Overview of NIT HMM-based speech synthesis system for Blizzard Challenge 2012

Shinji Takaki, Kei Sawada, Kei Hashimoto, Keiichiro Oura, and Keiichi Tokuda

Department of Scientific and Engineering Simulation, Nagoya Institute of Technology, Nagoya, JAPAN

Abstract

This paper describes a hidden Markov model (HMM) based speech synthesis system developed for the Blizzard Challenge 2012. In the Blizzard Challenge 2012, we focused on a design of contexts for using audio books as training data and duration modeling of silence between sentences for synthesizing paragraphs. It is well known that contextual factors affect speech. We use extended contexts for using audio books to construct appropriate model parameter tying structures. In addition, duration models of silence between sentences are created to synthesize more natural speech because connections between sentences are important for synthesizing paragraphs. Subjective evaluation results show that the system synthesized the high intelligible speech.

Index Terms: speech synthesis, hidden Markov model, context clustering

1. Introduction

A statistical parametric speech synthesis system based on hidden Markov models (HMMs) was recently developed. In HMM-based speech synthesis, the spectrum, excitation, and duration of speech are simultaneously modeled by HMMs, and speech parameter sequences are generated from the HMMs themselves [1]. Compared to other synthesis methods, this method has several advantages, 1) under its statistical training framework, it can learn statistical properties of speakers, speaking styles [2], emotions [3], etc, from the speech corpus; 2) many techniques developed for HMM-based speech recognition can be applied to speech synthesis [4, 5]; 3) voice characteristics of synthesized speech can be easily controlled by modifying acoustic statistics of HMMs [6, 7].

It is well known that contextual factors affect speech. Therefore, context-dependent acoustic models [8, 9] are widely used in HMM-based speech synthesis. Although a large number of context-dependent acoustic models can capture variations in speech data, too many model parameters lead to the over-fitting problem. Consequently, maintaining a good balance between model complexity and the amount of training data is very important for obtaining high generalization performance. The decision tree based context clustering [10] is an efficient method for dealing with the problem of data sparseness, for both estimating robust model parameters of context-dependent acoustic models and obtaining predictive distributions of unseen contexts. In HMM-based speech synthesis, the minimum description length (MDL) criterion is widely used as the criterion for model selection [11], and the context clustering was separately applied to distributions of the spectrum, F_0, aperiodicity measures, and state duration.

In this context clustering, questions about contexts are prepared beforehand and the model parameter tying structures are constructed by using these questions. Thus, the constructed state tying structures are strongly affected by questions about contexts. In conventional HMM-based speech synthesis, only contextual factors considering sentences are used for constructing the tying structures because speech data segmented into sentences is used. However, questions about only the such contexts would not be able to construct appropriate tying structures in using consecutive utterances such as audio books as training data. Furthermore, various reading styles are included in audio books because speaker read emphatically, emotionally, and so on. Contextual factors about such various reading styles would affect acoustic features. We use extended contexts for audio books and construct more appropriate tying structures to synthesize more natural speech.

In addition, duration models of silence for consecutive utterances such as a paragraph are created. Conventional HMM-based speech synthesis systems assume only synthesis for a sentence. However, synthesis of consecutive utterances is necessary for synthesizing expressive speech. Although consecutive utterances as training data is necessary for synthesizing paragraphs, because speech data segmented into sentences is usually used. Duration modeling of silence between sentences is one of the problems for synthesizing paragraphs. Silence durations between sentences are not modeled in conventional HMM-based speech synthesis system, though silence durations of the begininng and end of a utterance are modeled. Thus, appropriate silent length between sentences can not be estimated in synthesizing paragraphs and synthesized speech would be unnatural. To solve this problem, duration models of silence between sentences are created.

The rest of this paper is organized as follows. Section 2 describes our base speech synthesis system. Section 3 and 4 introduce new features of our system for the Blizzard Challenge 2012. Subjective listening test results are presented in Section 5. Concluding remarks and future work are presented in the final section.

2. Base system

2.1. HMM-based speech synthesis system

Figure 1 overviews a HMM-based speech synthesis system. It consists of training and synthesis parts.

