INTERSPEECH 2009 BRIGHTON

An improved minimum generation error based model adaptation for HMM-based speech synthesis

Yi-Jian Wu[†] , *Long Qin*[‡], *Keiichi Tokuda*[†]

[†] Nagoya Institute of Technology, Japan
[‡] Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA
yjwu@sp.nitech.ac.jp, lqin@cs.cmu.edu, tokuda@nitech.ac.jp

Abstract

A minimum generation error (MGE) criterion had been proposed for model training in HMM-based speech synthesis. In this paper, we apply the MGE criterion to model adaptation for HMM-based speech synthesis, and introduce an MGE linear regression (MGELR) based model adaptation algorithm, where the regression matrices used to transform source models are optimized so as to minimize the generation errors of adaptation data. In addition, we incorporate the recent improvements of MGE criterion into MGELR-based model adaptation, including state alignment under MGE criterion and using a log spectral distortion (LSD) instead of Euclidean distance for spectral distortion measure. From the experimental results, the adaptation performance was improved after incorporating these two techniques, and the formal listening tests showed that the quality and speaker similarity of synthesized speech after MGELR-based adaptation were significantly improved over the original MLLR-based adaptation.

Index Terms: Speech synthesis, HMM, speaker adaptation, minimum generation error, linear regression

1. Introduction

HMM-based speech synthesis method [1, 2] had been under developed for a decade, and shown its potential to realize a speech synthesis system with high quality and flexibility [3]. One of the unique capabilities of HMM-based speech synthesis is the ability to adapt the models in order to modify the characteristics of synthesized speech, including the change of speaker identity, speaking style, and so on. This is achieved by modifying the HMM parameters using model adaptation techniques. Several model adaptation algorithms, which were originally proposed for speech recognition, including Maximum a Posteriori (MAP), Maximum Likelihood Linear Regression (MLLR) [4], Constrained MLLR (CMLLR) [5], and so on, have been applied to HMM-based speech synthesis [6, 7]. It has also been demonstrated that speaker adaptation of an "Average Voice" model [8] is superior to speaker adaptation of a speaker-dependent model.

Recently, a minimum generation error (MGE) criterion [9] was proposed for HMM training in order to solve two issues related to ML-based HMM training for speech synthesis, which includes the mismatch between training and application of HMM, and the ignorance of constraint between static and dynamic features. In this new criterion, a generation error function using Euclidean distance was defined, and the HMM parameters were optimized so as to minimize the total generation errors of training data. In [10], a log spectral distortion (LSD) was adopted to replace the Euclidean distance for calculating the generation error between the original and generated line spectral pairs (LSPs) [11] in MGE training, and the quality of synthesized speech was significantly improved.

Since the ML criterion is also used for model adaptation, this paper continue to apply the MGE criterion to model adaptation for HMM-based speech synthesis, and introduce a MGE linear regression (MGELR) algorithm [12]. In order to effectively make use of limited adaptation data, the source models are firstly grouped into regression classes, where the models within one class share the same linear transformation matrix. After initialized using the MLLR-based model adaptation, the parameters of the transforms are re-estimated under the MGE criterion, where the parameters are optimized to minimize the total generation errors of the adaptation data. Furthermore, we incorporate the recent improvements of MGE criterion into MGELR-based model adaptation, including state alignment under MGE criterion and using the LSD instead of the Euclidean distance for spectral distortion measure, and investigate the effectiveness of these two techniques.

The rest of this paper is organized as follows. In section 2, we first briefly review the MGE criterion for HMM training. In section 3, we present the details of MGELR-based model adaptation algorithm. In section 3, we describe the experiments used to evaluate the performance of the MGELR-based speaker adaptation and present the results. Finally, our conclusions are given in section 4.

2. Minimum generation error criterion

The basic concept of MGE criterion is to calculate the generation errors by incorporating the parameter generation into training process, and then optimize the HMM parameters so as to minimize the total generation errors of training data.

