Improved Average-Voice-based Speech Synthesis Using Gender-Mixed Modeling and a Parameter Generation Algorithm Considering GV

Junichi Yamagishi[1], *Takao Kobayashi*[2], *Steve Renals*[1],
Simon King[1], *Heiga Zen*[3], *Tomoki Toda*[4], *Keiichi Tokuda*[3]

[1]University of Edinburgh, [2]Tokyo Institute of Technology,
[3]Nagoya Institute of Technology, [4]Nara Institute of Science and Technology,
jyamagis@inf.ed.ac.uk, takao.kobayashi@ip.titech.ac.jp, s.renals@ed.ac.uk,
simon.king@ed.ac.uk, zen@sp.nitech.ac.jp, tomoki@is.naist.jp, tokuda@nitech.ac.jp

Abstract

For constructing a speech synthesis system which can achieve diverse voices, we have been developing a speaker independent approach of HMM-based speech synthesis in which statistical average voice models are adapted to a target speaker using a small amount of speech data. In this paper, we incorporate a high-quality speech vocoding method STRAIGHT and a parameter generation algorithm with global variance into the system for improving quality of synthetic speech. Furthermore, we introduce a feature-space speaker adaptive training algorithm and a gender mixed modeling technique for conducting further normalization of the average voice model. We build an English text-to-speech system using these techniques and show the performance of the system.

1. Introduction

Recent concatenative speech synthesis approaches give us high quality synthetic speech. However, as is well known, these approaches always require large-scale speech corpora for generating natural sounding speech and as a consequence, become an inefficient choice and a major bottleneck when we need to quickly add new speakers' voices and construct a speech synthesizer which can simultaneously deal with many speakers' voices. To eliminate this bottleneck would lead to both cost reduction for building a new voices and many new applications for human-computer interfaces using speech input/output. In order to make such speech synthesis realistically feasible, we need to develop an approach in which synthetic speech comparable to that of a speaker-dependent system built using a large amount of speech data can be generated from a small amount of the speech data.

For this purpose, we have been developing speaker independent HMM-based speech synthesis in which "average voice models" are created using hidden semi-Markov models (HSMMs) and adapted with a small amount of speech data from the target speaker (e.g. [1, 2]). This speech synthesis method (Fig. 1) is referred to as "average-voice-based speech synthesis (AVSS)." By using this framework, we can obtain synthetic speech for a target speaker from even 100 utterances (about 6 minutes). Interestingly, we have shown that synthetic speech using this approach is perceived as being more natural sounding than that of the speaker-dependent (SD) system by many listeners because of the data-rich average voice model [3].

However, this system has similar drawbacks to the SD system: the synthetic speech has a "buzzy" quality, because the mel-cepstral vocoder with simple pulse or noise excitation of

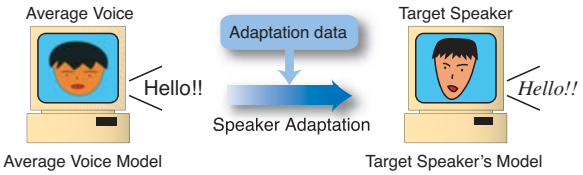

Figure 1: Average-voice-based speech synthesis.

this system is identical to that of the speaker-dependent system. In order to alleviate the problem, Zen et al. [4] incorporated a high-quality speech vocoding method, STRAIGHT with mixed excitation [5], and a parameter generation algorithm considering global variance (GV) [6] into the speaker dependent HMM system and drastically improved the quality of synthetic speech. These improvements made a great contribution to the system in an open evaluation of corpus-based text-to-speech (TTS) synthesis system, named Blizzard Challenge 2005 [7].

It is important to remember that the amount of speech data available from the target speaker is very limited in the AVSS system. To add several new parameters required for a new technique results in increase of the number of parameters to be estimated from the small amount of speech data. Therefore it would be, strictly speaking, a trade-off problem to additionally use the mixed excitation system and the parameter generation algorithm considering GV in the AVSS system. However, fortunately, the number of additional parameters for the mixed excitation system is relatively small, and that for the parameter generation algorithm considering GV is small enough to directly estimate from the adaptation data.