The training part is similar to that used in speech recognition. The main difference is that both spectrum (e.g., mel-cepstral coefficients and their dynamic features) and excitation (e.g., log F_0 and its dynamic features) parameters are extracted from a speech database and modeled by HMMs. Although the spectrum part can be modeled by continuous HMM, the F_0 part cannot be modeled by continuous or discrete HMM be-

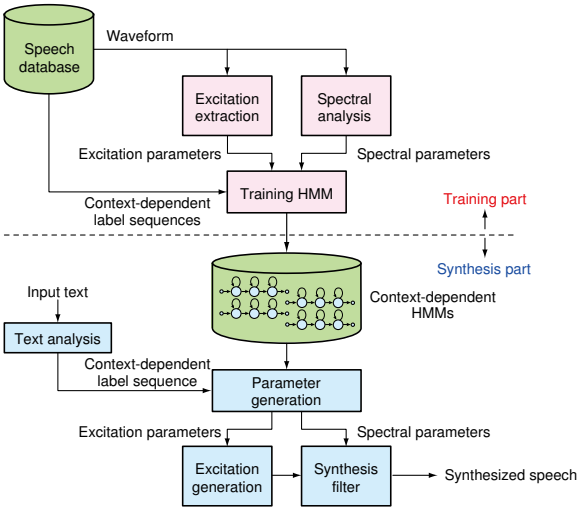

Figure 1: Overview of HMM-based speech synthesis system.

cause the observation sequence of F_0 is composed of a one-dimensional continuous value and discrete symbol which represents unvoiced. To model such observation sequence, multi-space probability distributions (MSDs) [12] are used for state-output distributions.

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 a sentence HMM is constructed by concatenating the context-dependent HMMs according to the label sequence. Second, state durations of the sentence HMM are determined based on the state-duration distributions. Third, the speech parameter generation algorithm generates sequences of spectral and excitation parameters that maximize their output probabilities under the constraints between static and dynamic features [13]. Finally, a speech waveform is synthesized directly from the generated spectral and excitation parameters using a speech synthesis filter. The most attractive part of this system is that voice characteristics, speaking styles, or emotions can easily be modified by transforming HMM parameters using various techniques, such as adaptation [5], interpolation [14], or eigenvoices [15].

2.2. Hidden semi-Markov model

In HMM-based speech synthesis, rhythm and tempo are controlled by state duration probability distributions. One of major limitations of HMMs is that they do not provide an adequate representation of the temporal structure of speech. This is because state duration probabilities decrease exponentially with time. To overcome this problem, the hidden semi-Markov model (HSMM) based speech synthesis framework [4] was used in our system. This framework introduces an HSMM, which is an HMM with explicit state duration probability distributions, into not only the synthesis part but also the training part of the HMM-based speech synthesis system. It makes possible to estimate state output and duration probability distributions simultaneously. The effectiveness of the HSMM-based approach has been reported in [4].

2.3. STRAIGHT vocoding

As a high-quality speech vocoding method, we use STRAIGHT, which is a vocoder type algorithm proposed by Kawahara *et al.* [16]. It consists of three main components; F_0 extraction, spectral and aperiodic analysis, and speech synthesis.

It is well known that extracting F_0 is difficult task because F_0 includes various errors. In this paper, F_0 used for acoustic features is created by filtering F_0 extracted from a number of F_0 extractor with a median filter. Since estimating voice or unvoice is necessary for extracting F_0, voice or unvoice is determined by the number of voice and unvoice candidates in a filter and only voice candidates are filtered with median filter to determine F_0 values in voice regions. We used 3 F_0 extractor STRAIGHT [16], RAPT [17], and SWIPE [18]. Filter length is 5 frame of current frame and 2 front and back frames in total.

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 domain to remove signal periodicity. As a spectral parameter, we use the 0th through 49th mel-cepstral coefficients to which the smoothed spectrum analyzed by the STRAIGHT is converted. An aperiodicity measure in the frequency domain [19] is also extracted.

2.4. Parameter generation algorithm considering global variance

The HMM-based speech synthesis method generates speech parameters from the HMMs directly, so that an output probability of the parameter is maximized under a constraint on an explicit relationship between static and dynamic features. Consequently, a smoothed parameter trajectory is generated but it is excessively smoothed due to the statistical processing. Therefore, the synthesized speech using over-smoothed parameters sounds muffled. To reduce this effect, we applied a parameter generation algorithm considering global variance (GV) of the generated parameters [20] 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. The probability density on GV is modeled using a Gaussian distribution with a diagonal covariance matrix. In parameter generation, first a parameter trajectory is generated with the speech parameter generation algorithm. Then, the generated trajectory is converted, so that its GV is equal to a mean of the Gaussian distribution. Using this converted trajectory as an initial value, the parameter trajectory is calculated iteratively to maximize a likelihood function with the Newton-Raphson method. This likelihood function consists of the output probability of the parameter sequence and that of its GV.

