An overview of nitech HMM-based speech synthesis system for Blizzard Challenge 2005

Conference Paper · January 2005

Source: DBLP

CITATIONS
98

READS
42

2 authors, including:

Heiga Zen
Google Inc.
118 PUBLICATIONS **7,402** CITATIONS

SEE PROFILE

An Overview of Nitech HMM-based Speech Synthesis System for Blizzard Challenge 2005

*Heiga Zen and Tomoki Toda**

Nagoya Institute of Technology, Gokiso-cho, Showa-ku, Nagoya, 466-8555 Japan
E-mail: zen@ics.nitech.ac.jp, tomoki@is.naist.jp

Abstract

In the present paper, hidden Markov model (HMM) based speech synthesis system developed in Nagoya Institute of Technology (Nitech-HTS) for a competition of text-to-speech synthesis systems using the same speech databases, named Blizzard Challenge 2005, is described. We show an overview of the basic HMM-based speech synthesis system and then recent developments to the latest one such as STRAIGHT-based vocoding, hidden semi-Markov model (HSMM) based acoustic modeling, and parameter generation considering global variance are illustrated. Constructed voices can synthesize speech around 0.3 xRT (real time ratio) and their footprints are less than 2 MB. The listening test results show that performances of our systems are much better than we expected.

1. Introduction

In recent years, we have developed a kind of corpus-based speech synthesis system based on hidden Markov models (HMMs) [1]. In this system, spectral and excitation parameters are extracted from speech database and modeled by context-dependent HMMs. In the synthesis part, spectral and excitation parameters are generated from HMMs themselves [2]. By filtering the excitation, a synthesis filter controlled by the spectral parameters outputs speech waveform. This system has the following features:

1) smooth and natural sounding speech can be synthesized,

2) the voice characteristics can be changed,

3) it is "trainable."

As for 1), by taking account of statistics of both static and dynamic features, the dynamics of the generated speech parameter sequence are constrained to be realistic. As for 2), by transforming HMM parameters appropriately, voice characteristics of synthesized speech can be changed since this system generates speech from the HMMs themselves. As for 3), this system can be automatically constructed.

In January 2005, Black and Tokuda conducted a competition of text-to-speech synthesis systems using the same speech databases, named Blizzard Challenge 2005. In the present paper, we describe technical details, building processes, and constructed voices of the latest HMM-based speech synthesis system developed in Nagoya Institute of Technology (Nitech-HTS) for the Blizzard Challenge 2005.

One of the limitations of the basic system is that synthesized speech is "buzzy" since it is based on a vocoding technique. To alleviate this problem, a high quality vocoder called STRAIGHT [3] is introduced. Other techniques such as hidden semi-Markov model based acoustic modeling [4] and parameter generation considering global variance [5], which have been proved to improve the basic system, are also integrated.

The rest of the present paper organized as follows. First a brief overview of the basic HMM-based speech synthesis system is given. This is followed by the new features integrated to the latest system and then by the details of our systems constructed for the Blizzard Challenge 2005. Finally concluding remarks are presented.

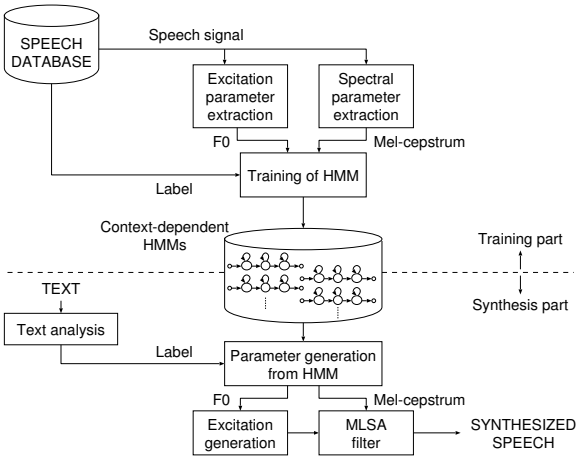

Figure 1: An overview of the basic HMM-based speech synthesis system.

2. Basic system

Figure 1 illustrates an overview of the basic HMM-based speech synthesis system [1]. In this system, feature vector consists of spectrum and F_0 parts. The spectrum part consists of mel-cepstral coefficients, their delta and delta-delta and F_0 part consists of $\log F_0$, its delta and delta-delta.

In the training part, feature vector sequences are modeled by context-dependent HMMs. Training procedure of the context-dependent HMMs is almost the same as that in speech recognition systems. The main differences are that not only phonetic contexts but also linguistic and prosodic ones are taken into account and state duration probabilities are explicitly modeled by single Gaussian distributions.

In the synthesis stage, first a given text to be synthesized is converted to a context-dependent label sequence and a sentence HMM is constructed by concatenating the context-dependent

*Presently, with the Graduate School of Information Science, Nara Institute of Science and Technology, Japan.