Therefore, we have incorporated these promising techniques into the AVSS system to improve the quality of synthetic speech. We have investigated that these techniques are effective even under condition of limited amount of speech data, based on the results of subjective evaluations. In addition to these techniques, we propose a feature-space speaker adaptive training (SAT) technique using HSMM and a gender mixed modeling technique for conducting further speaker normalization of the average voice model. Although we utilized an HSMM-based model-space SAT algorithm in our conventional system, an HSMM-based feature-space SAT algorithm is alternatively used in order to efficiently utilize both mean vectors and covariance matrices of Gaussian probability density functions (pdfs) for the normalization of the average voice model. Then, in order to reflect gender information of training speakers as a prior information in the training and adaptation stages, we develop a gender mixed modeling technique. In these experiments, we

apply the AVSS system using those techniques to U.S. English, build a new system named "AVSS 2006 " and compare the system with our conventional system. We furthermore compare the system with the speaker dependent system "Nitech-HTS 2005," which was the best system in the Blizzard Challenge 2005, in order to assess the performance of the AVSS system in the state-of-the-art TTS systems.

2. Details of the AVSS 2006 system

2.1. Speech Analysis using STRAIGHT

We use the STRAIGHT mel-cepstrum [4], $\log F_0$, and aperiodicity measures as acoustic features in the same manner as the speaker dependent system Nitech-HTS 2005. The mel-cepstral coefficients are obtained by STRAIGHT spectral analysis [5] in which F_0-adaptive spectral smoothing is carried out in the time-frequency region. The F_0 values are estimated using the following three-stage extraction to reduce error of F_0 extraction such as halving and doubling and to suppress voiced/unvoiced error. First, using IFAS-based method [8], the system extracted F_0 values for all speech data of each speaker within a common search range. Then, the F_0 range of each speaker was roughly determined based on a histogram of the extracted F_0 values. F_0 values were re-extracted in the speaker-specific range using the IFAS algorithm, fixed-point analysis [9], and ESPS get-F_0 [10]. Finally, a median value of the extracted F_0 values at each frame was utilized as an eventual F_0 value. The aperiodicity measures for mixed excitation are based on a ratio between the lower and upper smoothed spectral envelopes, and averaged on five frequency sub-bands. In addition to these static features, dynamic and acceleration features of each static feature are used.

2.2. Acoustic Models and Labels

As in the case of our conventional Japanese AVSS system, we utilize context-dependent multi-stream left-to-right MSD-HMM/HSMMs [11] in order to simultaneously model the above acoustic features and duration. Details of the phonetic and linguistic contexts for U.S. English are identical to [12]. In addition to this phonetic and linguistic information, we added gender information of speakers into the context labels for conducting the gender-mixed modeling technique in the training procedures described in the next section.

2.3. Speaker Adaptive Training

Using the above HMM/HSMMs, we trained average voice models from training data consisting of several speakers' speech. Training of the average voice model uses the SAT algorithm. Although we utilized a model-space SAT algorithms [13] using linear transformations of mean vectors of Gaussian pdfs in our conventional systems [1, 2], a feature-space SAT algorithm [14] is used as an alternative algorithm in the AVSS 2006 system to efficiently utilize both mean vectors and covariance matrices of the Gaussian pdfs for the speaker normalization of the average voice model. We can derive the feature-space SAT in the framework of HSMM in a similar way to [1]. Here we assume that each state of the HSMM has the following an output pdf $b_i(\boldsymbol{o})$ and a duration pdf $p_i(d)$:

$$b_i(\boldsymbol{o}) = \mathcal{N}(\boldsymbol{o}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \tag{1}$$

$$p_i(d) = \mathcal{N}(d; m_i, \sigma_i^2). \tag{2}$$

where $\boldsymbol{o}$ and d is an observation vector and a duration at state i, respectively. The feature-space SAT of the HSMM estimates the parameters of the Gaussian pdfs as follows:

$$\overline{\boldsymbol{\mu}}_i = \frac{\sum_{f=1}^{F}\sum_{t=1}^{T_f}\sum_{d=1}^{t}\gamma_t^d(i)\sum_{s=t-d+1}^{t}\overline{\boldsymbol{o}}_s^{(f)}}{\sum_{f=1}^{F}\sum_{t=1}^{T_f}\sum_{d=1}^{t} d\cdot\gamma_t^d(i)} \tag{3}$$

