomoki Toda*[1,2], *Satoshi Nakamura*[1]

[1]Graduate School of Information Science, Nara Institute of Science and Technology, Japan
[2]Information Technology Center, Nagoya University, Japan

{shinnosuke-t, kazuhiro-k, ko-t, s-nakamura}@is.naist.jp, tomoki@icts.nagoya-u.ac.jp

Abstract

This paper presents a text-to-speech (TTS) system developed at Nara Institute of Science and Technology (NAIST) for the Blizzard Challenge 2015. The tasks of this year's challenge are the mono-lingual speech synthesis (IH1 hub task) for 6 Indian languages including Bengali, Hindi, Malayalam, Marathi, Tamil, and Telugu, and the multi-lingual speech synthesis (IH2 spoke task) for Indian language and English. We have developed our TTS system based on a statistical parametric speech synthesis technique using a hidden Markov model (HMM). To improve quality of synthetic speech, we have newly implemented two techniques for the traditional HMM-based speech synthesis framework, 1) pre-processing for producing smooth parameter trajectories to be modeled with HMM and 2) speech parameter generation considering the modulation spectrum. The developed system has been submitted to the mono-lingual task and its performance has been demonstrated from the results of large-scaled subjective evaluation.

Index Terms: HMM-based speech synthesis, modulation spectrum, parameter trajectory smoothing, continuous F_0 contour, speech parameter generation

1. Introduction

In order to better understand different speech synthesis techniques to develop a corpus-based text-to-speech (TTS) system using a common dataset, Blizzard Challenge was devised in January 2005 [1] and has been held every year since then [2]. This year's Blizzard Challenge has two tasks, 1) a mono-lingual speech synthesis task (IH1 hub task) for 6 Indian languages consisting of Bengali, Hindi, Malayalam, Marathi, Tamil, and Telugu, and 2) a multi-lingual speech synthesis task (IH2 spoke task) for Indian language and English. The Indian datasets [3] provided in the challenge consist of speech waveforms and the corresponding texts only. The size of the speech data in each Indian language is about 4 hours for Hindi, Tamil and Telugu, and 2 hours for Bengali, Malayalam, and Marathi. They are sampled at 16 kHz. The text data is provided in UTF-8 format. As only the plain text data is provided without any additional information, such as a phoneme set, syllable definition, and prosodic labels, participants need to develop a natural language processing module (front-end) as well as a speech waveform generation module (back-end) to develop their own TTS systems.

Our research group, a speech synthesis group of Augmented Human Communication laboratory, Nara Institute of Science and Technology (NAIST), studies various speech synthesis techniques, such as high-quality statistical parametric speech synthesis techniques [4], real-time voice conversion techniques for augmented speech production [5] (e.g., voice/vocal effector [6] or a speaking aid system for laryngectomees [7]), towards the development of technologies to break down existing barriers in our speech communication. To submit a TTS system from our group to the Blizzard Challenge 2015, we have developed our own system, the NAIST TTS system based on a statistical parametric speech synthesis technique using hidden Markov model (HMM) [8]. To improve quality of synthetic speech, two techniques are newly implemented for the traditional HMM-based speech synthesis framework, 1) pre-processing for producing smooth parameter trajectories to be modeled with HMM and 2) speech parameter generation considering the modulation spectrum (MS) of speech parameters [9, 10]. The developed system has been submitted to the mono-lingual task and its performance has been demonstrated from the results of large-scaled subjective evaluations.

This paper describes details of the NAIST TTS system. We also briefly discuss the results of large-scaled subjective evaluations on naturalness, similarity to the original speaker, and intelligibility, which were provided from the organizers.

