Overview of NITECH HMM-based speech synthesis system for Blizzard Challenge 2013

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 2013. In the Blizzard Challenge 2013, audiobooks are provided as training data. In this paper, we focus on a construction of databases for training acoustic models from audiobooks. An automatic alignment technique based on speech recognition is used for obtaining pairs of audio and transcriptions. We also focus on training high natural and neutral acoustic models from audiobooks. Audiobooks consist of speech with various qualities, styles, emotions, etc. It is necessary to appropriately handle such data for training high quality acoustic models. We pruned unneutral and mistakable speech data from the aligned data with multiple techniques and trained acoustic models normalized differences of speaking styles, recording conditions, and file formats among chapters with adaptive training for each chapter. Subjective evaluation results show that the developed system synthesized the high natural and intelligible speech.

Index Terms: speech synthesis, hidden Markov model, audiobook, data pruning, adaptive training

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].

In the Blizzard Challenge 2013, audiobooks are provided as training data. The audiobooks consist of chapters, i.e., the speech data is not segmented into sentences, and there are mismatches between texts and speech data. An automatic alignment technique for such data is required because it is difficult to segment a large amount of data and fix mismatches with manpower. The accuracy of the alignment considerably impacts acoustic models trained for synthesizing speech. Therefore, this is the very important problem and under discussions. Techniques to handle the large speech corpora such as audiobooks for speech synthesis have been proposed [8, 9, 10]. In this paper, the lightly supervised technique is used for the alignment because there are helpful texts corresponding to audio in audiobooks. Using this technique, the pairs of transcriptions and audio are obtained.

Audiobooks consist of speech with various qualities, styles, emotions, etc. It is necessary to appropriately handle such data for training high quality acoustic models. In this paper, unneutral and mistakable speech data is pruned using multiple techniques based on confidence measures (WER), text features, speech features and phoneme confidence scores. The high natural and neutral acoustic models would be trained from the pruned training data. Adaptive training for each chapter also be applied for normalizing the remaining differences of speaking styles, recording conditions, and file formats among chapters after the data pruning processing. More stable acoustic models would be trained by chapter adaptive training.

The rest of this paper is organized as follows. Section 2 describes our base speech synthesis system. Section 3 and 4 introduce the techniques for the alignment of audiobooks and data handling for high natural and neutral acoustic models, respectively. 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 f0 and its dynamic fea-

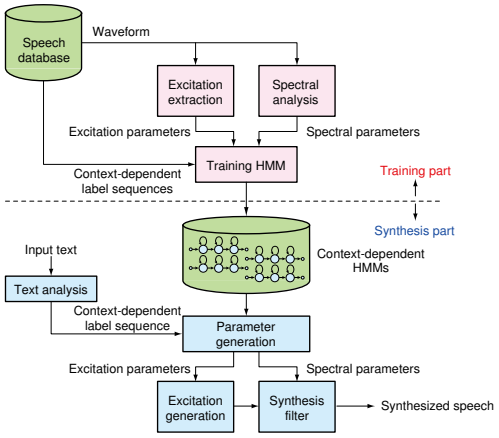

Figure 1: Overview of HMM-based speech synthesis system.

tures) parameters are extracted from a speech database and modeled by HMMs. In our system, the hidden semi-Markov model (HSMM) based speech synthesis framework [4] was used. It makes possible to estimate state output and duration probability distributions simultaneously. Although the spectrum part can be modeled by continuous HMM, the f0 part cannot be modeled by continuous or discrete HMM because the observation sequence of f0 is composed of a one-dimensional continuous value and discrete symbol which represents unvoiced. To model such observation sequence, multi-space probability distributions (MSDs) [11] 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 [12]. Finally, a speech waveform is synthesized directly from the generated spectral and excitation parameters using a speech synthesis filter.

2.2. STRAIGHT vocoding

As a high-quality speech vocoding method, we use STRAIGHT, which is a vocoder type algorithm proposed by Kawahara *et al.* [13]. It consists of three main components; f0 extraction, spectral and aperiodic analysis, and speech synthesis. Using the extracted f0, 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.

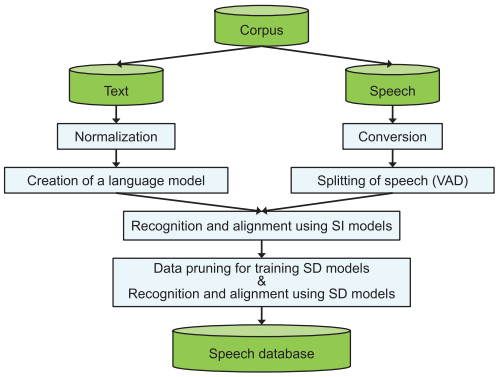

Figure 2: Overview of constructing a speech database from audiobooks.