2.1. Parameter generation

For a given HMM λ and the state sequence $\boldsymbol{q}$, the parameter generation algorithm [1] is to determine the speech parameter vector sequence $\boldsymbol{o} = [\boldsymbol{o}_1^\top, \boldsymbol{o}_2^\top, \ldots, \boldsymbol{o}_T^\top]^\top$ which maximizes $P(\boldsymbol{o}|\boldsymbol{q}, \lambda)$. In HMM-based speech synthesis, $\boldsymbol{o}_t = [\boldsymbol{c}_t^\top, \Delta^{(1)}\boldsymbol{c}_t^\top, \Delta^{(2)}\boldsymbol{c}_t^\top]^\top$ includes not only static but also dynamic features. The constraint between static and dynamic feature vector can be formulated as $\boldsymbol{o} = \boldsymbol{W}\boldsymbol{c}$, where $\boldsymbol{c} = [\boldsymbol{c}_1^\top, \boldsymbol{c}_2^\top, \ldots, \boldsymbol{c}_T^\top]^\top$, and $\boldsymbol{W}$ is a regression matrix [1] for calculating dynamic features.

Under this constraint, parameter generation is equivalent to determining $\boldsymbol{c}$ to maximize $P(\boldsymbol{o}|\lambda, \boldsymbol{q})$. By setting $\partial P(\boldsymbol{o}|\lambda, \boldsymbol{q})/\partial \boldsymbol{c} = 0$, we obtain

$$\bar{\boldsymbol{c}}_{\boldsymbol{q}} = \boldsymbol{R}_{\boldsymbol{q}}^{-1} \boldsymbol{r}_{\boldsymbol{q}}, \tag{1}$$

where

Yi-Jian Wu is currently with the TTS group of Microsoft Business Division at Beijing, China. Email: yijiwu@microsoft.com

$$\boldsymbol{R}_{\boldsymbol{q}} = \boldsymbol{W}^{\top}\boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1}\boldsymbol{W}, \quad \boldsymbol{r}_{\boldsymbol{q}} = \boldsymbol{W}^{\top}\boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1}\boldsymbol{\mu}_{\boldsymbol{q}}, \tag{2}$$

and $\boldsymbol{\mu}_{\boldsymbol{q}} = \left[\boldsymbol{\mu}_1^{\top}, \ldots, \boldsymbol{\mu}_T^{\top}\right]^{\top}$ and $\boldsymbol{\Sigma}_{\boldsymbol{q}} = \mathrm{diag}\left(\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_T\right)$ are the mean vector and covariance matrix related to $\boldsymbol{q}$, respectively.

2.2. Generation error

With the generated feature vector $\bar{\boldsymbol{c}}_{\boldsymbol{q}}$, we need to measure the distortion between the original and generated feature vector. In the baseline MGE criterion, the Euclidean distance was used to calculate the distortion

$$D(\boldsymbol{c}, \bar{\boldsymbol{c}}_{\boldsymbol{q}}) = \| \boldsymbol{c} - \bar{\boldsymbol{c}}_{\boldsymbol{q}} \|^2 . \tag{3}$$

Although the posterior probability can be used to weight the distance for all possible state sequence, it is computationally expensive for this direct calculation. Therefore, the representative n-best paths can be used to approximate the generation error. In our current implementation, only the optimal state sequence is used, and the generation error is defined as

$$e(\boldsymbol{c}, \lambda) = D(\boldsymbol{c}, \bar{\boldsymbol{c}}_{\hat{\boldsymbol{q}}}), \tag{4}$$

where $\hat{\boldsymbol{q}}$ is the optimal state sequence for $\boldsymbol{o}$. This refers to a Viterbi-type MGE training. In the rest of the paper, we use $\boldsymbol{q}$ to denote $\hat{\boldsymbol{q}}$.

