MINIMUM GENERATION ERROR LINEAR REGRESSION BASED MODEL ADAPTATION FOR HMM-BASED SPEECH SYNTHESIS

Long Qin[1*], *Yi-Jian Wu*[2], *Zhen-Hua Ling*[3], *Ren-Hua Wang*[3], *and Li-Rong Dai*[3]

[1]Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA
[2]Nagoya Institute of Technology, Nagoya, Japan
[3]University of Science and Technology of China, Hefei, P.R.China
lqin@cs.cmu.edu, yjwu@sp.nitech.ac.jp, zhling@ustc.edu, rhw@ustc.edu.cn, lrdai@ustc.edu.cn

ABSTRACT

Due to the inconsistency between the maximum likelihood (ML) based training and the synthesis application in HMM-based speech synthesis, a minimum generation error (MGE) criterion had been proposed for HMM training. This paper continues to apply the MGE criterion to model adaptation for HMM-based speech synthesis. We propose a MGE linear regression (MGELR) based model adaptation algorithm, where the regression matrices used to transform source models to target models are optimized to minimize the generation errors for the input speech data uttered by the target speaker. The proposed MGELR approach was compared with the maximum likelihood linear regression (MLLR) based model adaptation. Experimental results indicate that the generation errors were reduced after the MGELR-based model adaptation. And from the subjective listening test, the discrimination and the quality of the synthesized speech using MGELR were better than the results using MLLR.

Index Terms — Speech synthesis, minimum generation error, linear regression, model adaptation

1. INTRODUCTION

Over the past few years, the HMM-based speech synthesis has been developed [1]. In this method, spectrum, pitch and duration are modeled simultaneously in a unified framework of HMMs [2], and the parameters are generated from HMMs by using the dynamic features [3]. Meanwhile voice characteristics of the synthesized speech can be converted from one speaker to another by applying a model adaptation algorithm, such as maximum likelihood linear regression (MLLR), with a small amount of speech data uttered by the target speaker [4], [5]. Recently, several techniques had been introduced to improve the model adaptation performance, i.e. a context decision tree tying method had been utilized to group source HMMs into classes [6][7] and many other speaker adaptation algorithms, including SMAP, CMLLR, SMAPLR, and CSMAPLR, had been studied [8].

The minimum generation error (MGE) criterion [9] had been proposed to solve the two issues related to the maximum likelihood (ML) based HMM training: the inconsistency between training and synthesis, and the lack of mutual constraints between the statistic and the dynamic features. In this training method, the parameter generation was incorporated into HMM training procedure, and the probabilistic descent (PD) algorithm [10] was applied for parameter updating with the aim to minimize the generation errors for all training data. Furthermore, the MGE criterion had also been applied to the tree-based clustering for context dependent HMMs [11] and the whole HMM training procedure [12].

As the ML criterion was also used for model adaptation in HMM-based speech synthesis, this paper followed up the previous work and applied the MGE criterion for model adaptation. In order to effectively adapt the source models when only a small amount of training data is available, all the source models are firstly grouped into several classes, where the models within one class share the same regression matrix, which is initially estimated under the ML criterion. And then the parameters of the regression matrix are updated using the MGE criterion, i.e. the parameters are optimized to minimize the generation errors for the input speech data uttered by the target speaker.

In the following part of this paper, an overview of MGE criterion is presented in section 2. Section 3 describes the details of the MGELR-based model adaptation method. Section 4 presents the results of experiments including subjective and objective evaluations while section 5 is some discussions on the experiments. At last a final conclusion is provided in section 6.

2. MINIMUM GENERATION ERROR CRITERION

2.1. Parameter generation algorithm

For a given HMM λ and the state sequence $\boldsymbol{Q}$, the parameter generation procedure is to determine the speech parameter vector sequence $\boldsymbol{O}=[\boldsymbol{o}_1^{\mathrm{T}},\boldsymbol{o}_2^{\mathrm{T}},\ldots,\boldsymbol{o}_T^{\mathrm{T}}]^{\mathrm{T}}$ to maximize $P(\boldsymbol{O}\mid\boldsymbol{Q},\lambda)$. In order to generate smooth parameter sequence, the dynamic features including delta and delta-delta coefficients are introduced, i.e.

$$\boldsymbol{o}_t=[\boldsymbol{c}_t^{\mathrm{T}},\Delta\boldsymbol{c}_t^{\mathrm{T}},\Delta^2\boldsymbol{c}_t^{\mathrm{T}}]^{\mathrm{T}}, \tag{1}$$

where $\boldsymbol{c}_t$, $\Delta\boldsymbol{c}_t$ and $\Delta^2\boldsymbol{c}_t$ are the static, delta and delta-delta part of speech parameter vector, respectively. As the dynamic features can be calculated from the static features, then the speech parameter vector sequence $\boldsymbol{O}$ can be rewritten as

$$\boldsymbol{O}=\boldsymbol{W}\boldsymbol{C}, \tag{2}$$

where $\boldsymbol{C}=[\boldsymbol{c}_1^{\mathrm{T}},\boldsymbol{c}_2^{\mathrm{T}},\ldots,\boldsymbol{c}_T^{\mathrm{T}}]^{\mathrm{T}}$. Due to the limited space, the details of $\boldsymbol{W}$ is not given, which can be found in [3].

