The HMM-based Speech Synthesis System (HTS) Version 2.0

Heiga Zen[1], *Takashi Nose*[2], *Junichi Yamagishi*[23], *Shinji Sako*[14],
Takashi Masuko[2], *Alan W. Black*[5], *Keiichi Tokuda*[1]

[1]Nagoya Institute of Technology, [2]Tokyo Institute of Technology, [3]University of Edinburgh,
[4]Tokyo University, [5]Carnegie Mellon University

zen@sp.nitech.ac.jp, Takashi.Nose@ip.titech.ac.jp, jyamagis@inf.ed.ac.uk, sako@mmsp.nitech.ac.jp
awb@cs.cmu.edu, tokuda@nitech.ac.jp

Abstract

A statistical parametric speech synthesis system based on hidden Markov models (HMMs) has grown in popularity over the last few years. This system simultaneously models spectrum, excitation, and duration of speech using context-dependent HMMs and generates speech waveforms from the HMMs themselves. Since December 2002, we have publicly released an open-source software toolkit named HMM-based speech synthesis system (HTS) to provide a research and development platform for the speech synthesis community. In December 2006, HTS version 2.0 was released. This version includes a number of new features which are useful for both speech synthesis researchers and developers. This paper describes HTS version 2.0 in detail, as well as future release plans.

1. Introduction

Currently the most popular speech synthesis technique is unit selection [1–3], where appropriate sub-word units are selected from large speech databases. Over the last decade, this technique has been shown to synthesize high quality speech and is used for many applications. Although it is very hard to surpass the quality of the best examples of unit selection, it does have a limitation that the synthesized speech will strongly resemble the style of the speech recorded in the database. As we require speech which is more varied in voice characteristics, speaking styles, and emotions, we need to record larger and larger databases with these variations to achieve the synthesis we desire without degrading the quality [4]. However, recording such a large database is very difficult and costly [5].

Over the last few years, a statistical parametric speech synthesis system based on hidden Markov models (HMMs) has grown in popularity [6–10]. In this system, context-dependent HMMs are trained from databases of natural speech, and we can generate speech waveforms from the HMMs themselves. This system offers the ability to model different styles without requiring the recording of very large databases.

Figure 1 is an overview of this system. It consists of training and synthesis parts. The training part is similar to that used in speech recognition systems. The main difference is that both spectrum (mel-cepstral coefficients [11], and their dynamic features) and excitation (logarithmic fundamental frequencies ($\log F_0$) and its dynamic features) parameters are extracted from a speech database and modeled by context-dependent HMMs (phonetic, linguistic, and prosodic contexts are taken into account). To model variable dimensional parameter sequence such as $\log F_0$ with unvoiced regions properly, multi-space probability distributions (MSD) [12] are used. Each HMM has state duration probability density functions (PDFs) to capture the temporal structure of speech [13, 14]. As a result, the system models spectrum, excitation, and durations in a unified HMM framework [6]. The synthesis part does the inverse operation of speech recognition. First, an arbitrarily given text to be synthesized is converted to a context-dependent label sequence, and then an utterance HMM is constructed by concatenating the context-dependent HMMs according to the label sequence. Second, state durations of the utterance HMM are determined based on the state duration PDFs. Third, the speech parameter generation algorithm (typically, the Case 1 algorithm in [15] is used, please refer to Section 2.4 for detail) generates the sequence of spectral and excitation parameters that maximize their output probabilities. Finally, a speech waveform is synthesized directly from the generated spectral and excitation parameters using the corresponding speech synthesis filter (mel log spectrum approximation (MLSA) filter [16] for mel-cepstral coefficients).

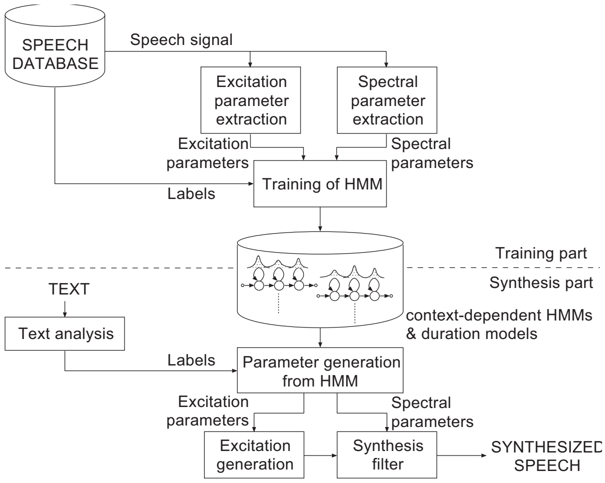

Figure 1: Overview of a typical HMM-based speech synthesis system.