2.3. Re-estimation of model parameters

Based on the generation error definition, the parameter generation process is incorporated into HMM training for calculating the total generation errors for all training data $\boldsymbol{c}_n$, which is

$$E(\lambda) = \sum_n e(\boldsymbol{c}_n, \lambda). \tag{5}$$

Finally, the objective of MGE criterion is to optimize the model parameters so as to minimize the total generation errors, i.e.,

$$\hat{\lambda} = \arg\min E(\lambda). \tag{6}$$

As direct solution for Eq. (6) is mathematically intractable, probabilistic descent (PD) [13] method was adopted for parameter optimization. The details of updating rules for mean and variance parameters in MGE training can be found in [9].

3. MGELR algorithm for model adaptation

The Maximum Likelihood Linear Regression (MLLR) algorithm had been successfully applied for model adaptation in HMM-based speech synthesis. Since the MGE criterion had been proposed to solve the two issues related to the ML-based HMM training, we introduce a corresponding MGELR algorithm for model adaptation in HMM-based speech synthesis.

3.1. Linear transformation for model parameters

In the MLLR-based model adaptation framework, the linear transformations for the mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ of one model are defined as

$$\hat{\boldsymbol{\mu}} = \boldsymbol{\Phi}\boldsymbol{\xi}, \tag{7}$$

$$\hat{\boldsymbol{\Sigma}}^{-1} = \boldsymbol{A}\boldsymbol{H}^{-1}\boldsymbol{A}^{\top}, \tag{8}$$

where $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ are the transformed mean vector and covariance matrix, $\boldsymbol{\Phi}$ and $\boldsymbol{H}$ are the transformation matrices for mean and variance parameters, $\boldsymbol{\xi} = \left[\varpi, \boldsymbol{\mu}^{\top}\right]^{\top}$ is an extended vector for $\boldsymbol{\mu}$, and $\boldsymbol{A}$ is the Cholesky decomposition factor of $\boldsymbol{\Sigma}^{-1}$, i.e.,

$$\boldsymbol{\Sigma}^{-1} = \boldsymbol{A}\boldsymbol{A}^{\top}. \tag{9}$$

For a state sequence $\boldsymbol{q} = [q_1, q_2, \ldots, q_T]$, the transformed mean vector sequence and covariance matrix can be calculated as

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{q}} = \boldsymbol{\Phi}_{\boldsymbol{q}}\boldsymbol{\xi}_{\boldsymbol{q}} \tag{10}$$

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{q}}^{-1} = \boldsymbol{A}_{\boldsymbol{q}}\boldsymbol{H}_{\boldsymbol{q}}^{-1}\boldsymbol{A}_{\boldsymbol{q}}^{\top} \tag{11}$$

where

$$\boldsymbol{\Phi}_{\boldsymbol{q}} = \mathrm{diag}\left[\boldsymbol{\Phi}_{q_1}, \boldsymbol{\Phi}_{q_2}, \ldots, \boldsymbol{\Phi}_{q_T}\right] \tag{12}$$

$$\boldsymbol{\xi}_{\boldsymbol{q}} = \left[\boldsymbol{\xi}_{q_1}^{\top}, \boldsymbol{\xi}_{q_2}^{\top}, \ldots, \boldsymbol{\xi}_{q_T}^{\top}\right]^{\top} \tag{13}$$

$$\boldsymbol{H}_{\boldsymbol{q}}^{-1} = \mathrm{diag}\left[\boldsymbol{H}_{q_1}^{-1}, \boldsymbol{H}_{q_2}^{-1}, \ldots, \boldsymbol{H}_{q_T}^{-1}\right] \tag{14}$$

$$\boldsymbol{A}_{\boldsymbol{q}} = \mathrm{diag}\left[\boldsymbol{A}_{q_1}, \boldsymbol{A}_{q_2}, \ldots, \boldsymbol{A}_{q_T}\right] \tag{15}$$

3.2. MGELR-based model adaptation

In the MGELR-based model adaptation, we incorporate the parameter generation into model adaptation process to calculate the generation errors of adaptation data, and then optimize the parameters of transformation matrices so as to minimize the total generation errors of adaptation data.