2. The NAIST HMM-based TTS system for mono-lingual task

The NAIST TTS system has 4 main modules; a *text processing module*, a *speech processing module*, a *training module*, and a *speech synthesis module*, as shown in Fig. 1. Context labels used for HMM training are generated using the existing toolkit or our developed rule-based grapheme-to-phoneme converter and syllable estimator in the text processing module. Smoothly varying speech parameter sequences are extracted in the speech processing module. The context-dependent phoneme HMMs and the MS probability density functions are trained using the context labels and the speech parameters in the training module. Finally, a speech waveform is generated from these trained models corresponding to a given text to be synthesized in the synthesis module.

2.1. Text processing module

2.1.1. Text analysis

Because the provided Indian datasets do not include any linguistic information, such as a phoneme set and prosodic labels, which is usually needed to describe speech parameters corresponding to a given text, it is indispensable to predict these information from the given text. In the last year's challenge, some participants used several techniques to cope with this issue, *e.g.*, the use of an existing speech recognizer for a different language to extract auxiliary linguistic information [11] or the development of a fully data-driven text analyzer [12].

In this year's challenge, we used hand-crafted text analyzers. We used text analyzers developed with language-specific recipes distributed by Festvox [13] for Bengali, Hindi, Tamil, and Telugu. Additionally, we also developed a text analyzer for

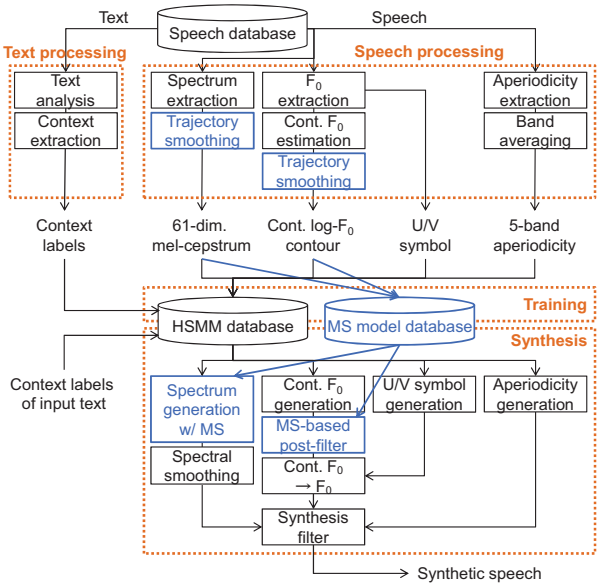

Figure 1: An overview of the NAIST TTS system for the Blizzard Challenge 2015. The orange-colored boxes indicate 4 main modules, a text processing module, a speech processing module, a training module, and a synthesis module. The blue-colored items are techniques newly implemented for the traditional HMM-based speech synthesis framework to improve synthetic speech quality, where "cont. F_0" and "MS" indicate the continuous F_0 and the modulation spectrum, respectively.

Marathi with the recipe for Hindi because Marathi has a certain similarity to Hindi. For Malayalam, we developed a rule-based grapheme-to-phoneme converter [14] dealing with chillus and a rule-based syllable estimator considering specific characteristics of Malayalam, such as dependent vowel signs.

2.1.2. Context label generation

The context labels are required to train the context-dependent phoneme HMMs. Our context labels were designed on the basis of the contextual factors used in HTS speaker adaptation/adaptive training demo for English [15]. An example of the contextual factors used in our context label definition is shown as follows:

- phoneme, syllable structure, and stress
- vowel/consonant, articulator position, and voicing/unvoicing
- position of phoneme, syllable, and word
- the number of phonemes, syllables, and words.

Note that stress information is not used for Malayalam because it is not extracted in our text analysis module.