The most attractive part of this system is that its voice characteristics, speaking styles, or emotions can easily be modified by transforming HMM parameters using various techniques such as adaptation [17, 18], interpolation [19, 20], eigenvoice [21], or multiple regression [22].

Since December 2002, we have publicly released an open-source software toolkit named HMM-based speech synthesis system (HTS) [23] to provide a research and development platform for speech synthesis community. Currently various organizations use it to conduct their own research projects, and we believe that it has contributed significantly to the success of HMM-based synthesis today. In December 2006, HTS version 2.0 was released. This version includes a number of new features which are useful for both speech synthesis researchers and developers. This paper describes relevant details of this system, and future release plans.

2. HTS: A toolkit for HMM-based speech synthesis system

2.1. Outline

The HMM-based speech synthesis system (HTS) has been being developed by the HTS working group as an extension of the HMM toolkit (HTK) [24]. The history of the main modifications we have made are listed below:

- Version 1.0 (December 2002)
  - Based on HTK-3.2.
  - Context clustering based on the minimum description length (MDL) criterion [25].
  - Stream-dependent context clustering [6].
  - Multi-space probability distributions (MSD) as state output PDFs [12].
  - State duration modeling and clustering [13].
  - Speech parameter generation algorithm (Case 1 in [15] only).
  - Demo using the CMU Communicator database.
- Version 1.1 (May 2003)
  - Based on HTK-3.2.
  - Small run-time synthesis engine.
  - Demo using the CSTR TIMIT database.
  - HTS voices for the Festival speech synthesis system [26].
- Version 1.1.1 (December 2003)
  - Based on HTK-3.2.1.
  - Demo using the CMU ARCTIC database [27].
  - Demo using the Nitech Japanese database.
  - Variance flooring for MSD-HMMs.
  - Post-filtering [28].
  - HTS voice for the Galatea toolkit [29].

The source code of HTS is released as a patch for HTK. Although the patch is released under a free software license similar to the MIT license, once the patch is applied users must obey the license of HTK.[1] Since version 1.1, a small run-time synthesis engine named `hts_engine` has been included. It works without the HTK libraries, hence it is free from the HTK license. Users can develop their own open or proprietary software based on the run-time synthesis engine. In fact, it has been integrated into ATR XIMERA [30] and Festival as an spectrum and prosody prediction modules and one of the speech synthesis modules, respectively. Although no text analyzers have been included, Festival (general) or the Galatea toolkit (Japanese) can be used. Of course users can use their own text analyzers. For example, Krstulovic et al. [31] used the text analysis provided by the MARY software [32] instead of the Festival. This toolkit has been used in various research groups to develop their own HMM-based speech synthesis systems [33–46].

There have been a variety of functional restrictions in HTS version 1.x releases. However, HTS version 2.0 has more flexibility and a number of new functions which we have proposed. The next section describes the detail of HTS version 2.0.

2.2. New features in version 2.0

After an interval of three years, HTS version 2.0 was released in December 2006. This is a major update and includes a number of new features and fixes, such as

- Based on HTK-3.4.
- Support GCC-4.
- Compilation without signal processing toolkit (SPTK).
- Terms about redistributions in binary form are added to the HTS license.
- `HCompV` (global mean and variance calculation tool) accumulates statistics in double precision. For large databases the previous version often suffered from numerical errors.
- `HRest` (Baum-Welch re-estimation tool for a single HMM) can generate state duration PDFs [13, 14] with the `-g` option.
- Phoneme boundaries can be given to `HERest` (embedded Baum-Welch re-estimation tool) using the `-e` option. This can reduce computational cost and improve phoneme segmentation accuracy [47]. We may also specify subset of boundaries (e.g, pause positions).
- Reduced-memory implementation of decision tree-based context clustering in `HHEd` (a tool for manipulating HMM definitions) with the `-r` option. For large databases the previous versions sometimes consumed huge memory.
- Each decision tree can have a name with regular expressions (`HHEd` with the `-p` option).
  e.g.,

  ```
  TB 000 {(*-a+*,*-i+*).state[2]}
  TB 000 {(*-t+*,*-d+*).state[3]}
  ```