HMMs according to the label sequence. Secondly, state durations maximizing their probabilities were determined. Thirdly, mel-cepstral coefficients and $\log F_0$ sequences maximizing their output probabilities for a given state sequence are generated by speech parameter generation algorithm (case 1 in [2]). Finally, speech waveform is synthesized directly from the generated mel-cepstral coefficients and $\log F_0$ sequences using Mel Log Spectrum Approximation (MLSA) filter.

3. New features

3.1. STRAIGHT vocoding

As a high-quality speech vocoding method, we employ STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum), which is a vocoder type algorithm proposed by Kawahara et al. [3]. It consists of three main components, i.e., F_0 extraction, spectral and aperiodic analysis, and speech synthesis.

The STRAIGHT automatically extracts F_0 with fixed-point analysis [6]. We employ a two-stage extraction to alleviate errors of the F_0 extraction, e.g., halving and doubling. Firstly, we perform the F_0 extraction for all of training data for each speaker under the condition in which a range to search is set to 40-600 Hz. Taking account of a histogram of the extracted F_0, we roughly estimate an F_0 range of each speaker. Then, F_0 is again extracted in the speaker-specific range.

Using the extracted F_0, the STRAIGHT performs 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 0th through 39th mel-cepstral coefficients to which the smoothed spectrum analyzed by the STRAIGHT is converted with a recursive filter. An aperiodicity measure on the frequency domain based on a ratio between the lower and upper smoothed spectral envelopes to represent the relative energy distribution of aperiodic components [7] is also extracted. As a parameter for constructing mixed excitation source in speech synthesis, average values of the aperiodicity measures on five frequency bands, i.e., 0-1 kHz, 1-2 kHz, 2-4 kHz, 4-6 kHz and 6-8 kHz are used.

The STRAIGHT designs the mixed excitation as the weighted sum of a pulse train with phase manipulation and Gaussian noise. The weighting process is performed in the frequency domain using the aperiodic measures. It is required to convert the mel-cepstrum to the linear-scaled spectrum since the STRAIGHT employs an FFT-based processing for synthesizing a speech waveform. However it increases computational complexity. To reduce computational cost, we apply MLSA filter to the STRAIGHT synthesis. Specifically, we generate one-pitch waveforms from mel-cepstral coefficients and the mixed-excitation with the MLSA filter, and then a synthetic waveform is generated with PSOLA.

3.2. Hidden semi-Markov model

In the HMM-based speech synthesis system, rhythm and tempo are controlled by the state duration probabilities modeled by the 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 reestimated in the Baum-Welch iteration. In the synthesis stage, we construct a sentence HMM and determine the state durations maximizing their probabilities. Then a speech parameter vector sequence is generated. However, there is an inconsistency: although parameters of HMMs are reestimated without explicit state duration probability distributions, speech parameter vector sequence is generated from the HMMs with explicit state duration probability distributions. This inconsistency might degrade the quality of outputs.

To resolve the discrepancy, we have introduced hidden semi-Markov model, which can be viewed as the HMMs with the explicit state duration probability distributions, into training part of this system [4]. It makes possible to reestimate state output and duration probability distributions simultaneously. Improvements in not only duration but also spectrum and F_0 in the synthetic speech have been reported [4].

3.3. Parameter generation considering global variance

Usually, speech parameter vector sequences generated from the HMMs are smoothed excessively. Synthesized speech using over-smoothed parameters sounds muffled. To reduce this effect, we use a parameter generation algorithm considering global variance (GV) of the generated parameters [5].

We apply that algorithm to both spectral and F_0 parameter generation processes [5]. One GV is calculated from a parameter sequence over the entire of one utterance. Note that only voiced frames are used for calculating GV of F_0 parameters. Probability density on GV is modeled by a Gaussian distribution with a diagonal covariance matrix.

In the parameter generation, we firstly generate a parameter trajectory with the speech parameter generation algorithm (case 1 in [2]). Then, we convert the generated trajectory so that its GV is equal to a mean of the Gaussian distribution. Using the converted one 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.

4. Constructing voices for Blizzard Challenge 2005

4.1. Preparing training data

For the Blizzard Challenge 2005, we built 4 voices (speakers BDL, CLB, RMS, and SLT) using the CMU ARCTIC databases. They consist of 1132 phonetically balanced utterances for each speaker and includes speech waveforms recorded at 16 kHz, phoneme segmentations, utterance information files, and pitch marks in the Festvox style. Two databases (speakers SLT and BDL) were released in advance. We used them to explore the best setting of our system, e.g., the order of mel-cepstral analysis, training procedure, and GV weight. When remaining 2 databases (speakers CLB and RMS) were released, we used the system settings developed with the first 2 databases. Therefore, building processes of the later 2 voices were completely automatic.