3.2.1. Generation error after transformation

After model transformation, the generation error for a feature vector sequence $\boldsymbol{c}$ in adaptation data is defined as

$$\hat{e}(\boldsymbol{c}, \lambda) = D(\boldsymbol{c}, \hat{\boldsymbol{c}}_{\boldsymbol{q}}) = \|\hat{\boldsymbol{c}}_{\boldsymbol{q}} - \boldsymbol{c}\|^{\top}, \tag{16}$$

where $\hat{\boldsymbol{c}}_{\boldsymbol{q}}$ is the generated feature vector sequence using the transformed models, which is calculated as

$$\hat{\boldsymbol{c}}_{\boldsymbol{q}} = \hat{\boldsymbol{R}}_{\boldsymbol{q}}^{-1}\hat{\boldsymbol{r}}_{\boldsymbol{q}}, \tag{17}$$

where

$$\hat{\boldsymbol{R}}_{\boldsymbol{q}} = \boldsymbol{W}^{\top}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{q}}^{-1}\boldsymbol{W}, \tag{18}$$

$$\hat{\boldsymbol{r}}_{\boldsymbol{q}} = \boldsymbol{W}^{\top}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{q}}^{-1}\hat{\boldsymbol{\mu}}_{\boldsymbol{q}}. \tag{19}$$

Similarly, we can get the total generation error by accumulating the generation error for all adaptation data, which is

$$\hat{E}(\lambda) = \sum_n \hat{e}(\boldsymbol{c}_n, \lambda). \tag{20}$$

3.2.2. Optimization of transformation matrices

Under the MGELR framework, the parameters of transformation matrices are optimized in order to minimize the total generation errors of adaptation data. Here, the PD-based method is adopted for parameter optimization. For each adaptation data $\boldsymbol{c}_\tau$, the parameter set is updated as

$$\lambda(\tau+1) = \lambda(\tau) - \epsilon_\tau \boldsymbol{B}_\tau \left.\frac{\partial \hat{e}(\boldsymbol{c}_\tau, \lambda)}{\partial \lambda}\right|_{\lambda=\lambda_\tau}. \tag{21}$$

where $\boldsymbol{B}_\tau$ is a positive definite matrix, and ϵ_τ is a learning rate that decrease when utterance index τ increase.

For the transformation matrices of the mean vector and covariance matrix related to $\hat{\boldsymbol{c}}_{\boldsymbol{q}}$, the gradients of the generation error function are calculated as

$$\frac{\partial D}{\partial \boldsymbol{\Phi}_{\boldsymbol{q}}} = 2\hat{\boldsymbol{\Sigma}}_{\boldsymbol{q}}^{-1}\boldsymbol{W}\hat{\boldsymbol{R}}_{\boldsymbol{q}}^{-1}(\hat{\boldsymbol{c}}_{\boldsymbol{q}} - \boldsymbol{c})\boldsymbol{\xi}_{\boldsymbol{q}}^{\top}, \tag{22}$$

$$\frac{\partial D}{\partial \boldsymbol{H}_{\boldsymbol{q}}^{-1}} = 2\boldsymbol{A}_{\boldsymbol{q}}^{\top}(\hat{\boldsymbol{\mu}}_{\boldsymbol{q}} - \boldsymbol{W}\hat{\boldsymbol{c}}_{\boldsymbol{q}})(\hat{\boldsymbol{c}}_{\boldsymbol{q}} - \boldsymbol{c})^{\top}\hat{\boldsymbol{R}}_{\boldsymbol{q}}^{-1}\boldsymbol{W}^{\top}\boldsymbol{A}_{\boldsymbol{q}}. \tag{23}$$

Finally, the updating rules for the transformation matrices of whole parameter set can formulated correpondingly.

3.2.3. Adaptation procedure

The whole model training and adaptation procedure based on the MGELR algorithm is implemented as follows:

1) Train the source voice model using the source speech database.
2) Conduct the MLLR-based model adaptation, and initialize the transformation matrices.
3) Obtain the optimal state alignments for all adaptation data using the MLLR-adapted HMMs.
4) Iteratively optimize the parameters of transformation matrices based on MGELR algorithm.
5) Apply the optimized transformation matrices to the source voice model.