In order to improve the estimation accuracy of GV models, we use the GV features calculated from only speech region excluding silence and pause regions and estimate the context-dependent GV models instead of a single global GV model. The silence and pause regions are determined by the automatic phone aligner using HSMMs [21] included in the latest HTS.

The context-dependent GV models are tied by the decision-tree based context clustering method in a similar way to acoustic model parameter tying. The number of leaf nodes of the decision trees is automatically determined by the minimum description length (MDL) criterion [11]. In this paper, to simplify the implementation, only sentence-level contextual features (e.g., number of phonemes in a sentence) were used.

3. Design of contexts for audio books

Speech parameters such as spectrum, excitation, and duration depend on a variety of contextual factors such as phoneme identities, accent, parts-of-speech, etc. In the HMM-based speech synthesis system, context dependent models are generally used to capture these contextual factors. If combinations of these contextual factors are taken into account, we can obtain more accurate models. However, as the number of contextual factors increases, the number of possible combinations also increases exponentially. It is difficult to robustly estimate model parameters due to the lack of the training data. Furthermore, it is impossible to cover every possible combinations of contextual factors for a finite set of the training data. Various parameter tying techniques have been proposed to prevent this problem. A decision tree based context clustering technique has been widely used [11]. In this technique, top-down clustering is performed to maximize the likelihood of model parameters with respect to the training data by using questions about contexts. Then, parameters of all states belonging to the same leaf node are tied. Unseen models can be generated by traversing the decision trees.

In this context clustering, questions about contexts are prepared beforehand and the model parameter tying structures are constructed by using these questions. Thus, the constructed state tying structures are strongly affected by questions about contexts. In conventional HMM-based speech synthesis, only context factors considering sentences are used to construct tying structures. Table 1 shows the conventional context factors considering sentences.

Table 1: Contextual factors extracted from a sentence

the {phoneme before the previous, previous, current, next, phoneme after the next} phoneme identity
position of the current {phoneme identity / syllable / word} in the current {syllable / word, phrase / phrase}
whether the {previous, current, next} syllable {stressed, accented}
the number of phonemes in the {previous, current, next} syllable
the number of {stressed, accented} syllables {before, after} the current syllable in the current phrase
the number of syllables from the previous {stressed, accented} syllable to the current syllable
the number of syllables from the current syllable to the next {stressed, accented} syllable
name of the vowel of the current syllable
guess part-of-speech of the {previous, current, next} word
the number of syllables in the {previous, current, next} word
the number of content words {before, after} the current word in the current phrase
the number of words from the {previous content / current} word to the {current / next content} word
the number of {syllables, words} in the {previous, current, next} phrase
position of the current phrase in this utterance
TOBI endtone of the current phrase
the number of {syllables, words, phrases} in this utterance

Table 2: Additional contextual factors

book
the number of chapter
position of the paragraph in the chapter
the number of sentences between the double quotes
position of the sentence between the double quotes

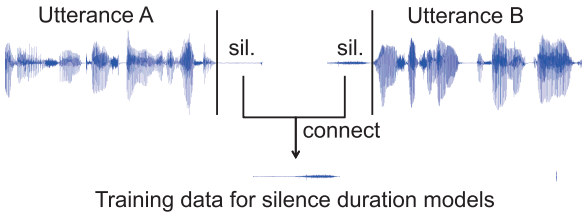

Figure 2: How to create training data for silence duration models.

However, questions about only the such contexts would not be able to construct appropriate tying structure in using audio books as training data. Furthermore, various reading styles are included in audio books because speaker read emphatically, emotionally, and so on. Hence, we use extended contexts to construct appropriate model parameter tying structures for audio books as training data . Table 2 shows additional context factors for audio books. There is a big amount of training data in audio books , so acoustic features would be affected by context factors about books, chapters, and paragraphs even if speaker is the same. Furthermore, reading styles are affected by double quotes because double quotes mean lines, emphasis, characters, and so on. Since speeches in the double quotes are different from normal speech, the appropriate tying structures would be constructed with context factors about double quotes.

In synthesis phase, there are unclear context factors about books, chapters, and paragraphs, so “no” is applied about questions for such contexts.