[*] The work of this paper was done in the iFLYTEK Speech Laboratory in the University of Science and Technology of China.

Under the condition (2), determining $\boldsymbol{O}$ to maximize $P(\boldsymbol{O} \mid \boldsymbol{Q},\lambda)$ is equivalent to determining $\boldsymbol{C}$ to maximize $P(\boldsymbol{O} \mid \boldsymbol{Q},\lambda)$. So by setting $\frac{\partial}{\partial \boldsymbol{C}} \log P(\boldsymbol{O} \mid \boldsymbol{Q},\lambda) = 0$, we obtain

$$\tilde{\boldsymbol{C}} = \left(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{U}^{-1}\boldsymbol{W}\right)^{-1} \boldsymbol{W}^{\mathrm{T}}\boldsymbol{U}^{-1}\boldsymbol{\mu} = \boldsymbol{R}^{-1}\boldsymbol{r} \,, \tag{3}$$

where

$$\boldsymbol{R} = \boldsymbol{W}^{\mathrm{T}}\boldsymbol{U}^{-1}\boldsymbol{W} \,,\ \boldsymbol{r} = \boldsymbol{W}^{\mathrm{T}}\boldsymbol{U}^{-1}\boldsymbol{\mu} \,, \tag{4}$$

and

$$\boldsymbol{\mu} = [\boldsymbol{\mu}_{q_1}^{\mathrm{T}}, \boldsymbol{\mu}_{q_2}^{\mathrm{T}}, ..., \boldsymbol{\mu}_{q_T}^{\mathrm{T}}]^{\mathrm{T}} \,, \tag{5}$$

$$\boldsymbol{U}^{-1} = diag[\boldsymbol{U}_{q_1}^{-1}, \boldsymbol{U}_{q_2}^{-1}, ..., \boldsymbol{U}_{q_T}^{-1}]^{\mathrm{T}} \,, \tag{6}$$

are the mean vector and covariance matrix, respectively.

2.2. Minimum generation error criterion

For a given speech parameter vector sequence $\boldsymbol{O} = \boldsymbol{W}\boldsymbol{C}$, the optimal state sequence $\boldsymbol{Q}_{opt}$ obtained by the Viterbi algorithm was used for parameter generation, and then the generation error $e(\boldsymbol{C},\lambda)$ is defined as the distortion between the original and generated feature vector, i.e. feature distortion. Here the Euclidean distance was adopted to calculate the distortion

$$e(\boldsymbol{C},\lambda) = D(\boldsymbol{C},\tilde{\boldsymbol{C}}) = \left\| \boldsymbol{C} - \tilde{\boldsymbol{C}}_{Q_{opt}} \right\|^2 . \tag{7}$$

With the measure of generation error, the parameter generation process is incorporated into HMM training for calculating the total generation errors for all training data, which is

$$E(\lambda) = \sum_{n=1}^{N} e(\boldsymbol{C}_n, \lambda) \,, \tag{8}$$

where N is the total number of training utterances.

Finally, we defined the object of MGE criterion, which is to optimize the parameters of HMMs so as to minimize the total generation errors

$$\hat{\lambda} = \arg\min E(\lambda) \,. \tag{9}$$

As direct solution for Eq.(9) is mathematically intractable, probabilistic descent (PD) [10] method is adopted for parameter optimization.

3. MGELR-BASED MODEL ADAPTATION

In the MGELR-based model adaptation, the parameter generation is incorporated into the model adaptation procedure to calculate the generation errors and the parameters of the regression matrix are optimized to minimize the generation errors.

3.1. Linear regression based model adaptation

Usually there has no sufficient training data from target speaker to calculate the regression matrix for each HMM, and thus the linear regression (LR) based model adaptation method is employed. In LR-based model adaptation, a context decision tree is used to group all HMMs into several classes [7], in which all the models share the same regression matrix. In order to guarantee that the data associated with each regression class is sufficient to estimate the regression matrix, an empirically derived minimum occupation count is applied. The parameters of the regression matrices are initialized using MLLR and then updated under the MGE criterion. In this paper, only the regression matrix for the mean vector of Gaussian distribution is re-estimated.

For a mean vector $\boldsymbol{\mu}$ of a Gaussian distribution, we define the extended mean vector $\boldsymbol{\xi}$ as $[\varpi, \mu_1, ..., \mu_D]^T$, where ϖ is the offset term for the regression. Under the LR-based model adaptation, the estimate of the adapted mean vector $\hat{\boldsymbol{\mu}}$ can be given by

$$\hat{\boldsymbol{\mu}} = \boldsymbol{M}\boldsymbol{\xi} \,, \tag{10}$$

where $\boldsymbol{M}$ is the $D \times (D+1)$ regression matrix [4], and D is the dimension of speech parameter vector.