$$\overline{\boldsymbol{\Sigma}}_i = \frac{\sum_{f=1}^{F}\sum_{t=1}^{T_f}\sum_{d=1}^{t}\gamma_t^d(i)\sum_{s=t-d+1}^{t}(\overline{\boldsymbol{o}}_s^{(f)}-\overline{\boldsymbol{\mu}}_i)(\overline{\boldsymbol{o}}_s^{(f)}-\overline{\boldsymbol{\mu}}_i)^\top}{\sum_{f=1}^{F}\sum_{t=1}^{T_f}\sum_{d=1}^{t} d\cdot\gamma_t^d(i)} \tag{4}$$

$$\overline{m}_i = \frac{\sum_{f=1}^{F}\sum_{t=1}^{T_f}\sum_{d=1}^{t}\gamma_t^d(i)\cdot\overline{d}^{(f)}}{\sum_{f=1}^{F}\sum_{t=1}^{T_f}\sum_{d=1}^{t}\gamma_t^d(i)} \tag{5}$$

$$\overline{\sigma}_i^2 = \frac{\sum_{f=1}^{F}\sum_{t=1}^{T_f}\sum_{d=1}^{t}\gamma_t^d(i)\cdot(\overline{d}^{(f)}-\overline{m}_i)^2}{\sum_{f=1}^{F}\sum_{t=1}^{T_f}\sum_{d=1}^{t}\gamma_t^d(i)} \tag{6}$$

where F is number of the training speakers, T_f is total number of frames of a speaker f, and $\gamma_t^d(i)$ is the state occupancy probability at state i of the HSMM. Note that $\overline{\boldsymbol{o}}_s = \boldsymbol{\zeta}\boldsymbol{o}_s + \boldsymbol{\epsilon}$ and $\overline{d} = \chi d + \nu$ are linearly transformed observation vector and duration, respectively. These transformation matrices ($\boldsymbol{W} = [\boldsymbol{\zeta}, \boldsymbol{\epsilon}]$ and $\boldsymbol{X} = [\chi, \nu]$) are simultaneously estimated using the HSMM-based CMLLR algorithm [15]. This technique can be viewed as a generalized version of several normalization techniques such as CMN, CVN, VTLN, and bias removal of F_0 and duration. Since this HSMM-based feature-space SAT algorithm requires a lot of computation, we basically train the acoustic models using the HMM-based feature-space SAT algorithm and apply the HSMM-based SAT algorithm in the final embedded training procedures (see Fig. 2).

Another advantage of this feature-space SAT is feasibility. As reported in [14], in the the model-space SAT algorithms, it is necessary to store a full matrix for each Gaussian pdf, or store statistics for each Gaussian component for every speaker. In our *speaker-independent* HMM-based speech synthesis system, the number of the Gaussian pdfs reaches $\mathcal{O}(10^7)$ or more, and it partly makes the parameter estimation impractical. In particular, the embedded training procedures in which we could use the model-space SAT were restricted to the training procedures in which the parameters of the Gaussian pdfs were tied among several pdfs. On the other hand, we can apply the feature-space SAT algorithm to all the embedded training procedures and conduct further normalization in the training of the average voice model.

2.4. Gender-Mixed Modeling

In general, speech data weaves speaker-dependent characteristics with gender-dependent characteristics in addition to phonetic and prosodic features. We must reproduce both the gender-dependent characteristics as well as the speaker-dependent characteristics of the target speaker in our system. If large amounts of training data for both genders are available, it would be the most efficient choice to use gender-dependent average voice models using enough training data as an initial model of the speaker adaptation. However, in practice, we encounter common problems from the amount of the training data available from either gender or both genders being limited. In such cases, it would not be the best choice to use gender-dependent average voice models. In addition to this, it is not straightforward to clarify that how many training sentences and speakers are enough for constructing the appropriate gender-dependent average voice models in any condition.