  As a result, two different trees can be constructed for consonants and vowels respectively.
- Flexible model structures in `HMGenS` (speech parameter generation tool). In the previous versions, we assumed that the first HMM stream is mel-cepstral coefficients and the others are for $\log F_0$. Now we can specify model structures using the configuration variables `PDFSTRSIZE` and `PDFSTRORDER` Non-left-to-right model topologies (e.g., ergodic HMM), Gaussian mixtures, and full covariance matrices are also supported.
- Speech parameter generation algorithm based on the expectation-maximization (EM) algorithm (the Case 3 algorithm in [15], please refer to Section 2.4 for detail) in `HMGenS`. Users can select generation algorithms using the `-c` option.

[1] The HTK license prohibits redistribution and commercial use.

- Random generation algorithm [48] in HMGenS. Users can turn on this function by setting a configuration variable RNDPG=TRUE.
- State or phoneme-level alignments can be given to HMGenS.
- The interface of HMGenS has been switched from HHEd-style to HERest-style.
- Various kinds of linear transformations for MSD-HMMs in HERest.
  - Constrained and unconstrained maximum likelihood linear regression (MLLR) based adaptation [49].
  - Adaptive training based on constrained MLLR [49].
  - Precision matrix modeling based on semi-tied covariance matrices [50].
  - Heteroscedastic linear discriminant analysis (HLDA) based feature transform [51].
  - Phonetic decision trees can be used to define regression classes for adaptation [52,53].
  - Adapted HMMs can be converted to the run-time synthesis engine format.
- Maximum a posteriori (MAP) adaptation [54] for MSD-HMMs in HERest.
- Speed improvements in many parts.
- Many bug fixes.

The most significant new features are speaker adaptation for MSD-HMMs and the speech parameter generation algorithm based on the EM algorithm. In the following section, we describe these features in more detail.

2.3. Adaptation and adaptive training

As discussed in Section 1, one of the major advantages of the HMM-based speech synthesis approach over the unit-selection approach is its flexibility: we can easily modify its voice characteristics, speaking style, or emotions by transforming HMM parameters appropriately.

Speaker adaptation is the most successful example. By adapting HMMs with only a small number of utterances, we can synthesize speech with voice characteristics of a target speaker [17, 18]. MLLR and MAP-based speaker adaptation for single-stream HMMs have been supported since HTK-2.2. However, we could not support them in the official HTS releases because our internal implementation of adaptation for multi-stream MSD-HMMs was not portable. In HTK-3.4 alpha, most of adaptation-related parts in HTK were rewritten. This change made porting adaptation for multi-stream MSD-HMMs straightforward.

In HTS version 2.0, MLLR mean (MLLRMEAN), diagonal variance (MLLRVAR), full variance (MLLRCOV), and constrained mean and variance (CMLLR) adaptations for MSD-HMMs are implemented. Unfortunately, adaptation of state duration PDFs [55, 56] is not supported yet. MAP estimation for mixture weights, means, variances, and transition probabilities are also supported. In addition, HTS version 2.0 includes adaptive training (CMLLR) [49], semi-tied covariance matrices [50], and HLDA, which have recently been implemented in HTK. The use of adaptive training enables us to estimate better canonical models for speaker adaptation and improves the performance of the average voice-based speech synthesis system [57]. Recently semi-tied covariance models were applied to HMM-based speech synthesis and we have achieved some improvement over diagonal covariance models if it is used with the speech parameter generation algorithm considering global variance [58]. These efficient full covariance modeling methods (would) become essential when we want to model highly correlated features such as articulatory movements. The use of HLDA enables us to derive a linear projection that best decorrelates training data associated with each particular class [51]. Although HLDA may not be effective in speech synthesis, it would be beneficial in recognition tasks.