3.3. Improvements

Here, we incorporate two recent techniques of MGE criterion into the MGELR-based model adaptation, including state alignment under MGE criterion and using the LSD instead of the Euclidean distance for spectral distortion measure, and investigate the effectiveness of these two techniques.

3.3.1. State alignment under MGE criterion

As we mentioned in Sect. 2.2, only the optimal state sequence is used for generation error definition. Under the MGE criterion, the optimal state sequence should be calculated by

$$\hat{\boldsymbol{q}} = \arg\min_{\boldsymbol{q}} e(\boldsymbol{c}, \lambda) = \arg\min_{\boldsymbol{q}} D(\boldsymbol{c}, \bar{\boldsymbol{c}}_{\boldsymbol{q}}), \tag{24}$$

However, the parameter generation process depends on the whole state sequence, which makes it intractable to search for the optimal state sequence directly using the Viterbi algorithm. In previous implementation, we used the optimal state sequence calculated under the ML criterion for approximation, which is

$$\hat{\boldsymbol{q}} \approx \arg\max_{\boldsymbol{q}} P(\boldsymbol{c}, \boldsymbol{q}|\lambda) \tag{25}$$

Under this approximation, the Viterbi algorithm can be applied. However, such approximation reduce the effect of MGE criterion.

Recently, we proposed a heuristic method to search for the optimal state sequence under MGE criterion, which is as follows:

1) Initialize the state alignment for the input utterance under Eq. (25) by using the Viterbi algorithm.
2) For each state boundary in the state alignment, try to shift it to the left (or right), and calculate the generation errors before and after shifting the state boundary.
3) If the generation errors decrease, keep the new state boundary and go back to the step 2); otherwise terminate the process.

In this procedure, we need to re-generate the whole utterance after each attempt of boundary shifting, which introduces the excessively high computational cost . Due to this, we make an approximation and only re-generate the feature vector sequence inside the window centered at the boundary, which means that the boundary location is optimized locally in each step. In this paper, we set the window size to 50 frames, which is enough to keep the accuracy.

3.3.2. LSD for spectral distortion measure

Since we used the LSPs as spectral feature for HMM modeling, the Euclidean distance between two LSPs is not so convincing as a spectral distortion measure. A log spectral distortion (LSD) was adopted in [10] to replace the Euclidean distance to calculate the distortion for generated LSPs, i.e.,

$$D_{lsd}(\boldsymbol{c}_t, \bar{\boldsymbol{c}}_t) = \frac{1}{\pi}\int_0^{\pi} [\log|A_{\boldsymbol{c}_t}(\omega)| - \log|A_{\bar{\boldsymbol{c}}_t}((\omega)|]^2 \, d\omega. \tag{26}$$

where $A_{\boldsymbol{c}_t}(\omega)$ and $A_{\bar{\boldsymbol{c}}_t}(\omega)$ are the spectra derived from the original and generated LSPs at t-th frame, respectively.

Since it is difficult to formulate the direct solution for the integration in Eq. (26). An alternative is to use a numerical integration to approximate the integral, which is calculated by accumulating the values of integrand at certain sampling points. Then Eq. (26) can be rewritten as

$$D_{lsd}(\boldsymbol{c}_t, \bar{\boldsymbol{c}}_t) = \frac{1}{N_s}\sum_{j=1}^{N_s} [\log|A_{\boldsymbol{c}}(\omega_j)| - \log|A_{\bar{\boldsymbol{c}}}(\omega_j)|]^2, \tag{27}$$

where ω_j is the location of each sampling point and N_s is the total number of sampling points.