Another practical approach is to use a gender-independent average voice model (or the opposite gender-dependent model

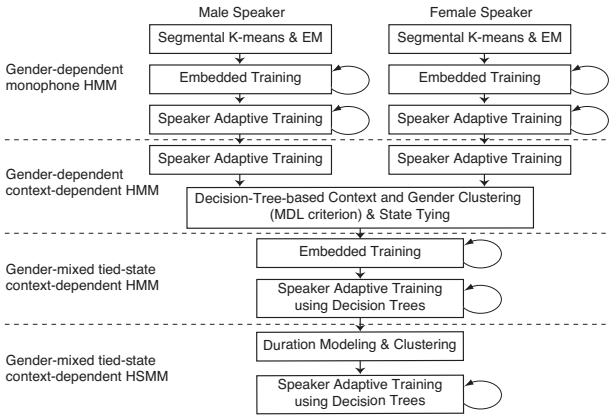

Figure 2: Details of gender-mixed modeling. This modeling technique consists of the speaker adaptive training and the decision-tree-based context and gender clustering.

using enough training data) as an initial model, instead of the correct gender-dependent average voice model. However, we have shown that naturalness and similarity of the synthetic speech using those average voice models becomes significantly worse than that of the synthetic speech using the correct gender-dependent average voice model [16]. This is a logical conclusion because we have to adapt not only speaker-dependent characteristics but also gender-dependent characteristics of the average voice model based on a small amount of the adaptation data. An alternative approach is to simultaneously use the gender-dependent average voice models to complement one another and to perform soft decisions in the speaker adaptation [16]. However, there was no significant improvements between the results of the simultaneous use of the gender-dependent average voice models and those of the single gender-dependent average voice model. Although the simultaneous use of the gender-dependent average voice models could complement one another, it required twice as many parameters for the adaptation as the gender-dependent average voice model, and it seemed to suffer from "curse of dimensionality." In summary, we are required to develop an approach which satisfies the following three conditions: 1) it reflects the gender-dependent characteristics as a prior information, 2) it makes the best possible use of the training data from both genders and complements one other if necessary, and 3) it does not increase the number of parameters required for the speaker adaptation.

To achieve this, we propose a *gender-mixed modeling* technique. The key idea of this gender-mixed modeling is similar to *style-mixed modeling* proposed in [17]. The gender-mixed modeling technically includes the speaker adaptive training and a decision-tree-based context and gender clustering technique. The actual training procedures for the modeling were conducted as follows (see Fig. 2). In order to conduct both normalization of the speaker-dependent characteristics and conservation of the gender-dependent characteristics, we first train gender-dependent monophone HMMs using the SAT algorithm. Then we convert them into gender-dependent context-dependent HMMs, and re-estimate the model parameters using the SAT algorithm again. Then, using the state occupancy probabilities obtained in the SAT framework, the decision-tree-based context clustering technique using minimum description length (MDL) criterion is applied to the HMMs, and the model parameters of the HMMs at each leaf node of the decision trees

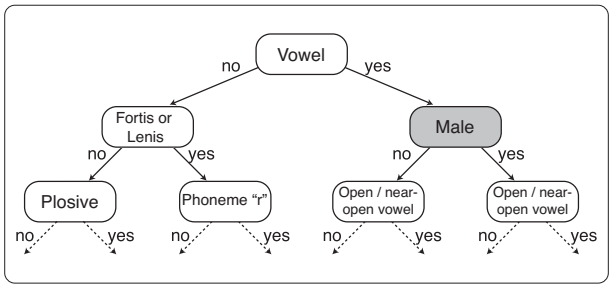

Figure 3: Part of a constructed decision tree in the gender-mixed modeling. Genders of training speakers are split by using gender-related questions as well as other contexts.

are tied. In the clustering, gender information of each speaker is treated as one of contexts for the clustering, and the clustering technique is applied to both the gender-dependent models at the same time. As a result, the gender information is included in a single acoustic model. Note that the decision trees were separately constructed for each state of mel-cepstrum, $\log F_0$, aperiodicity measures, and duration parts. Hence, when the target feature is generally gender-specific, such as $\log F_0$, the gender would be automatically split at around a root node of the tree by using gender-related questions, and the pdfs of the feature can keep the gender-dependent characteristics if required. Then, when dependency on gender of the target feature locally occurs such as duration, the gender information are automatically split as well as other contexts during the construction of a decision tree, and thereby we can make use of the training data from both genders laconically. We refer to the resulting model as a gender-mixed average voice model. Figure 3 shows a part of the constructed decision tree for the mel-cepstral part in the fifth state of the HMMs.