Usually, MLLR transforms are shared across similar Gaussian distributions clustered by a regression class tree [59]. However, this method has a disadvantage: we can adapt segment level features only [60]. This is because the regression class tree is constructed based on a distribution distance in a bottom-up fashion and does not reflect connections between distributions on the time axis. To address this problem, phonetic decision trees have been applied to define regression classes [52, 53]. This enables us to adapt both segmental and suprasegmental features, and in this way significant improvements over the regression class trees have been reported. In HTS version 2.0, HHEd has a command DT for converting phonetic decision trees into a regression class tree. Converted decision trees can be used as a regression class tree to estimate MLLR transforms.[2]

To use adaptation transforms in synthesis, we can use both HMGenS and hts_engine. HMGenS can load and apply adaptation transforms in the same way used in HERest. For hts_engine, first model sets are transformed by adaptation transforms using the AX command of HHEd. Then adapted model sets are converted into the hts_engine format using the CT and CM commands.[3]

2.4. Speech parameter generation algorithm based on the EM algorithm

In [15], three types of speech parameter generation algorithms are described. These algorithms aim to solve the following three problems

Case 1. Maximize $P(\boldsymbol{o} \mid \boldsymbol{q}, \boldsymbol{i}, \lambda)$ w.r.t. $\boldsymbol{o}$,

Case 2. Maximize $P(\boldsymbol{o}, \boldsymbol{q}, \boldsymbol{i} \mid \lambda)$ w.r.t. $\boldsymbol{q}$, $\boldsymbol{i}$, and $\boldsymbol{o}$,

Case 3. Maximize $P(\boldsymbol{o} \mid \lambda)$ w.r.t. $\boldsymbol{o}$,

under the constraints between static and dynamic features ($\boldsymbol{o} = \boldsymbol{W}\boldsymbol{c}$), where λ is an utterance HMM and corresponding state duration models, $\boldsymbol{o} = \left[\boldsymbol{o}_1^\top, \ldots, \boldsymbol{o}_T^\top\right]^\top$ is a speech parameter trajectory including both static and dynamic features, $\boldsymbol{c} = \left[\boldsymbol{c}_1^\top, \ldots, \boldsymbol{c}_T^\top\right]^\top$ is a static feature vector sequence, $\boldsymbol{W}$ is a window matrix to calculate dynamic features from static features, $\boldsymbol{q} = \{q_1, \ldots, q_T\}$ is a state sequence, $\boldsymbol{i} = \{i_1, \ldots, i_T\}$ is a mixture component sequence, and T is the number of frames. For Case 1, it is simply required to solve a set of linear equations. However, recursive search and EM algorithm-based iterative optimization are required for Cases 2 and 3 respectively.

In the previous versions, only the algorithm for Case 1 was implemented: state and mixture component sequences were as-

[2] In the speaker adaptation demo script released with HTS version 2.0, this function is turned off to reduce computational complexity.

[3] Covariance matrices of adapted model sets are approximated by their diagonal elements.

sumed to be provided. In HTS version 2.0, we have additionally implemented the algorithm for Case 3,[4] in which we assume that the state and mixture component sequences or a part of them are hidden. We can select the algorithm to be used using the `-c` option. If the `-c 0` option is specified, the Case 1 algorithm is used (both $\boldsymbol{q}$ and $\boldsymbol{i}$ are given). If `-c 1`, the Case 3 algorithm with a fixed state sequence is used ($\boldsymbol{q}$ is given but $\boldsymbol{i}$ is hidden). With the `-c 2` option, the Case 3 algorithm is used (both $\boldsymbol{q}$ and $\boldsymbol{i}$ are hidden). It should be noted that although the Case 1 algorithm cannot use Gaussian mixtures, it is much more computationally efficient than the Case 2 and Case 3 algorithms.

2.5. Demonstrations and documentation

HTS version 2.0 comes with two demo scripts for training speaker-dependent systems (English and Japanese) and a demo script for a speaker-adaptation system (English). The English demo scripts use the CMU ARCTIC databases and generate model files for Festival and `hts_engine`. The Japanese demo script uses the Nitech database and generates model files for the Galatea toolkit. These scripts demonstrate the training processes and the functions of HTS. We recommend that users first try to run these demos and read the scripts themselves. Six voices for Festival trained by the CMU ARCTIC databases have also been released. Each HTS voice consists of model files trained by the demo script, and can be used as a voice for Festival without any other HTS tools.

Currently no documentation for HTS is available. However, the interface and functions of HTS are almost the same as those of HTK. Therefore, users who are familiar with HTK can easily understand how to use HTS. The manual of HTK [24] is also very useful. Most of questions we have been asked have their answers in this manual. There is an open mailing list for the discussion of HTS (`hts-users@sp.nitech.ac.jp`). If you have any questions or trouble with HTS, please first search the mailing list archive and read the HTK manual, and then ask on the mailing list.