2.3. Parameter generation algorithm considering global variance

We applied a parameter generation algorithm considering global variance (GV) of the generated parameters [14] to both spectral and f0 parameter generation processes. 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 context-dependent GV models are tied by the decision-tree based context clustering method in a similar way to acoustic model parameter tying.

3. Alignment of audiobooks

This section shows a technique of constructing a database for training acoustic models from audiobooks. Figure 2 overviews procedures of constructing a database. In this figure, a corpus consists of audio and texts segmented into chapters in audiobooks. The goal of this technique is to obtain speech data segmented into sentences and corresponding transcriptions.

In the procedure, text and speech processing is firstly performed. In the text processing, texts are normalized and a language model is created using normalized text. We used the Festival speech synthesis system [15] for the text normalization and SRILM [16] for the language model creation. The audiobook language model using normalized texts is interpolated with the general language model, and the ratio for interpolation of the audiobook and general language models is nine to one. Since texts in audiobooks include many unknown words, pronunciations of such words estimated by the Festival speech synthesis system [15] are added into the dictionary. In the speech processing, speech is down-sampled to 16 kHz rate that is used in many speech recognition systems. Then, the converted speech is split into sentences

by voice activity detection (VAD). We used SHoUT [17] for VAD.

In the next procedure, the recognition of the split audio is performed using the created language model and speaker independent (SI) models. TIMIT, WSJ0 and WSJ1 data sets are used for training SI models. The acoustic feature vector consists of 39 components comprised of 12-dimension mel-frequency cepstrul coefficients (MFCCs) including the 0th order coefficient with first and second order derivatives. Trained GMMs have 32 mixtures for silence and 16 mixtures for the others. We used HDecode (HTK version 3.4.1) for the recognition. After the recognition, a normalized audiobook text is aligned with a word sequence obtained by connecting all recognition results for each chapter. Transcriptions and alignments for each sentence are obtained from the alignment result. In this procedure, the word error rate (WER) is calculated as the confidence measure for each sentence.

After the recognition using SI models, the recognition is performed using speaker dependent (SD) models trained from the sentences obtained by the above step. SI models would be inappropriate for the recognition of audiobooks, and the quality of recognition would be improved using SD models. In this procedure, sentences for training SD models are pruned using the confidence measure calculated with SI models. SD models are trained by the same way as SI model training from the sentences whose WER is 0%. The recognition and the alignment for all data are re-preformed with SD models as the SI model case. The confidence measures using SD models can also be calculated for each sentence. Thus, a speech database for speech synthesis is constructed.

4. Data handling for training high natural and neutral acoustic models

Speech included in audiobooks is various in terms of qualities, speaking styles, emotions, etc. It is necessary to appropriately handle such data for training high quality acoustic models.

4.1. Data pruning

Data pruning techniques are used for training high natural and neutral models. The confidence measure (WER), text features, speech features and the phoneme confidence scores are used for the data pruning processing.

The data pruning processing based on the confidence measure is performed in sentence and chapter levels. Low confidence sentences, whose WER calculated with the SD models is 0%, are pruned in sentence level pruning. Moreover, chapter level data is pruned using the confidence measure. The ratio of pruned sentences in a chapter would be higher if speech data of the chapter is lower quality. In this paper, a chapter is pruned when the ratio of pruned sentences is more than 20%.

Table 1: Number of pruning sentences. *DQ*, *F0*, *Power* and *Conf* mean double quotes, f0 features, power features, and phoneme confidence scores, respectively. *OR* mean logical add of four pruning techniques.

	DQ	*f0*	*Power*	*Conf*	*OR*
EH1	18,679	4,083	5,689	2,952	24,969
EH2	2913	332	475	218	3413

Table 2: Numbers of sentences before and after pruning.