We re-estimate the clustered HMMs using SAT algorithm with piecewise linear regression functions. To determine regression classes for the piecewise linear regression, the decision trees constructed for the gender-mixed model are used, since use of the decision tree automatically reflects both differences of gender information and phonetic and linguistic information, and it is expected that more appropriate normalization for the average voice model is conducted. We then calculate initial duration pdfs from trellises of the HMMs [18], and conduct the decision-tree-based context and gender clustering for the duration pdfs. Using the tied duration pdfs, we perform the HSMM-based SAT algorithm with piecewise linear regression functions in order to normalize speaker characteristics included in the duration pdfs as well as other acoustic features. In each iteration of these SAT stages, we first estimated transformation matrices three times, and then updated mean vectors of both output and duration pdfs, their covariance matrices, weight for MSD, and transition matrices five times. Then we repeated the iterations three times in each SAT stage.

In the speaker adaptation stage, we adapt the gender-mixed average voice model to that of the target speaker by using a small amount of speech data with gender information of the target speaker. We utilize a combined algorithm of HSMM-based constrained structural maximum a posteriori linear regression (CSMAPLR) [19] and maximum a posteriori (MAP) adaptation [3]. In the CSMAPLR adaptation, the decision trees for the gender-mixed average voice model are used for the same reason as the above SAT algorithm with piecewise linear regression functions.

2.5. Parameter Generation Considering Global Variance

In the synthesis stage, input text is first transformed into a sequence of context-dependent phoneme labels with the gender information of the target speaker. Based on the label sequence, a sentence HSMM is constructed by concatenating context-dependent HSMMs. From the sentence HSMM, mel-cepstrum, $\log F_0$, and aperiodicity-measure sequences are obtained using the parameter generation algorithm considering GV [6], in which phoneme durations are determined using the duration pdfs. The parameter generation algorithm is a penalized maximum likelihood method in which the GV pdf (a Gaussian pdf for the variance of the trajectory at utterance level) acts as a penalty for the likelihood function. The algorithm tries to keep the global variance of the generated trajectory as wide as that of the target speaker, while maintaining an appropriate parameter sequence in the sense of maximum likelihood. It is possible to adapt the GV pdf from a speaker-independent model to that of a target speaker using MAP adaptation. However, the number of parameters of a GV pdf is very small. Specifically, it is equal to the dimensionality of the static features. Hence we directly estimate the GV pdf from the adaptation data. The generation method for speech waveforms is identical to that of Nitech-HTS 2005. A one-pitch waveform is synthesized from STRAIGHT mel-cepstral coefficients and the mixed excitation with the MLSA filter, and then a synthesized waveform was generated with PSOLA.

3. Experiments

3.1. Experimental conditions

We carried out several subjective and objective evaluation tests to assess the performance of the AVSS 2006 system. We used the CMU-ARCTIC speech database, which contains a set of about a thousand phonetically balanced sentences uttered by 4 male speakers (AWB, BDL, JMK, RMS) and 2 female speakers (CLB, SLT), and a speech database, which was released from ATR for the purpose of the Blizzard Challenge 2007 and contains the same sentences as that of CMU-ARCTIC speech database and additional sentences uttered by a male speaker EM001. To model the synthesis units, we used the "radio" phone set of the Festival speech synthesis system, and took the phonetic and linguistic contexts included in the utterance files of the Festival speech synthesis system into account.

Speech signals were sampled at a rate of 16 kHz and windowed by an F_0-adaptive Gaussian window with a 5-ms shift. The feature vectors consisted of 25 STRAIGHT mel-cepstral coefficients (including the zeroth coefficient), $\log F_0$, aperiodicity measures, and their dynamic and acceleration coefficients. We used 5-state left-to-right context-dependent multi-stream MSD-HSMMs without skip paths. Each state had a single Gaussian pdf with a diagonal covariance matrix. In the speaker adaptation, the transformation matrices were triblock diagonal corresponding to the static, dynamic, and acceleration coefficients.