3. Other applications

Although HTS has been developed to provide a research platform for HMM-based speech synthesis, it has also been used in various other ways, such as

- Human motion synthesis [61–63],
- Face animation synthesis [64],
- Audio-visual synthesis and recognition [65–67],
- Acoustic-articulatory inversion mapping [68],
- Prosodic event recognition [69,70],
- Very low-bitrate speech coder [71],
- Acoustic model adaptation for coded speech [72],
- Training data generation for ASR systems to obtain domain-specific acoustic models [73].
- Automatic evaluation of ASR systems [74].
- Online handwriting recognition [75].

We hope that HTS will contribute to progress in other research fields as well as speech synthesis.

4. Conclusions and future release plans

This paper described the details of the HMM-based speech synthesis system (HTS) version 2.0. This version includes a number of new features and fixes such as adaptation, adaptive training, and the speech parameter generation algorithm based on the EM algorithm.

Internally, we have developed a number of variants of HTS, e.g.,

- Hidden semi-Markov models (HSMMs) [76].
- Speech parameter generation algorithm considering global variance [58].
- Variational Bayes [77].
- Trajectory HMMs [78].
- Interpolation [19,20].
- Shared tree construction [79].
- Advanced adaptation and adaptive training [80,81].
- Eigenvoice [21].
- Multiple linear regression HMMs [22].

Some of these have been applied to our Blizzard Challenge systems and achieved successful results [7]. Hopefully, we can integrate valuable features of these variants into future HTS releases. The current plan for future releases is as follows:

- Version 2.0.1 (August 2007)
  - Bug fixes.
  - C/C++ API for `hts_engine`.
  - Speaker interpolation.
- Version 2.1 (March 2008)
  - HSMM training and adaptation.
  - Speech parameter generation algorithm considering global variance.
  - Advanced adaptation.

HTS version 2.1, with the STRAIGHT analysis/synthesis technique [82], will provide the ability to construct the state-of-the-art HMM-based speech synthesis systems developed for the past Blizzard Challenge events [7,83].

5. Acknowledgments

The authors are grateful to Dr. Korin Richmond of the university of Edinburgh for his helpful comments on earlier drafts of this paper. We also thank people who contributed to the development of HTS version 2.0. This work was partly supported by the MEXT e-Society project, the Japan Society for the Promotion of Science (No.18800019), and the Hori information science promotion foundation.

6. References

[1] A.W. Black and P. Taylor, "CHATR: a generic speech synthesis system," in *Proc. COLING94*, 1994.

[2] A. Hunt and A.W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. ICASSP*, 1996, pp. 373–376.

[3] R.E. Donovan and P.C. Woodland, "Automatic speech synthesizer parameter estimation using HMMs," in *Proc. ICASSP*, 1995, pp. 640–643.

[4] Only `HMGenS` provides algorithm for Case 3.

[4] E. Eide, A. Aaron, R. Bakis, W. Hamza, M. Picheny, and J. Pitrelli, "A corpus-based approach to <AHEM/> expressive speech synthesis," in *Proc. ISCA SSW5*, 2004.

[5] A.W. Black, "Unit selection and emotional speech," in *Proc. Eurospeech*, 2003, pp. 1649–1652.

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

[7] H. Zen, T. Toda, M. Nakamura, and K. Tokuda, "Details of Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 1, pp. 325–333, Jan. 2007.

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

[9] A.W. Black, H. Zen, and K. Tokuda, "Statistical parametric speech synthesis," in *Proc. ICASSP*, 2007, pp. 1229–1232.

[10] J. Yu, M. Zhang, J. Tao, and X. Wang, "A novel HMM-based TTS system using both continuous HMMs and discrete HMMs," in *Proc. ICASSP*, 2007, pp. 709–712.

[11] T. Fukada, K. Tokuda, Kobayashi T., and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. ICASSP*, 1992, pp. 137–140.

[12] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling," in *Proc. ICASSP*, 1999, pp. 229–232.

[13] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Duration modeling for HMM-based speech synthesis," in *Proc. ICSLP*, 1998, pp. 29–32.