	Before	After
EH1	112,387	87,418
EH2	9,734	6,321

After the data pruning processing based on the confidence measure, text feature, speech feature and phoneme confidence score pruning are performed in sentence level. In the text feature pruning, sentences that include double quotes in the text are pruned because speaking styles of such sentences would be unneutral. In the speech feature pruning, maximum, mean and variance of f0 and power are used as features for each sentence and we determined the threshold for each features. Phoneme confidence scores also be used for the data pruning. The phoneme confidence scores are calculated as follows: 1) Monophone HSMMs are trained using the training data before pruning. 2) The phone regions are determined by the automatic phone aligner using HSMMs included in the latest HTS. 3) The phoneme confidence scores for all phonemes including incorrect phonemes are calculated in each obtained region. These phoneme confidence scores were used for detecting mismatches between texts and pronunciations. We pruned the sentences which the number of phone regions that the confidence score of the assigned phoneme is small, i.e., the confidence score of the assigned phoneme is included in the top five highest score, is more than 20%. Table 1 and 2 show the number of sentences pruned by each technique and the numbers of sentences before and after pruning.

4.2. Chapter adaptive training

Differences of speaking styles, recording conditions, and file formats would remain after the data pruning processing. Adaptive training based on MLLR [18] is applied for normalizing such differences. Although speakers are used as chunks in conventional techniques for adaptive training, books, chapters, paragraphs and sentences can be used as chunks for normalizing such differences. We used chapters as chunks for adaptive training based on MLLR in view of the amount of training data for each chunk. More stable average acoustic models trained by adaptive training were used for speech synthesis.

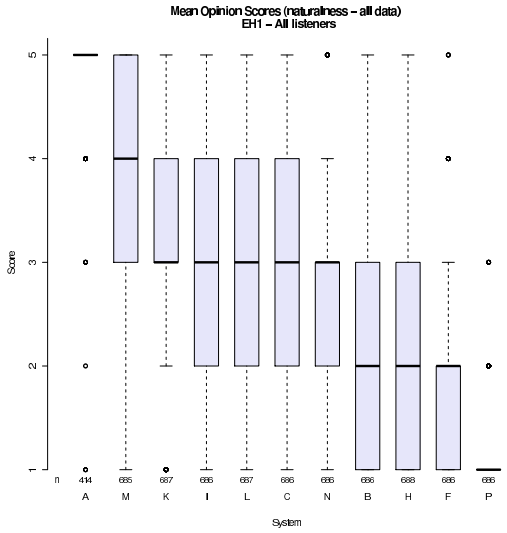

Figure 3: Results of MOS on naturalness (EH1).

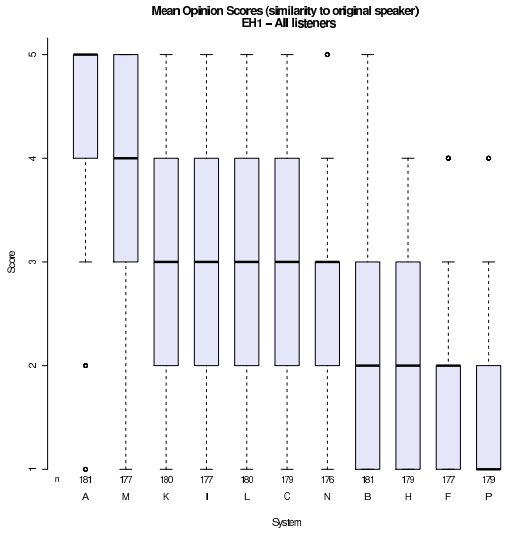

Figure 4: Results of MOS on speaker similarity (EH1).

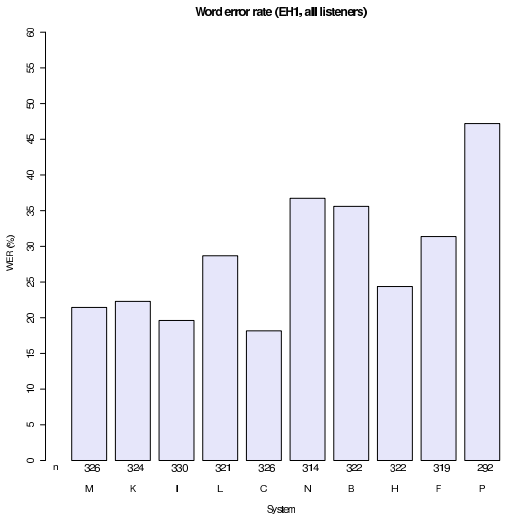

Figure 5: Results of WER (EH1).