3.2. Evaluation of the AVSS 2006 system

First, we evaluated naturalness and similarity of the synthetic speech generated from the adapted model. We chose a male speaker AWB as a target speaker of the speaker adaptation and used 3 male speakers (BDL, JMK, RMS) and 2 female speakers (CLB, SLT) of CMU-ARCTIC database as training speakers for the average voice model. The number of training data

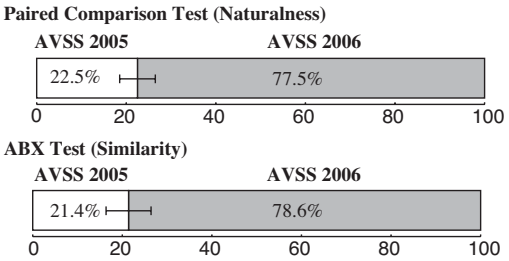

Figure 4: The average preference scores of the paired comparison test and the ABX test using our conventional system (AVSS 2005 system) and the proposed system (the AVSS 2006 system).

from each speaker was about 1000 sentences and the number of the adaptation sentences from the target speaker was 100 sentences selected from the corpus randomly. Then, ten test sentences which were not included in either the training or the adaptation data were used for the subjective evaluations. We constructed our conventional system (AVSS 2005 system) [2] and the AVSS 2006 system using the above training data and adapted the resulting average voice models of each system to the target speaker using the above adaptation data. Note that the shared-decision-tree-based context clustering algorithm was not used in both systems, since the algorithm is a directly-opposed idea from that of gender mixed modeling.

We then conducted a paired comparison test to investigate that these techniques are effective even under condition of limited amount of speech data. We compared the synthesized speech generated from the adapted models using the AVSS 2005 or 2006 systems. The subjective evaluations were conducted via the Internet. 28 subjects were presented a pair of synthetic speech utterances generated from the adapted models in random order, and asked which speech sounded more natural. At the same moment, we conducted an ABX comparison test to assess adaptation performance of the average voice models of both systems. In the ABX test, the subjects were presented a reference speech in addition to the above pair of synthesized speech, and asked to select the first or second synthetic speech as being similar to the reference speech. The reference speech was the recorded original speech. The same test sentences as the paired comparison test were used.

Figure 4 shows the average preference scores with 95% confidence interval of the paired comparison test and the ABX test. From this figure, we can see that naturalness and similarity of the synthetic speech generated from the adapted model using the AVSS 2006 system are drastically improved compared to our conventional system. In order to analyze which technique brings this good result, we separately investigated effects of STRAIGHT, feature-space SAT, gender mixed modeling, and parameter generation algorithm considering GV using preliminary evaluations. From the preliminary evaluations, we confirmed that each method had some effect, and above all the parameter generation algorithm considering GV made a huge contribution to the improvements in these subjective evaluations. However, it is interesting to note that objective measures such as mel-cepstral distance or RMSE of $\log F_0$ between synthetic speech using GV and real speech became worse than those between synthetic speech without GV and real speech. Since the experimental results for the STRAIGHT and the parameter generation algorithm considering GV were similar to the results of speaker-dependent system [4], we report the effect of the feature-space SAT and gender mixed modeling in the next subsections.

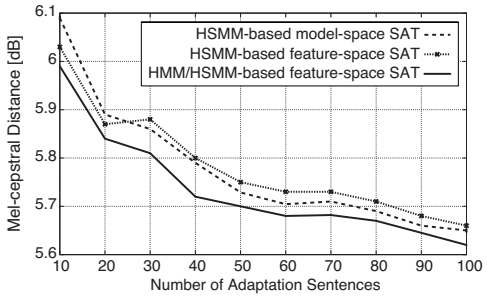

Figure 5: Objective evaluation of the SAT algorithm: Average mel-cepstral distance.

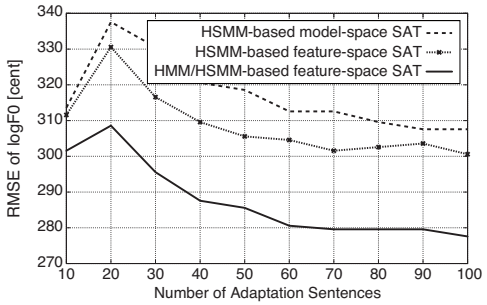

Figure 6: Objective evaluation of the SAT algorithm: RMSE of $\log F_0$.