[14] H. Zen, T. Masuko, T. Yoshimura, K. Tokuda, T. Kobayashi, and T. Kitamura, "State duration modeling for HMM-based speech synthesis," *IEICE Trans. on Inf. & Syst.*, vol. E90-D, no. 3, pp. 692–693, 2007.

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

[16] S. Imai, "Cepstral analysis synthesis on the mel frequency scale," in *Proc. ICASSP*, 1983, pp. 93–96.

[17] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Voice characteristics conversion for HMM-based speech synthesis system," in *Proc. ICASSP*, 1997, pp. 1611–1614.

[18] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR," in *Proc. ICASSP*, 2001, pp. 805–808.

[19] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Speaker interpolation in HMM-based speech synthesis system," in *Proc. Eurospeech*, 1997, pp. 2523–2526.

[20] M. Tachibana, J. Yamagishi, T. Masuko, and T. Kobayashi, "Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing," *IEICE Trans. Inf. & Syst.*, vol. E88-D, no. 11, pp. 2484–2491, 2005.

[21] K. Shichiri, A. Sawabe, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Eigenvoices for HMM-based speech synthesis," in *Proc. ICSLP*, 2002, pp. 1269–1272.

[22] T. Nose, J. Yamagishi, and T. Kobayashi, "A style control technique for speech synthesis using multiple regression HSMM," in *Proc. Interspeech*, 2006, pp. 1324–1327.

[23] K. Tokuda, H. Zen, J. Yamagishi, T. Masuko, S. Sako, A.W. Black, and T. Nose, "The HMM-based speech synthesis system (HTS)," `http://hts.sp.nitech.ac.jp/`

[24] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X.-Y. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The Hidden Markov Model Toolkit (HTK) version 3.4*, 2006, `http://htk.eng.cam.ac.uk/`

[25] K. Shinoda and T. Watanabe, "Acoustic modeling based on the MDL criterion for speech recognition," in *Proc. Eurospeech*, 1997, pp. 99–102.

[26] A.W. Black, P. Taylor, and R. Caley, "The festival speech synthesis system," `http://www.festvox.org/festival/`

[27] J. Kominek and A.W. Black, "CMU ARCTIC databases for speech synthesis," Tech. Rep. CMU-LTI-03-177, Carnegie Mellon University, 2003.

[28] 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.

[29] Galatea Project, "Galatea – An open-source toolkit for anthropomorphic spoken dialogue agent," `http://hil.t.u-tokyo.ac.jp/galatea/`

[30] H. Kawai, T. Toda, J. Yamagishi, T. Hirai, J. Ni, T. Nishizawa, M. Tsuzaki, and K. Tokuda, "XIMERA: A concatenative speech synthesis system with large scale corpora," *IEICE Trans. Inf. & Syst. (Japanese Edition)*, vol. J89-D, no. 12, pp. 2688–2698, Dec. 2006.

[31] S. Krstulovic, A. Hunecke, and M. Schroeder, "An HMM-based speech synthesis system applied to German and its adaptation to a limited set of expressive football announcements," in *Proc. of Interspeech*, 2007.

[32] M. Schröder and J. Trouvain, "The German text-to-speech synthesis system MARY: A tool for research, development and teaching," *International Journal of Speech Technology*, vol. 6, pp. 365–377, 2003.

[33] Y.-J. Wu and R.H. Wang, "HMM-based trainable speech synthesis for Chinese," *Journal of Chinese Information Processing*, vol. 20, no. 4, pp. 75–81, 2006.

[34] Y. Qian, F. Soong, Y. Chen, and M. Chu, "An HMM-based Mandarin Chinese text-to-speech system," in *Proc. of ISCSLP*, 2006.

[35] S.-J. Kim, J.-J. Kim, and M.-S. Hahn, "Implementation and evaluation of an HMM-based Korean speech synthesis system," *IEICE Trans. Inf. & Syst.*, vol. E89-D, pp. 1116–1119, 2006.

[36] C. Weiss, R. Maia, K. Tokuda, and W. Hess, "Low resource HMM-based speech synthesis applied to German," in *ESSP*, 2005.

[37] M. Barros, R. Maia, K. Tokuda, D. Freitas, and F. Resende Jr., "HMM-based European Portuguese speech synthesis," in *Interspeech*, 2005, pp. 2581–2584.

[38] A. Lundgren, *An HMM-based text-to-speech system applied to Swedish*, Master thesis, Royal Institute of Technology (KTH), 2005.