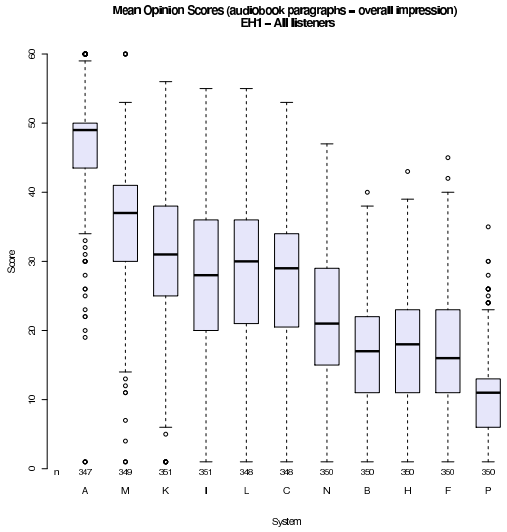

Figure 6: Results of MOS for audiobook paragraphs on overall impression (EH1).

5. Blizzard Challenge 2013 evaluation

5.1. Experimental conditions

We used 87,418 utterances (VAD results) of 664 chapters for EH1 task and 6,321 utterances of 97 chapters for EH2 task after the data pruning processing. Since we used the provided data, which was segmented into utterances, in EH2 task, we did not use the alignment technique for audiobooks. Speech signals were sampled at a 44.1 kHz rate and windowed by an f0-adaptive Gaussian window with a 5 ms shift. Feature vectors comprised 228-dimensions: 49-dimension STRAIGHT [13] mel-cepstral coefficients (plus the zero-th coefficient), log f0, 24-dimension mel-cepstral analysis aperiodicity measures, and their dynamic and acceleration coefficients. We used 5-state left-to-right context-dependent multi-stream MSD-HSMMs [4, 11] without skip transitions as acoustic models. Each state output probability distribution was composed of spectrum, f0, and aperiodicity streams. The spectrum and aperiodicity streams were modeled by single multivariate Gaussian distributions with diagonal covariance matrices. The f0 stream was modeled by a multi-space probability distribution consisting of a Gaussian distribution for voiced frames and a discrete distribution for unvoiced frames. State durations were modeled by a one-dimensional Gaussian distribution.

5.2. Experimental results

To evaluate naturalness and similarity, 5-point mean opinion score (MOS) tests were conducted. The scale for the naturalness was 5 for “completely natural” and

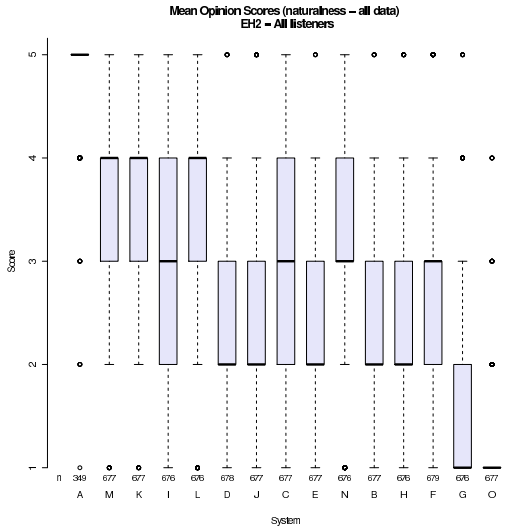

Figure 7: Results of MOS on naturalness (EH2).

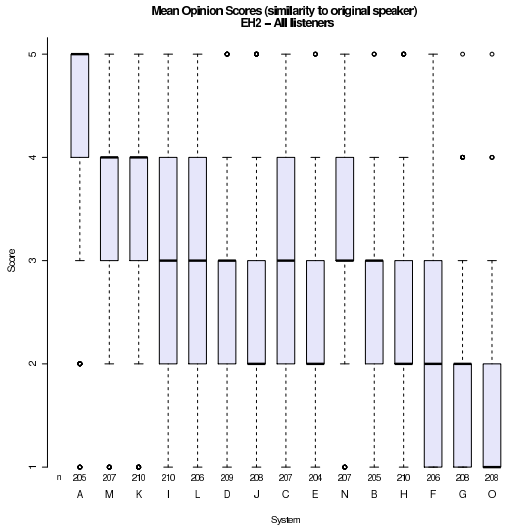

Figure 8: Results of MOS on speaker similarity (EH2).