3.3. Evaluation of the Feature-Space SAT

We evaluated the feature-space SAT algorithm using two types of objective evaluations based on the average mel-cepstral distance and RMSE of $\log F_0$. In these evaluations, we chose a male speaker EM001 as a target speaker of the speaker adaptation and used 4 male speakers (AWB, BDL, JMK, RMS) and 2 female speakers (CLB, SLT) of CMU-ARCTIC database as training speakers for the average voice model. We constructed three kinds of the gender-independent average voice model using HSMM-based model-space SAT and HMM/HSMM-based feature-space SAT, and adapted the resulting average voice models of each system to the target speaker. The amount of training data from each speaker was about 1100 sentences. The adaptation data was from 10 sentences to 100 sentences. 1000 test sentences were used for the evaluations, and these were included in neither the training nor the adaptation data. For the calculation of the average mel-cepstral distance and the RMSE of $\log F_0$, the state duration of each HSMM was adjusted after Viterbi alignment with the target speakers' real utterance.

Figure 5 shows the average mel-cepstral distance between spectra generated from the adapted model and spectra obtained by analyzing target speakers' real utterance. Figure 6 shows the RMSE of $\log F_0$ between F_0 patterns of synthetic and real speech. Silence, pause, and consonant regions were eliminated from the mel-cepstral distance calculation. Since F_0 is not observed in the unvoiced region, the RMSE of $\log F_0$ was calculated in the region where both the generated and the real F_0 were voiced. Comparing HSMM-based model-space and feature-space SAT only, one sees that the feature-space SAT gives slightly better results in the adaptation of the F_0 parameter, whereas the error of the feature-space SAT partly becomes slightly worse in the adaptation of the spectral parameters. However, we can also see that when we consistently apply the feature-space SAT to all the embedded training procedures for HMMs and HSMMs, both the mel-cepstral distance and RMSE of $\log F_0$ significantly decrease.

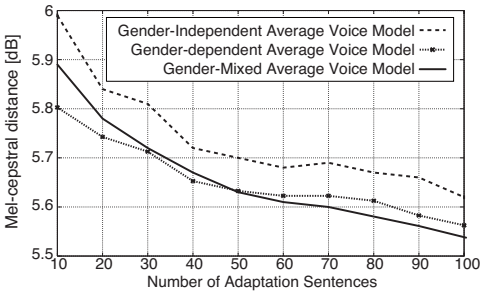

Figure 7: Objective evaluation of the gender-mixed modeling: Average mel-cepstral distance.

3.4. Evaluation of the Gender-Mixed Modeling

Then, we evaluated the gender-mixed modeling using the mel-cepstral distance. We constructed the gender-independent, gender-dependent, and gender-mixed average voice models and adapted these average voice models to the target speaker using the same adaptation data. The experimental condition on the speech data in this subsection is the same as 3.3.

Figure 7 shows the average mel-cepstral distance between spectra generated from the adapted model and spectra obtained by analyzing target speakers' real utterance. Silence, pause, and consonant regions were eliminated from the mel-cepstral distance calculation. Comparing the gender-dependent and gender-mixed average voice models, we can see that from 10 to 50 adaptation sentences, the gender-dependent modeling is generally better, whereas the gender-mixed modeling becomes better from the 50 to 100 adaptation sentences. We believe that this is because the gender-mixed average voice model has many more pdfs than the gender-dependent model, although we need to perform further experiments to investigate it.

3.5. Comparison with Nitech-HTS 2005

Finally, we conducted a comparison category rating (CCR) test and assessed the performance of the AVSS system with the state-of-the-art TTS systems. For this purpose, we compared the synthesized speech generated from the AVSS 2006 system with that of the speaker-dependent system Nitech-HTS 2005. The only difference between this Nitech-HTS 2005 system and a system reported in [4] is dimension of mel-cepstral coefficients. In [4], 39 mel-cepstral coefficients were used. However, this increases the number of parameters of the matrix for linear transformation. Hence we consistently utilize 24 mel-cepstral coefficients for both systems. The experimental condition on the training data in this subsection is the same as 3.3. We constructed the AVSS 2006 system using the training data and adapted the resulting average voice model to the target speaker using 100 sentences of the target speaker EM001. The speaker-dependent system Nitech-HTS 2005 was built using 1000 sentences of the target speaker EM001. For reference, we compared synthesized speech generated from an adapted model using the same 1000 sentences of the target speaker EM001 as adaptation data. 25 subjects were first presented with synthetic speech of Nitech-HTS 2005 as a reference speech and then with synthesized speech from the adapted models using 100 sentences or 1000 sentences in random order. Then the subjects were asked to comprehensively evaluate the synthetic speech generated from the adapted models compared with the reference speech. The evaluation was done on a 5-point scale, that is, 2 for better, 1 for slightly better, 0 for almost the same, -1 for slightly worse, and 2 for worse than the reference speech.