[39] T. Ojala, *Auditory quality evaluation of present Finnish text-to-speech systems*, Master thesis, Helsinki University of Technology, 2006.

[40] M. Vainio, A. Suni, and P. Sirjola, "Developing a Finnish concept-to-speech system," in *2nd Baltic conference on HLT*, 2005, pp. 201–206.

[41] B. Vesnicer and F. Mihelic, "Evaluation of the Slovenian HMM-based speech synthesis system," in *TSD*, 2004, pp. 513–520.

[42] S. Martincic-Ipsic and I. Ipsic, "Croatian HMM-based speech synthesis," *Journal of Computing and Information Technology*, vol. 14, no. 4, pp. 307–313, 2006.

[43] O. Abdel-Hamid, S. Abdou, and M. Rashwan, "Improving Arabic HMM based speech synthesis quality," in *Interspeech*, 2006, pp. 1332–1335.

[44] J. Latorre, K. Iwano, and S. Furui, "Polyglot synthesis using a mixture of monolingual corpora," in *ICASSP*, 2005, vol. 1, pp. 1–4.

[45] X. Gonzalvo, I. Iriondo, J. Socor, F. Alas, and C. Monzo, "HMM-based Spanish speech synthesis using CBR as F0 estimator," in *ITRW on NOLISP*, 2007.

[46] S. Chomphan and T. Kobayashi, "Implementation and evaluation of an HMM-based Thai speech synthesis system," in *Proc. of Interspeech*, 2007.

[47] D. Huggins-Daines and A. Rudnicky, "A constrained Baum-Welch algorithm for improved phoneme segmentation and efficient training," in *Proc. of Interspeech*, 2006, pp. 1205–1208.

[48] K. Tokuda, H. Zen, and T. Kitamura, "Reformulating the HMM as a trajectory model," in *Proc. Beyond HMM – Workshop on statistical modeling approach for speech recognition*, 2004.

[49] M.J.F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech & Language*, vol. 12, no. 2, pp. 75–98, 1998.

[50] M.J.F. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 272–281, 1999.

[51] M.J.F. Gales, "Maximum likelihood multiple projection schemes for hidden Markov models," *IEEE Trans. Speech & Audio Process.*, vol. 10, no. 2, pp. 37–47, 2002.

[52] J. Yamagishi, M. Tachibana, T. Masuko, and T. Kobayashi, "Speaking style adaptation using context clustering decision tree for HMM-based speech synthesis," in *Proc. ICASSP*, 2004, pp. 5–8.

[53] M. Tachibana, J. Yamagishi, T. Masuko, and T. Kobayashi, "A style adaptation technique for speech synthesis using HSMM and suprasegmental features," *IEICE Trans. Inf. & Syst.*, vol. E89-D, no. 3, pp. 1092–1099, 2006.

[54] J.L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. on Speech & Audio Process.*, vol. 2, no. 2, pp. 291–298, 1994.

[55] J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 2, pp. 533–543, 2007.

[56] J. Yamagishi, K. Ogata, Y. Nakano, J. Isogai, and T. Kobayashi, "HSMM-based model adaptation algorithms for average-voice-based speech synthesis," in *Proc. ICASSP*, 2006, pp. 77–80.

[57] J. Yamagishi, *Average-Voice-Based Speech Synthesis*, Ph.D. thesis, Tokyo Institute of Technology, 2006.

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

[59] M.J.F. Gales, "The generation and use of regression class trees for MLLR adaptation," Tech. Rep. CUED/F-INFENG/TR263, Cambridge University, 1996.

[60] M. Ostendorf and I. Bulyko, "The impact of speech recognition on speech synthesis," in *Proc. the IEEE Workshop on Speech Synthesis*, 2002, CD-ROM proceeding.

[61] K. Mori, Y. Nankaku, C. Miyajima, K. Tokuda, and T. Kitamura, "Motion generation for Japanese finger language based on hidden Markov models," in *Proc. FIT*, 2005, vol. 3, pp. 569–570, (in Japanese).

[62] N. Niwase, J. Yamagishi, and T. Kobayashi, "Human walking motion synthesis with desired pace and stride length based on HSMM," *IEICE Trans. Inf. & Syst.*, vol. E88-D, no. 11, pp. 2492–2499, 2005.