2.2. Speech analysis module

2.2.1. Speech feature extraction

A high-quality speech analysis-synthesis system is required to develop a high-quality TTS synthesizer. We conducted preliminary evaluation to compare analysis-synthesized speech quality by STRAIGHT [16, 17] and WORLD [18, 19] as a high-quality analysis-synthesis system. From the result of this preliminary

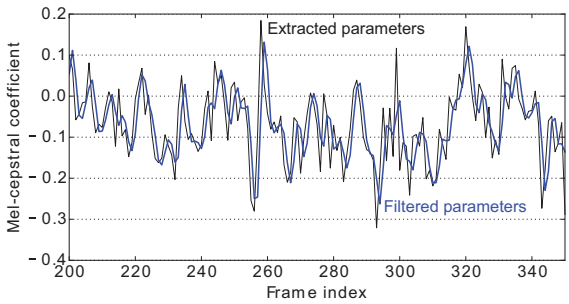

Figure 2: An example of the 20-th mel-cepstral coefficient sequences before and after the low pass filtering that removes the MS components over than 50 Hz. We can see that some fluctuations have been removed.

evaluation, we decided that spectral envelope and aperiodicity were extracted with STRAIGHT, given F_0 extracted with WORLD. They were parameterized into the 0^{th}-through-60^{th} mel-cepstral coefficients, band aperiodicity, and log-scaled F_0, where the band aperiodicity was calculated by averaging aperiodicity of each frequency component in 5 frequency bands [20]. The shift length was set to 5 ms. Moreover, the continuous F_0 contour [21] was additionally produced from the extracted F_0 contour. The spline-based interpolation algorithm was used to estimate F_0 values at unvoiced regions [9].

2.2.2. Parameter trajectory smoothing

Many fluctuations are usually observed over a time sequence of some speech parameters, such as mel-cepstral coefficients. They are represented as the MS of the temporal parameter sequence, *i.e.*, power spectrum of the parameter sequence. As described in [9], we have found that the effect of the MS components in high MS frequency bands on quality of analysis-synthesized speech is negligible, *e.g.*, more than 50 Hz MS frequency components for the mel-cepstral coefficient sequence and more than 10 Hz MS frequency components for the continuous F_0 contour.[1] To make the HMMs focus on the modeling of only auditory informal components, low-pass filter (LPF) was applied to each parameter sequence. The cutoff frequency of LPF was set to 50 Hz for the mel-cepstral coefficients and 10 Hz for the continuous F_0 contour, respectively. An example of this parameter trajectory smoothing for the mel-cepstral coefficients is shown in Fig. 2.

2.3. Training module

2.3.1. Hidden semi-Markov model training

The context-dependent phoneme hidden semi-Markov models (HSMMs) were trained on the basis of a maximum likelihood criterion in a unified framework to model individual speech components [23]. Five-state left-to-right HSMMs were used for every Indian language. The feature vector consisted of mel-cepstral coefficients (61 dimensions), continuous log-scaled F_0 contour (1 dimension), band aperiodicity (5 dimensions), and their delta and delta-delta features, and discrete log-scaled F_0 contour (1 dimension) consisting of unvoiced symbols. The total dimensionality of the feature vector is 202. Only for Hindi, we used the 0^{th}-through-24^{th} mel-cepstral coefficients as we found that the spectral parameter because the 61-dimentional

[1] Micro-prosody is captured by these components [22].

Table 1: Number of training utterances for each language. No external data was used.

Bengali	Marathi	Hindi	Tamil	Malayalam	Telugu
1304	1197	1709	1462	1289	2481

mel-cepstral coefficients were not well modeled in the HSMMs. The spectrum, continuous F_0, band aperiodicity components were modeled with the multi-stream continuous distributions. The discrete F_0 contour was additionally modeled with the multi-space distributions [24] to determine the voiced/unvoiced region of the continuous F_0 contour in the synthesis module. The tree-based clustering with the minimum description length (MDL) criterion [25] was employed. The stream weights were set to 1.0 (spectrum), 1.0 (continuous F_0), 1.0 (discrete F_0)[2], and 0.0 (aperiodicity). The number of utterances included in training data is listed in Table 1.