Two sampling strategies, including the equidistance sampling and the sampling at LSP frequencies, were investigated in [10], and the experimental results showed that using the LSDs calculated by sampling at LSP frequencies outperformed that with the equidistance sampling strategy in the MGE-LSD training. In this paper, we adopt the sampling strategy by sampling at LSP frequencies, i.e.,

$$\omega_j = c_{t,j}, \quad j = 1, 2, \ldots, p, \tag{28}$$

where $c_{t,j}$ is the j-th coefficient of the original LSP vector $\boldsymbol{c}_t$.

4. Experiments

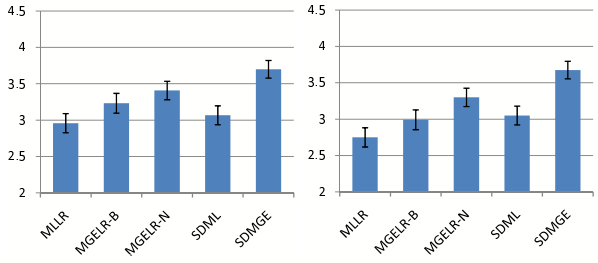
4.1. Experimental conditions

We used the CMU-ARCTIC English database [14] in the experiment. Speech data (about 1 hour) from each of 4 males (awb, bdl, rms, jmk) and 1 female (clb) was used to train the source Average Voice model. 100 utterances of speech data from another female speaker (slt) were used to adapt the source model. The acoustic features include F0 and LSP coefficients, where LSP coefficients were calculated based on spectra extracted by STRAIGHT [15]. The feature vector consists of static features (including 24-th LSP coefficients, logarithm of gain and logarithm of F0), and their delta and delta-delta coefficients. A 5-state left-to-right no-skip HMM was used, and MSD-HMM [16] was adopted for F0 modeling. In synthesis, the STRAIGHT synthesis filter was used to synthesize the speech waveform.

While using the full matrix instead of diagonal matrix for model adaptation usually improved the speaker similarity of generated speech after adaptation, it also resulted in the disorder problem of generated LSPs [17]. The band-diagonal matrix is a compromised solution considering both speaker similarity and stability of generated speech after adaptation. Here we adopted the band-diagonal matrix whose diagonal bandwidth is set to 3. In the experiment, we compared the following configurations for model adaptation:

a) MLLR: MLLR-based model adaptation;
b) MGELR-B: basic MGELR-based model adaptation with Euclidean distance;
c) MGELR-N: new MGELR-based model adaptation by incorporating LSD and state alignment by MGE;

In addition, two speaker dependent models using 1-hour speech data from the target speaker were trained under the ML-based training (SDML) and MGE-based training (SDMGE) procedures, respectively. Note that the LSD and state alignment by MGE are also incorporated into the MGE-based training. Here the speaker dependant models are regarded the upper bounds of model adaptation.



(a) MOS (b) DMOS

Figure 1: MOS and DMOS scores of synthesized speech from the adapted models and speaker dependent models

4.2. Experimental results

Two formal subjective listening tests were conducted. The first test evaluated the quality of synthesized speech using the MOS score, and the second one evaluated the speaker similarity between the natural target speech and the synthesized speech from the adapted models and speaker dependent models using DMOS score. 40 sentences, which were not included in the training data, were synthesized from the adapted models (MLLR, MGELR-B, MGELR-N) and speaker dependent models (SDML, SDMGE). Eight Japanese listeners participated in the test. Each listener evaluated 15 sets of samples consisting of five synthesized speech samples, and gave the MOS and DMOS scores for each sample. The speech samples were randomly selected for each listener from the 40 test sentences.