To prepare training data, mel-cepstral coefficients, F_0, and aperiodicity parameters were extracted from the databases in the way described in Section 3.1. Feature vector consisted of spectrum, F_0 and aperiodicity measure parameter vectors: the spectrum parameter vector consisted of 40 mel-cepstral coefficients including the zeroth coefficient, their delta and delta-delta coefficients, the F_0 parameter vector consisted of $\log F_0$, its delta and delta-delta, and the aperiodicity measure parameter vector consisted of 5 average values of the aperiodicity measures and their delta and delta-delta.

Table 1: The number of distributions (leaf nodes) after decision-tree based context clustering.

	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 2: Voice building time (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

4.2. Acoustic model training

We used 5-state left-to-right HSMM structure. Each state output probability distribution was consisted of 5 streams: mel-cepstral coefficients with their delta and delta-delta, $\log F_0$, $\Delta \log F_0$, $\Delta^2 \log F_0$ and aperiodicity measures with their delta and delta-delta. The spectrum and aperiodicity streams were modeled by a single Gaussian distribution with diagonal covariance matrix. The $\log F_0$, $\Delta \log F_0$, $\Delta^2 \log F_0$ streams were modeled by a multi-space probability distribution (MSD) consisted of a single Gaussian distribution with diagonal covariance matrix (voiced space) and a single discrete distribution which outputs only one symbol (unvoiced space). Each state duration probability distribution was modeled by a single Gaussian distribution whose dimensionality was equal to the number of HSMM states.

A modified version of HMM-based speech synthesis software toolkit [8], which was developed as a patch code for HTK, was used to train acoustic models. After training monophone HSMMs, they were converted to context-dependent ones. In this work, contextual factors described in [9] were taken into account. They were extracted from the utterance information files included in the databases using feature extraction functions of the Festival speech synthesis system. The context-dependent HSMMs were reestimated (one iteration) and then a decision-tree based context clustering technique based on an Minimum Description Length (MDL) criterion was applied to distributions for spectrum, F_0, aperiodicity, and state duration. After reestimating clustered HSMMs (four iteration), parameter sharing structure were untied. Then untied HSMMs were reestimated again (one iteration). We applied the decision-tree based context clustering again. Table 1 shows the number of distributions after second context clustering. Re-clustered HSMMs were reestimated (five iteration), and then converted to input files for our speech synthesis engine[1].

Tables 2 and 3 show total building times[2] and footprints of constructed voices, respectively. Table 3 indicates that the footprints of constructed voices were less than 2 MB. The pdf files include the parameter values of state output and duration probability distributions saved in binary integer and single precision floating point number. The tree files contain the decision trees of spectrum, F_0, aperiodicity measure and duration saved in ASCII (HTK format). We can reduce their footprints without any quality degradation by eliminating redundant informations. Further reduction is also available with small quality degradation by vector quantization, saving pdf files in fixed point number instead of floating point one, or pruning the decision trees.

[1]This speech synthesis engine does not include text processing part.

[2]Training was run on a 3.2 GHz Pentium 4 machine.

Table 3: Footprints of constructed systems (KB).

	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

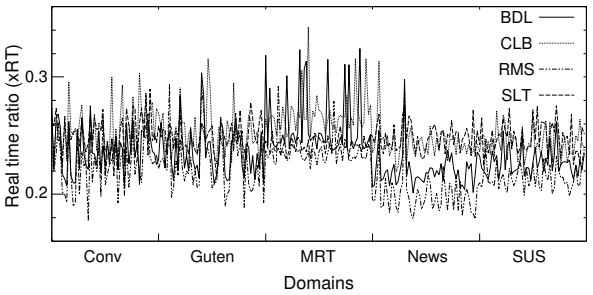

Figure 2: Real time factors for synthesizing speech.

4.3. Synthesizing speech

The test sentences provided by 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),
- Semantically unpredictable sentences (SUS).

For each sentence, we converted it into context-dependent label sequence using the Festival speech synthesis system. We did not provide any tags which specifies accents, stresses or pronunciations to help text analyzer, and outputs of text analyzer were not manually corrected. Then our engine synthesized a speech waveform according to given context-dependent label sequence.

Figure 2 illustrates the real-time ratio of each system to synthesize speech waveform for given label sequence on a 1.6 GHz Pentium 4 machine. This figure indicates that the constructed voices can synthesize speech around 3 times faster than the real time even on a little bit less state-of-the-art machine.

We had built 2 voices based on the basic system described in Section 2 using the databases of speakers BDL and SLT. Compared with the voices based on the basic system, quality of the voices based on the latest system were totally improved.