2.3.2. MS model training

Gaussian distributions were also trained as the context-independent MS models for the spectrum and continuous F_0 contour. The utterance-level mean was first subtracted from the temporal parameter sequence, and then its MS was calculated. The length of discrete Fourier transform to calculate the MS was set to cover the maximum utterance length of the training data. These MS models were used in the synthesis module to reproduce the MS components, which were not well reproduced from the HSMMs only.

2.4. Synthesis module

2.4.1. Speech parameter generation

In the synthesis module, the context labels were first generated in the text processing module, and then the sentence HSMM corresponding to the text to be synthesized were constructed to generate the spectrum, continuous F_0, aperiodicity, and voiced/unvoiced regions. The spectral parameters were generated based on the speech parameter generation algorithm considering the MS components lower than 50 Hz [10]. The other parameters were generated based on the ML-based parameter generation [27]. Additionally, we applied the MS-based post-filter [9] to the generated continuous F_0 contour.[3] The MS was not considered in the aperiodicity component because there was no quality gain by the MS modification. An example of the generated mel-cepstrum sequences is illustrated in Fig. 3. We can find that more fluctuations are observed on the mel-cepstal sequence generated with the MS than that without the MS. Note that the global variance (GV) [28] is also recovered because the MS can also represent the GV.

2.4.2. Spectral smoothing in silence frames

After speech parameter generation, we further performed spectral parameter smoothing at silence frames. The averaged spectral parameter at a current silence region was calculated and the spectral parameters in the silence region were replaced with the averaged one.

[2]This stream setting is similar to the duplicated feature training [26] and the stream weights for continuous F_0 and discrete F_0 should be determined. We informally evaluated synthetic speech quality using some stream weight settings and chose this setting.

[3]No significant quality difference was observed between the continuous F_0 contour generated by speech parameter generation considering the MS and that filtered by the MS-based post-filter.

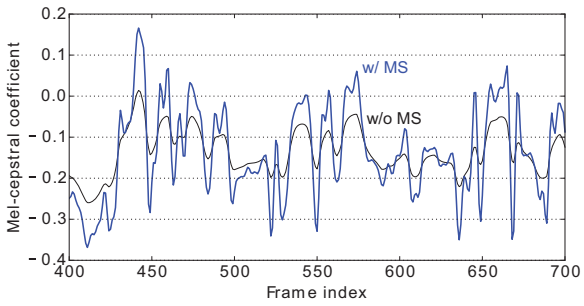

Figure 3: An example of the 20-th mel-cepstral coefficient sequence generated without considering the MS [27] ("w/o MS") and that with considering the MS ("w/ MS").

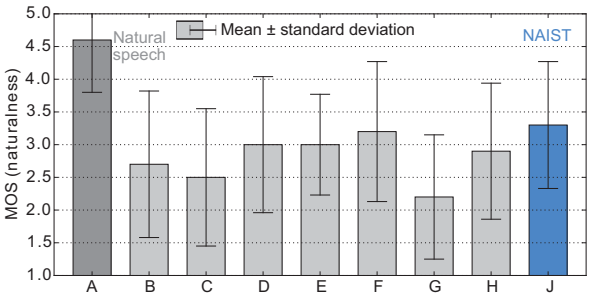

Figure 4: Result of MOS test on naturalness in Marathi.

3. Experimental results

To submit the NAIST TTS system to the Blizzard Challenge 2015, we synthesized 50 reading texts (RD) and 50 semantically unpredictable sentences (SUS) in each language. The following 3 subjective evaluations were conducted in the challenge: (1) a mean opinion score (MOS) test on naturalness, (2) a degradation MOS (DMOS) test on similarity to the original speaker, and (3) a manual dictation test on intelligibility to calculate the word error rate (WER).

Because of the limited space, we only show the results of the naturalness evaluation using RD sentences (Fig. 4), the similarity evaluation using RD sentences (Fig. 5), and the intelligibility evaluation (Fig. 6) in Marathi. Alphabets "A" and "J" indicate natural speech and our system, respectively. The other alphabets indicate the other participants' systems. We have found that our system was ranked in the highest group among the submitted systems in terms of naturalness in most of Indian languages but the gap between natural speech and synthetic speech was still large. Although our system was evaluated as the best in terms of intelligibility in Marathi (which was better than natural speech), such a result was not observed consistently over the other languages. Finally, our system was usually ranked in the middle group among the submitted systems in terms of similarity.