4. Duration models of silence between sentences

Although consecutive utterances such as audio books as training data is necessary for synthesizing paragraphs, modeling of such consecutive utterances is difficult because speech data is segmented into sentences in HMM-based speech synthesis. Duration modeling of silence between sentences is one of the problems for synthesizing paragraphs. Silence durations between sentences are not modeled in conventional HMM-based speech synthesis system, though silence durations of the begininng and end of a utterance are modeled.

In this work, duration models of silence between sentences are created using silence data connected the beginning and end silence of a utterance front and back. Figure 2 shows that how to create training data for duration models of silence between sentences. The alignment information of silence regions is required for creating training data for duration models of silence. The silence regions are determined by the HSMM-based automatic phoneme aligner [21] included in the latest HTS. In training phase, HSMMs are trained using created silence data and duration models of this HSMMs are used as duration models of silence between sentences.

Figure 3 shows speech synthesis procedure using duration models of silence between sentences. The procedure for synthesis with duration models of silence between sentences is de-

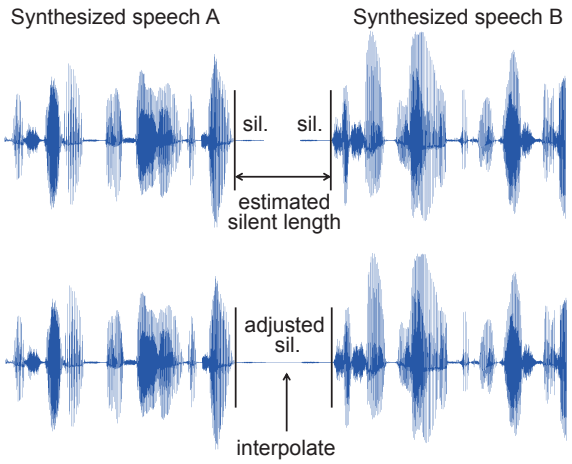

Figure 3: Speech synthesis procedure using estimated silent length with silence duration models.

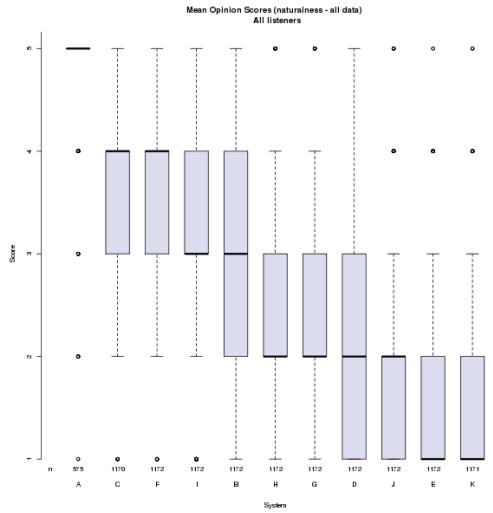

Figure 4: Results of MOS on naturalness of sentences.

fined as follows. 1) Each sentence of a paragraph is synthesized with conventional acoustic models. 2) Silence length of the beginning and end of synthesized speech is compared with estimated silence length with duration models of silence between sentences. 3) When silence length of synthesized speech is shorter than estimated silence length with silence duration models, silence between sentences is adjusted by interpolating silence. Otherwize two synthesized speech are connected without adjustment.

Duration models of silence between sentences would estimates appropriate silence length and synthesized speech would become more natural by interpolating silence.

5. Blizzard Challenge 2012 evaluation

We used 11,441 utterances, which were selected according to the alignment likelihood from "A Tramp Abroad", "The Adventures of Tom Sawyer", and "The Man that Corrupted Hadleyburg". Speech signals were sampled at a 44.1 kHz rate and windowed by an F_0-adaptive Gaussian window with a 5 ms shift. Feature vectors comprised 303-dimensions: 49-dimension STRAIGHT [16] mel-cepstral coefficients (plus the zero-th coefficient), log F_0, 49 band-filtered aperiodicity measures, and their dynamic and acceleration coefficients. We used 5-state left-to-right context-dependent multi-stream MSD-HSMMs [4, 12] without skip transitions as acoustic models. Each state output probability distribution was composed of spectrum, F_0, and aperiodicity streams. The spectrum and aperiodicity stream was modeled by single multi-variate Gaussian distributions with diagonal covariance matrices. The F_0 stream was modeled by a multi-space probability distribution consisting of a Gaussian distribution for voiced frames and a discrete distribution for unvoiced frames. Each state duration distribution was modeled by a one-dimensional Gaussian distribution.