Figures 3 and 4 show average Mean Opinion Scores (MOSs) and average Word Error Rates (WERs) of natural speech, the Nitech-HTS and the best of other participants, respectively. The performance of our systems was much better

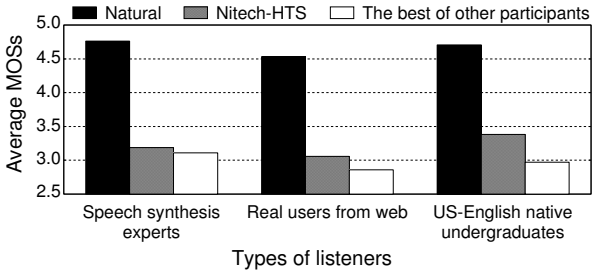

Figure 3: Average Mean Opinion Scores (MOSs) of the natural speech, the Nitech-HTS, and the best of other participants.

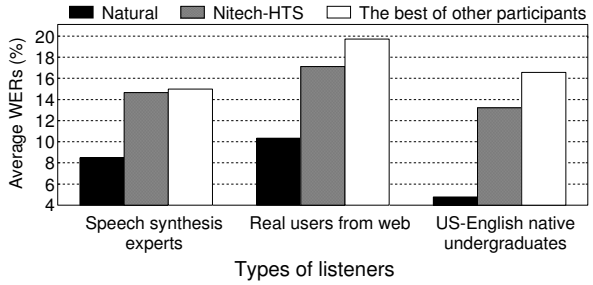

Figure 4: Average Word Error Rates (WERs) of the natural speech, the Nitech-HTS and the best of other participants.

than we expected, though it was still far from natural speech. These figures indicate that the Nitech-HTS achieved the highest MOSs and the lowest WERs in all types of listeners. Please see [10] for the detail.

In the unit selection approach, the generated speech has a high quality at waveform level, especially in limited domain speech synthesis because it concatenates speech waveforms directly. Although unit selection approach sometimes gives excellent results, it sometimes gives very bad ones too. On the other hand, in the HMM-based approach, it has a quality of "vocoded speech" but sounds smooth and stable. Furthermore, it has the advantages of being small and making it possible to change voice characteristics easily by applying a speaker adaptation technique used in speech recognition.

In this competition, relatively small databases (around 1 hour for each speaker) were used. For such amount of data, the HMM-based approach may be more appropriate than the unit selection one to build voices because it can potentially cover the given training data more effectively [11]. Hence, larger speech databases (e.g., more than 10 hours) are usually used in state-of-the-art unit selection systems. It would be worthy of exploring the size of speech database where unit selection approach overcomes HMM-based one.

Sometimes it is difficult to collect large speech database enough to build good unit selection system. For example, Black have indicated that recording emotional or emphasized speech consistently has been difficult [12]. For such case, the HMM-based approach might be very useful because it does not require large amount of training data and can reestimate new voices with only a few utterances from existing models trained by large data using speaker adaptation, speaker interpolation, or eigenvoice technique.

5. Conclusion

In the present paper, hidden Markov model (HMM) based speech synthesis system developed in the Nagoya Institute of Technology (Nitech-HTS) for a competition of text-to-speech synthesis systems using the same speech databases, named Blizzard Challenge 2005, was described. We showed an overview of the basic HMM-based speech synthesis system and then recent developments to the latest one such as STRAIGHT-based vocoding, hidden semi-Markov model (HSMM) based acoustic modeling, and parameter generation considering global variance were illustrated. Constructed voices could synthesize speech around 0.3 xRT (real time ratio) and their footprints were less than 2 MB. The listening test results showed that performances of our systems were much better than we expected.

6. Acknowledgments

Authors would like to thank Dr. Keiichi Tokuda of Nagoya Institute of Technology for helpful discussions. The core of this work originated with his pioneering ideas, which led us to new research idea. The authors are also grateful to Dr. Hideki Kawahara of Wakayama University for permission to use STRAIGHT vocoding method.

7. References

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

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

[3] 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 Communication*, vol. 27, pp. 187–207, 1999.

[4] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Hidden semi-Markov model based speech synthesis," in *Proc. of ICSLP*, 2004, pp. 1185–1180.

[5] T. Toda and K. Tokuda, "Speech parameter generation algorithm considering global variance for HMM-based speech synthesis," in *Proc. of Eurospeech*, 2005.

[6] 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," in *Proc. of Eurospeech*, 1999, pp. 2781–2784.

[7] 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*, 2001, pp. 13–15.

[8] 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/`.

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

[10] C. Bennett and A. Syrdal, "Large scale evaluation of corpus-based synthesizers: results and lessons from the 2005 blizzard challenge," 2005, submitted to Eurospeech.

[11] A. Black, "Perfect synthesis for all of the people all of the time," in *Proc. of IEEE Speech Synthesis Workshop*, 2002.

[12] ——, "Unit selection and emotional speech," in *Proc. of Eurospeech*, 2003, pp. 1649–1652.