4. Summary

This paper has presented the NAIST TTS system for the Blizzard Challenge 2015. The pre-processing for smoothing parameter trajectories and the speech parameter generation considering the modulation spectrum have been implemented in our system. The results in the challenge have demonstrated that our system is capable of synthesizing naturally sounding speech.

Acknowledgements: Part of this work was supported by JSPS KAKENHI Grant Number 26280060, Grant-in-Aid for JSPS Fellows Grant Number 26·10354, and "JSPS Strategic Young Researcher Over-

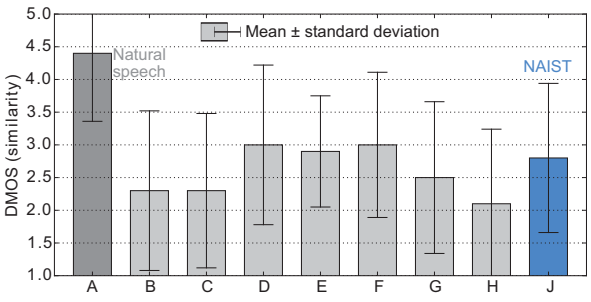

Figure 5: Result of MOS test on similarity to the original speaker in Marathi.

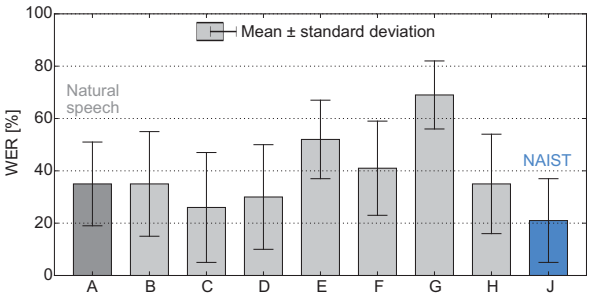

Figure 6: Result of intelligibility test in Marathi.

seas Visits Program for Accelerating Brain Circulation." The authors are grateful to Emeritus Prof. Hideki Kawahara of Wakayama Univ., Japan, for permission to use the STRAIGHT and Dr. Masanori Morise of Yamanashi Univ., Japan, for meaningful comments for the WORLD.

5. References

[1] A.W. Black, K. Tokuda, "The Blizzard Challenge - 2005: Evaluating corpus-based speech synthesis on common datasets," *Proc. INTERSPEECH*, pp. 77–80, Lisbon, Portugal, Sep. 2005.

[2] "Blizzard Challenge http://synsig.org/index.php/Blizzard_Challenge."

[3] H.A. Patil, T.B. Patel, N.J. Shah, H.B. Sailor, R. Krishnan, G.R. Kasthuri, T. Nagarajan, L. Christina, N. Kumar, V. Raghavendra, S.P. Kishore, S.R.M. Prasanna, N. Adiga, S.R. Singh, K. Anand, P. Kumar, B.C. Singh, S.L. Binil Kumar, T.G. Bhadran, T. Sajini, A. Saha, T. Basu, K.S. Rao, N.P. Narendra, A.K. Sao, R. Kumar, P. Talukdar, P. Acharyaa, S. Chandra, S. Lata, H.A. Murthy, "A syllable-based framework for unit selection synthesis in 13 Indian languages," *Proc. O-COCOSDA*, pp. 1–8, Gurgaon, India, Nov. 2013.

[4] S. Takamichi, T. Toda, Y. Shiga, S. Sakti, G. Neubig, and S. Nakamura, "Parameter generation methods with rich context models for high-quality and flexible text-to-speech synthesis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 2, pp. 239–250, 2014.