The average values and their 95% confidence interval of each adapted model in the CCR tests were **0.140±0.145** for 100 sentences and **0.424±0.08** for 1000 sentences, respectively. The values indicate that the AVSS 2006 system can synthesize speech of almost the same quality as the Nitech-HTS 2005 system from just 100 sentences, that is, 10% of the training data for the speaker-dependent systems. This is a very meaningful result since the Nitech-HTS 2005 system was evaluated as a best system in the Blizzard Challenge 2005, and we can say that the synthetic speech using the AVSS 2006 system bears comparison with other state-of-the-art TTS systems. Furthermore, we can see that the synthetic speech generated from the AVSS 2006 system using 1000 sentences is judged to be slightly better than those using 100 sentences and Nitech-HTS 2005 system. This result implies that this average voice approach is no longer just a speaker conversion system and it has the potential to surpass the common speaker-dependent approach.

4. Conclusions

In this paper, we incorporated a high-quality speech vocoding method STRAIGHT and a parameter generation algorithm with GV into the AVSS system for improving quality of synthetic speech. In addition to these techniques, we also proposed a feature-space SAT algorithm using the HSMM and a gender mixed modeling technique for conducting further speaker normalization of the average voice model. We applied the AVSS system using these techniques to U.S. English and built a new system named AVSS 2006 system. From the subjective evaluations, we shown that naturalness and similarity of the synthetic speech of the AVSS 2006 system were drastically improved compared to our conventional system, and then the AVSS 2006 can synthesize speech of the almost the same quality as the Nitech-HTS 2005 system from just 100 sentences.

Our future work is to develop a modeling technique for dealing with several dialects of English in the framework of the average voice model. We will also focus on developing an unsupervised speaker adaptation algorithm for speech synthesis.

5. Acknowledgments

This research was conducted for the purpose of the Blizzard Challenge 2006 and 2007. The authors would like to thank Dr. Yasser Hifny Abdel-Haleem of IBM T.J. Watson research center for his original idea on gender-mixed modeling.

6. References

[1] 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, Feb. 2007.

[2] J. Yamagishi, T. Kobayashi, M. Tachibana, K. Ogata, and Y. Nakano, "Model adaptation approach to speech synthesis with diverse voices and styles," in *Proc. ICASSP 2007*, Apr. 2007, pp. 1233–1236.

[3] K. Ogata, M. Tachibana, J. Yamagishi, and T. Kobayashi, "Acoustic model training based on linear transformation and MAP modification for HSMM-based speech synthesis," in *Proc. ICSLP 2006*, Sept. 2006, pp. 1328–1331.

[4] 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.

[5] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigné, "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.

[6] 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, May 2007.

[7] A.W. Black and K. Tokuda, "The Blizzard Challenge – 2005: Evaluating corpus-based speech synthesis on common datasets," in *Proc. EUROSPEECH 2005*, Sept. 2005, pp. 77–80, http://festvox.org/blizzard/blizzard2005.html.

[8] D. Arifianto, T. Tanaka, T. Masuko, and T. Kobayashi, "Robust F0 estimation of speech signal using harmonicity measure based on instantaneous frequency," *IEICE Trans. Information and Systems*, vol. E87-D, no. 12, pp. 2812–2820, Dec. 2004.

[9] H. Kawahara, H. Katayose, A. Cheveigné, and R. Patterson, "Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity," in *Proc. EUROSPEECH 1999*, Sept. 1999, pp. 2781–2784.

[10] Entropic Research Laboratory Inc, *ESPS Programs Version 5.0*, 1993.

[11] 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, May 2007.

[12] K. Tokuda, H. Zen, and A.W. Black, "An HMM-based speech synthesis system applied to English," in *Proc. of IEEE Speech Synthesis Workshop*, Sept. 2002.

[13] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Proc. ICSLP-96*, Oct. 1996, pp. 1137–1140.

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

[15] 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*, May 2006, pp. 77–80.

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

[17] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "Acoustic modelingof speaking styles and emotional expressions in HMM-based speech synthesis," *IEICE Trans. Inf. & Syst.*, vol. E88-D, no. 3, pp. 503–509, Mar. 2005.

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

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