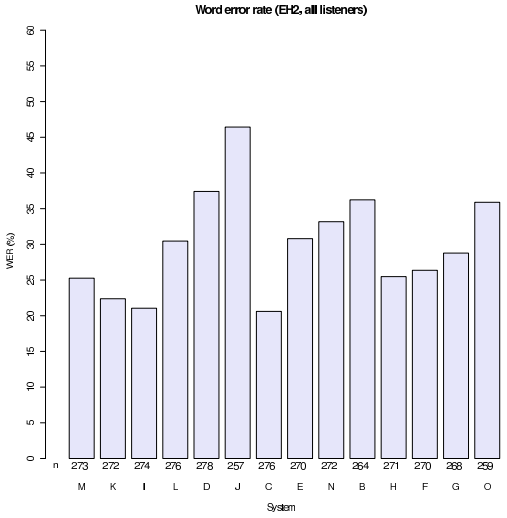

Figure 9: Results of WER (EH2).

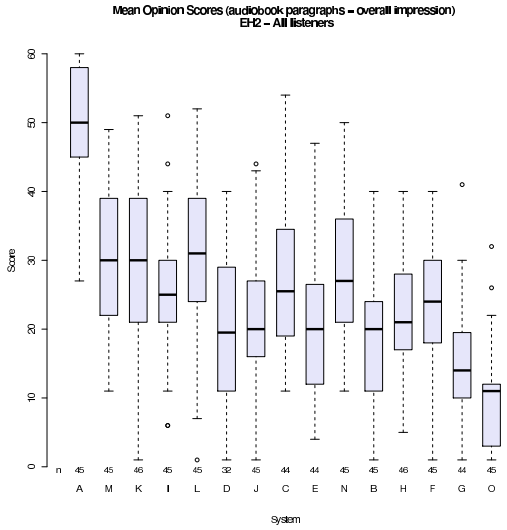

Figure 10: Results of MOS for audiobook paragraphs on overall impression (EH2).

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 3, 4, 5, and 6 show the evaluation results for EH1 task on naturalness, similarity, intelligibility and overall impression for audiobook paragraphs, respectively. Figure 7, 8, 9, and 10 show the evaluation results for EH2 task. In these figures, “A”, “B”, “C”, and “I” correspond as follows.

- A: Natural speech
- B: Festival unit selection benchmark
- C: HTS statistical parametric benchmark
- I: The 2013 NITECH HMM-based speech synthesis system

The results of listening tests showed that our system “I” outperformed the benchmark unit-selection system “B” and was as good as the HTS benchmark system “C” in naturalness and similarity for EH1 and EH2 tasks. In terms of intelligibility, our system “I” also outperformed the benchmark unit-selection system “B”. Furthermore, our system “I” was high-ranking in naturalness and intelligibility evaluation in all institutions. These results indicate that our system “I” generated the high natural and intelligible speech. In the evaluation results of audiobook paragraphs, however, our system “I” was worse than other high-ranking institutions. This is because our

HMM-based speech synthesis system generates less expressive speech, though synthesized speech was smooth. Although stable acoustic models also were constructed using the techniques of data pruning and chapter adaptive training, expressive expressions might be removed in our system "I". Synthesizing expressive speech will be future work for our system.

6. Conclusion

We described HMM-based speech synthesis system developed at the Nagoya Institute of Technology (NITECH) for the Blizzard Challenge 2013. The techniques of alignment for audiobooks, data pruning and chapter adaptive training were applied for synthesizing natural speech. The results of listening tests showed that our system generated high natural and intelligible speech. Synthesizing expressive speech will be future work.

7. Acknowledgments

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*, 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] Y. Zhao, D. Peng, L. Wang, M. Chu, Y. Chen, P. Yu, and J. Guo, "Constructing stylistic synthesis databases from audio books," *Proceedings of Interspeech*, pp. 1750–1753, 2006.

[9] K. Prahallad, R. Toth, A., and A. Black, "Automatic building of synthetic voices from large multi-paragraph speech databases," *Proceedings of Interspeech*, pp. 2901–2904, 2007.

[10] N. Braunschweiler, M. Gales, and S. Buchholz, "Lightly supervised recognition for automatic alignment of large coherent speech recordings," *Proceedings of Interspeech*, pp. 2222–2225, 2010.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] Festival. [Online]. Available:, http://www.festvox.org/festival/.

[16] A. Stolcke, "SRILM – an extensible language modeling toolkit," *Proceedings of Intl. Conf. Spoken Language Processing*, vol. 2, 2002.

[17] M. A. H. Huijbregts, "Segmentation, diarization and speech transcription: Surprise data unraveled," *PhD thesis, University of Twente.*

[18] T. Anastasakos, J. McDonough, and J. Makhoul, "Speaker adaptive training: a maximum likelihood approach to speaker normalization," *Proceedings of ICASSP 1997*, pp. 813–816, 1997.