The results are shown in Fig. 1, with the vertical line indicating the 95% confidence intervals. In this figure, it can be seen that both speech quality and speaker similarity were improved over the MLLR-based adaptation after applying the basic MGELR adaptation, and were further improved after incorporating the two improvements of MGE criterion into the MGELR adaptation. Usually, the performance of ML-based SD model training can be regarded as the upper bound of MLLR adaptation. From the figure, the MOS and DMOS scores of MGELR-N (i.e. improved MGELR adaptation) is higher than that of SDML (i.e. ML-trained SD model), which means the performance of the improved MGELR adaptation was even over the original upper bound of MLLR adaptation. However, the performance of MGELR adaptation is still worse than that of MGE-based SD model training, which can be regarded as the upper bound of the new MGELR adaptation. Furthermore, it can be seen that MGELR-B has higher MOS score but slightly lower DMOS score comparing to the scores of SDML in the figure, which means the basic MGELR adaptation is very effective to improve the speech quality, but less effective to improve the speaker similarity.

5. Conclusions

This paper introduce an improved MGE linear regression (MGELR) based model adaptation algorithm, where two recent improvements of MGE criterion, including state alignment under MGE criterion and using a log spectral distortion (LSD) instead of Euclidean distance for spectral distortion measure, are incorporated into MGELR-based model adaptation. From the experimental results, the adaptation performance was improved after incorporating these two techniques, where the quality and speaker similarity of synthesized speech after MGELR-based adaptation were significantly improved over the original MLLR-based adaptation, and even over the ML-based speaker dependent model training.

6. Acknowledgements

This work was partly supported by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 213845 (the EMIME project).

7. References

[1] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Speech synthesis from HMMs using dynamic features," in *Proc. of ICASSP*, pp. 389-392, 1996.

[2] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proc. of ICASSP*, vol. 5, pp. 2347-2350, 1999.

[3] J. Yamagishi, H. Zen, T. Toda, and K. Tokuda, "Speaker-independent HMM-based speech synthesis system - HTS-2007 system for the Blizzard Challenge 2007", in *Blizzard Challenge 2007*.

[4] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," in *Computer Speech and Language*, vol.9, no.2, pp. 171-185, 1995.

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

[6] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Speaker adaptation for HMM-based speech synthesis system using MLLR," in *The Third ESCA/COCOSDA Workshop on Speech Synthesis*, pp. 273-276, 1998.

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

[8] J. Yamagishi, M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "A training method of average voice model for HMM-based speech synthesis," in *IEICE Trans. of Fundamentals*, vol. E86-A, no. 8, pp. 1956-1963, 2003.

[9] Y.-J. Wu and R.H. Wang, "Minimum generation error training for HMM-based speech synthesis," in *Proc. of ICASSP*, vol. 1, pp. 889–892, 2006.

[10] Y.-J. Wu and K. Tokuda, "Minimum generation error training with direct log spectral distortion on LSPs for HMM-based speech synthesis," in *Proc. of Interspeech*, pp. 577–580, 2008.

[11] F. Itakura, "Line spectrum representation of linear predictive coefficients of speech signals," in *J. Acoust. Soc. Amer.*, 1975, vol. 57, p. 535(a), p. s35(A).

[12] L. Qin, Y.-J. Wu, Z.-H. Ling, R.-H. Wang, and L.-R. Dai, "Minimum generation error lineal regression based model adaptation for HMM-based speech synthesis," in *Proc. of ICASSP*, pp. 3953-3956, Mar. 2008.

[13] S. Amari, "A theory of adaptive pattern classifiers," *IEEE Trans. Electron. Comput.*, vol. EC-16, no. 3, pp. 299–307, 1967.

[14] J. Kominek and A. Black, "The CMU ARCTIC speech databases for speech synthesis research," Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, Tech. Rep. CMULTI-03-177, http://festvox.org/cmu_arctic/, 2003.

[15] H. Kawahara, I. Masuda-Katsuse and A. deCheveigne, "Restructuring speech representations using pitch-adaptive time-frequency smoothing and an instanta-neous-frequency-based F0 extraction: possible role of a repetitive structure in sounds," in *Speech Communication*, vol. 27, pp. 187-207, 1999.

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

[17] L. Qin, Y.-J. Wu, Z.H. Ling and R.H. Wang, "Improving the performance of HMM-Based voice conversion using context clustering decision tree and appropriate regression matrix format," in *Proc. of Interspeech*, pp. 2250-2253, 2006.