[5] T. Toda, "Augmented speech production based on real-time statistical voice conversion," *Proc. GlobalSIP*, pp. 755–759, Atlanta, U.S.A., Dec. 2014

[6] K. Kobayashi, T. Toda, H. Doi, T. Nakano, M. Goto, G. Neubig, S. Sakti, and S. Nakamura, "Voice timbre control based on perceived age in singing voice conversion," *IEICE Trans. on Inf. and Syst.*, vol. E97-D, no. 6, pp. 1419–1428, Jun. 2014.

[7] K. Tanaka, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "A hybrid approach to electrolaryngeal speech enhancement based on noise reduction and statistical excitation generation," *IEICE Trans. on Inf. and Syst.*, vol. E97-D, no. 6, pp. 1429–1437, Jun. 2014.

[8] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden Markov models," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234–1252, 2013.

[9] S. Takamichi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "A postfilter to modify modulation spectrum in HMM-based speech synthesis," in *Proc. ICASSP*, Florence, Italy, May 2014, pp. 290–294.

[10] S. Takamichi, T. Toda, A. W. Black, and S. Nakamura, "Parameter generation algorithm considering modulation spectrum for HMM-based speech synthesis," in *Proc. ICASSP*, Brisbane, Australia, Apr. 2015.

[11] K. Sawada, S. Takaki, K. Hashimoto, K. Oura, and K. Tokuda, "Overview of NITECH HMM-based text-to-speech system for Blizzard Challenge 2014," in *Proc. Blizzard Challenge 2014 Workshop*, Singapore, Sep. 2014.

[12] A. Suni, T. Raitio, D. Gowda, R. Karhila, M. Gibson, and O. Watts, "The Simple4All entry to the Blizzard Challenge 2014," in *Proc. Blizzard Challenge 2014 Workshop*, Singapore, Sep. 2014.

[13] "Festvox http://festvox.org/download.html."

[14] S.S. Nair, R.C. Rechitha, and C.S. Kumar, "Rule-based grapheme to phoneme converter for malayalam," *International Journal of Computational Linguistics and Natural Language Processing*, vol. 2, no. 7, pp. 417–420, Jul. 2013.

[15] "HMM-based speech synthesis system (HTS) http://hts.sp.nitech.ac.jp/."

[16] H. Kawahara, I. Masuda-Katsuse, and A.D. 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 Commun.*, vol. 27, no. 3–4, pp. 187–207, 1999.

[17] 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 *MAVEBA 2001*, Firentze, Italy, Sept. 2001, pp. 1–6.

[18] M. Morise, "CheapTrick, a spectral envelope estimator for high-quality speech synthesis," *Speech Commun.*, vol. 67, pp. 1–7, 2015.

[19] M. Morise, H. Kawahara, and H. Katayose, "Fast and reliable F0 estimation method based on the period extraction of vocal fold vibration of singing voice and speech," in *Proc. AES 35th International Conference*, London, U.K., Feb. 2009.

[20] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation," in *Proc. INTERSPEECH*, Pittsburgh, U.S.A., Sep. 2006, pp. 2266–2269.

[21] K. Yu and S. Young, "Continuous F0 modeling for HMM based statistical parametric speech synthesis," *IEEE Trans. Audio, Speech and Language*, vol. 19, no. 5, pp. 1071–1079, 2011.

[22] P. Taylor, *Text-To-Speech synthesis*. Cambridge Univ. Press, 2009.

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

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

[25] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," *J. Acoust. Soc. Jpn.(E)*, vol. 21, no. 2, pp. 79–86, 2000.

[26] S. Kang and H. Meng, "Statistical parametric speech synthesis using weighted multi-distribution deep belief network," in *Proc. INTERSPEECH*, Max Atria, Singapore, Sep. 2014, pp. 1959–1963.

[27] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP*, Istanbul, Turkey, June 2000, pp. 1315–1318.

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