In order to improve the estimation accuracy of GV models, we used the GV features calculated from only speech region excluding silence and pause regions and estimated the context-dependent GV models instead of a single global GV model. The decision tree-based context clustering technique was also applied to the context-dependent GV models. The decision tree was automatically selected by the MDL criterion. In this system, only sentence-level contextual features (e.g., number of phonemes in a sentence) were used.

5.1. Experimental results

To evaluate naturalness and similarity of sentences, 5-point mean opinion score (MOS) tests were conducted. The scale for the naturalness was 5 for "completely natural" and 1 for "completely unnatural". The scale for the similarity was 5 for "sounds like exactly the same person" and 1 for "sounds like a totally different person" compared to a few natural example sentences from the reference speaker. To evaluate naturalness of paragraphs, 60-point MOS tests were conducted (for example "bad"=10 and "excellent"=50).

Figure 4 and 7 shows the evaluation results on naturalness and similarity of sentences respectively. Figure 5 and 6 shows the evaluation results on naturalness and naturalness of speech pauses of paragraphs respectively. Figure 8 shows the evaluation results on intelligibility.

In these figure, "A", "B", and "H" correspond as follows.

- A: Natural speech.
- B: A Festival benchmark system. This system is a standard unit-selection voice built using the same method as used in the CSTR entry to Blizzard 2007.
- H: The 2012 NIT HMM-based speech synthesis system.

The results of listening tests showed that our system "H" was worse than the benchmark unit-selection system "B" in naturalness of speech and speaker similarity. However, in terms of speech pauses of paragraphs our system "H" outperformed the benchmark unit-selection system "B".

In terms of intelligibility, our system "H" outperformed the benchmark unit-selection system "B". [22] also showed that a HMM-based speech synthesis system was significantly more intelligible than a unit-selection based speech synthesis system.

These results indicate that our system "H" can generate the high intelligible speech although the naturalness and speaker similarity do not reach high enough levels. It seems that the buzziness of speech cause these results. Therefore, we need to improve the excitation model and feature extraction.

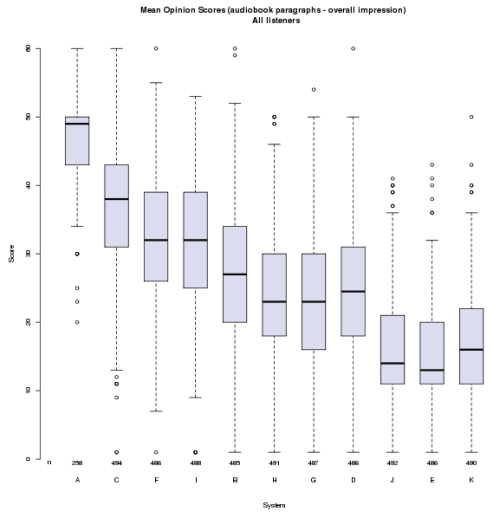

Figure 5: Results of MOS on naturalness of paragraphs.

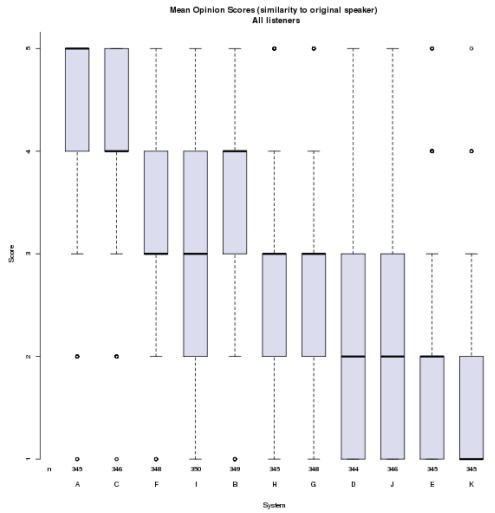

Figure 7: Results of MOS on speaker similarity.