[63] G. Hofer, H. Shimodaira, and J. Yamagishi, "Speech driven head motion synthesis based on a trajectory model," in *Proc. SIGGRAPH*, 2007, (submitted).

[64] O. Govokhina, G. Bailly, G. Breton, and P. Bagshaw, "TDA: a new trainable trajectory formation system for facial animation," in *Proc. Interspeech*, 2006, pp. 1274–1247.

[65] M. Tamura, S. Kondo, T. Masuko, and T. Kobayashi, "Text-to-audio-visual speech synthesis based on parameter generation from HMM," in *Proc. Eurospeech*, 1999, pp. 959–962.

[66] S. Sako, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "HMM-based text-to-audio-visual speech synthesis," in *Proc. ICSLP*, 2000, pp. 25–28.

[67] T. Ishikawa, Y. Sawada, H. Zen, Y. Nankaku, C. Miyajima, K. Tokuda, and T. Kitamura, "Audio-visual large vocabulary continuous speech recognition based on early integration," in *Proc. FIT*, 2002, pp. 203–204, (in Japanese).

[68] K. Richmond, "A trajectory mixture density network for the acoustic-articulatory inversion mapping," in *Proc. of Interspeech*, 2006, pp. 577–580.

[69] K. Emoto, H. Zen, K. Tokuda, and T. Kitamura, "Accent type recognition for automatic prosodic labeling," in *Proc. Autumn Meeting of ASJ*, 2003, vol. I, pp. 225–226, (in Japanese).

[70] H.-L. Wang, Y. Qian, F.K. Soong, J.-L. Zhou, and J.-Q. Han, "A multi-space distribution (MSD) approach to speech recognition of tonal languages," in *Proc. of Interspeech*, 2006, pp. 125–128.

[71] T. Hoshiya, S. Sako, H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Improving the performance of HMM-based very low bitrate speech coding," in *Proc. ICASSP*, 2003, vol. 1, pp. 800–803.

[72] K. Tanaka, S. Kuroiwa, S. Tsuge, and F. Ren, "An acoustic model adaptation using HMM-based speech synthesis," in *Proc. NLPKE*, 2003, vol. 1, pp. 368–373.

[73] M. Ishihara, C. Miyajima, N. Kitaoka, K. Itou, and K. Takeda, "An approach for training acoustic models based on the vocabulary of the target speech recognition task," in *Proc. Spring Meeting of ASJ*, 2007, pp. 153–154, (in Japanese).

[74] R. Terashima, T. Yoshimura, T. Wakita, K. Tokuda, and T. Kitamura, "An evaluation method of ASR performance by HMM-based speech synthesis," in *Proc. Spring Meeting of ASJ*, 2003, pp. 159–160, (in Japanese).

[75] L. Ma, Y.-J. Wu, P. Liu, and F. Soong, "A MSD-HMM approach to pen trajectory modeling for online handwriting recognition," in *Proc. ICDAR*, 2007.

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

[77] Y. Nankaku, H. Zen, K. Tokuda, T. Kitamura, and T. Masuko, "A Bayesian approach to HMM-based speech synthesis," in *Tech. rep. of IEICE*, 2003, vol. 103, pp. 19–24, (in Japanese).

[78] H. Zen, K. Tokuda, and T. Kitamura, "Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences," *Computer Speech & Language*, vol. 21, no. 1, pp. 153–173, 2006.

[79] J. Yamagishi, M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "A context clustering technique for average voice models," *IEICE Trans. Inf. & Syst.*, vol. E86-D, no. 3, pp. 534–542, 2003.

[80] J. Isogai, J. Yamagishi, and T. Kobayashi, "Model adaptation and adaptive training using ESAT algorithm for HMM-based speech synthesis," in *Proc. Interspeech*, 2005, pp. 2597–2600.

[81] Y. Nakano, M. Tachibana, J. Yamagishi, and T. Kobayashi, "Constrained structural maximum a posteriori linear regression for average-voice-based speech synthesis," in *Proc. Interspeech*, 2006, pp. 2286–2289.

[82] H. Kawahara, I. Masuda-Katsuse, and A.de 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.

[83] H. Zen, T. Toda, and K. Tokuda, "The Nitech-NAIST HMM-based speech synthesis system for the Blizzard Challenge 2006," in *Blizzard Challenge Workshop*, 2006.