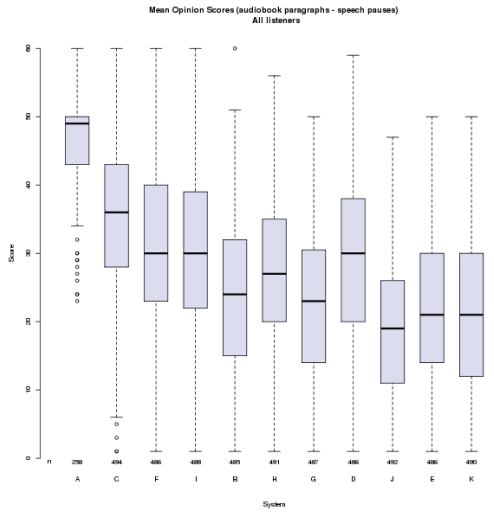

Figure 6: Results of MOS on speech pauses of paragraphs.

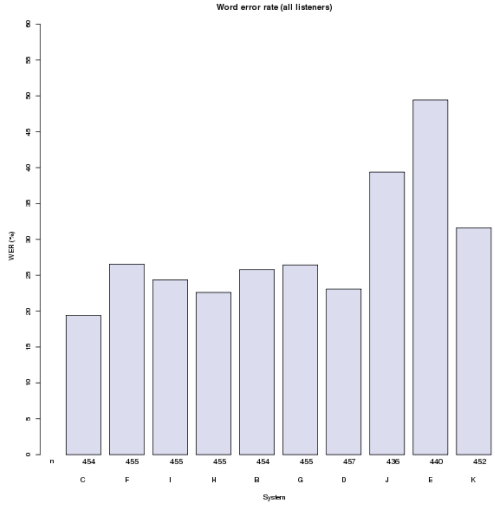

Figure 8: Results of WER.

6. Conclusions

We described HMM-based speech synthesis system developed at the Nagoya Institute of Technology (NIT) for the Blizzard Challenge 2012. we used extented contexts for using audio books as training data to construct appropriate model parameter tying structures. In addition, duration models of silence between sentences for estimating appropriate silence length between sentences were created. The results of listening tests showed that that our system was worse than the benchmark unit-selection system in naturalness of speech and speaker similarity. However, in terms of speech pauses of paragraphs our system "H" outperformed the benchmark unit-selection system "B".

In terms of intelligibility, our system competed with natural speech and outperformed the benchmark unit-selection system although there was no significant difference. These results indicate that our system can generate the high intelligible speech although the naturalness and speaker similarity do not reach high enough levels.

7. Acknowledgements

The research leading to these results was partly funded by the Core Research for Evolutional Science and Technology (CREST) from Japan Science and Technology Agency (JST).

8. 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," *Proceedings of Eurospeech 1999*, pp. 2347–2350, 1999.

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

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

[4] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura,

"Hidden semi-Markov model based speech synthesis," *Proceedings of ICSLP*, pp. 1185–1180, 2004.

[5] J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training," *IEICE Transactions on Information & Systems*, vol. E90-D, no. 2, pp. 533–543, 2007.

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

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

[8] K. Lee, "Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 38, no. 4, pp. 599–609, 1990.

[9] J. Odell, "The use of context in large vocabulary speech recognition," *PhD dissertation, Cambridge University*, 1995.

[10] S. Young, J. Odell, and P. Woodland, "Tree-based state tying for high accuracy acoustic modelling," *Proceedings of ARPA Workshop on Human Language Technology*, pp. 307–312, 1994.

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

[12] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Multi-space probability distribution HMM," *IEICE Transactions on Information & Systems*, vol. E85-D, no. 3, pp. 455–464, 2002.

[13] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," *Proceedings of ICASSP 2000*, pp. 936–939, 2000.

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

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

[16] 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.

[17] D. Talkin, W. kleijn, and K. Paliwal, "A robust algorithm for pitch tracking," *Speech Coding and Synthesis, Elsevier Science*, pp. 495–518, 1995.

[18] A. Camacho and J. Harris, "A sawtooth waveform inspired pitch estimator for speeh and music," *J. Acoust. SOC. Am*, vol. 124, no. 3, pp. 1638–1652, 2008.

[19] 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," *Proceedings of MAVEBA*, pp. 13–15, 2001.

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

[21] K. Oura, H. Zen, N. Yoshihiko, A. Lee, and K. Tokuda, "A fully consistent hidden semi-Markov model-based speech recognition system," *IEICE Transactions on Information & Systems*, vol. E91-D, no. 11, pp. 2693–2700, 2008.

[22] M. Wolters, K. Isaac, and S. Renals, "Evaluating speech synthesis intelligibility using amazon mechanical turk," *Proceedings of SSW7*, pp. 136–141, 2010.