From Eq.(3), for the state sequence $\boldsymbol{Q}$ of an adaptation utterance, the generated parameter vector from the adapted HMMs is

$$\begin{aligned} \tilde{\boldsymbol{C}}_Q &= \left(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{U}^{-1}\boldsymbol{W}\right)^{-1} \boldsymbol{W}^{\mathrm{T}}\boldsymbol{U}^{-1}\hat{\boldsymbol{\mu}}_Q \\ &= \left(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{U}^{-1}\boldsymbol{W}\right)^{-1} \boldsymbol{W}^{\mathrm{T}}\boldsymbol{U}^{-1}\boldsymbol{M}_Q\boldsymbol{\xi}_Q \,, \\ &= \boldsymbol{R}^{-1}\boldsymbol{r}_Q \end{aligned} \tag{11}$$

where

$$\boldsymbol{r}_Q = \boldsymbol{W}^{\mathrm{T}}\boldsymbol{U}^{-1}\boldsymbol{M}_Q\boldsymbol{\xi}_Q \,, \tag{12}$$

with

$$\boldsymbol{M}_Q = diag[\boldsymbol{M}_{q_1}, \boldsymbol{M}_{q_2}, \cdots, \boldsymbol{M}_{q_T}] \,, \tag{13}$$

$$\boldsymbol{\xi}_Q = [\boldsymbol{\xi}_{q_1}^{\mathrm{T}}, \boldsymbol{\xi}_{q_2}^{\mathrm{T}}, \cdots, \boldsymbol{\xi}_{q_T}^{\mathrm{T}}]^{\mathrm{T}} \,, \tag{14}$$

are the regression matrix and extended mean vector for the whole state sequence $\boldsymbol{Q}$.

3.2. Parameter updating under MGE criterion

The MGELR-based model adaptation is to minimize the total generation errors for all speech data uttered by the target speaker

$$\hat{\lambda} = \arg\min E(\lambda) = \arg\min \sum_{n=1}^{N} e(\boldsymbol{C}_n, \lambda) \,. \tag{15}$$

where N is the total number of the training utterances.

For a sample C_n in the adaptation training set, the updating rule of the regression matrix parameters is

$$\lambda(n+1) = \lambda(n) - \varepsilon_n \frac{\partial e(\boldsymbol{C}_n;\lambda)}{\partial \lambda}\Big|_{\lambda=\lambda(n)} \,, \tag{16}$$

with respect to the regression matrix $\boldsymbol{M}_Q$, where ε_n is the step size for parameter updating.

From Eq.(7) and Eq.(11), the updating of the regression matrix parameters can be formulated as

$$\frac{\partial e(\boldsymbol{C},\lambda)}{\partial m_{t,i,j}} = 2 \cdot \left(\tilde{\boldsymbol{C}}_Q - \boldsymbol{C}\right)^{\mathrm{T}} \frac{\partial \tilde{\boldsymbol{C}}_Q}{\partial m_{t,i,j}} \,, \tag{17}$$

where

$$\frac{\partial \tilde{\boldsymbol{C}}_Q}{\partial m_{t,i,j}} = \boldsymbol{R}^{-1}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{U}^{-1}\boldsymbol{Z}_\mu\boldsymbol{\xi}_Q \,. \tag{18}$$

Finally,

$$m_{t,i,j}(n+1) = m_{t,i,j}(n) - 2\varepsilon_n(\tilde{\boldsymbol{C}}_Q - \boldsymbol{C})^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{U}^{-1}\boldsymbol{Z}_\mu\boldsymbol{\xi}_Q \,, \tag{19}$$

where $m_{t,i,j}$ is the element of the i th row and the j th column in the regression matrix related to the t-th frame, and

$$\boldsymbol{Z}_\mu = [\underset{1st}{0_{D*(D+1)}^{\mathrm{T}}}, \cdots, \underset{t-th}{K_{D*(D+1)}^{\mathrm{T}}}, \cdots, \underset{T-th}{0_{D*(D+1)}^{\mathrm{T}}}]^{\mathrm{T}} \,, \tag{20}$$

with each element in K_t is $k_{t,x,y} = \begin{cases} 1 & x=i, y=j \\ 0 & else \end{cases}$. (21)

4. EXPERIMENTS

4.1. Experimental conditions

We collected 1000 phonetically balanced sentences of a female speaker and 200 sentences of a male speaker from a Chinese speech database. The speech data of the female speaker was used to train the source HMMs, while the speech data of the male speaker was kept for model adaptation and evaluation. All speech data was sampled at a rate of 16KHz. Spectrum and pitch were obtained by the STRAIGHT analysis [13], and were converted to the line spectral pair (LSP) coefficients and the logarithm F0 respectively. Finally, the feature vector of spectrum and pitch is composed of the 41-order LSP coefficients including the gain coefficient, the logarithm F0, as well as their delta and delta-delta coefficients. We use the 5-state left-to-right no-skip HMMs, in which the spectral part was modeled by single Gaussian distribution and F0 part was modeled by multi-space distribution (MSD) [2]. The duration feature vector is a 5-dimensional vector, corresponding to the 5-state HMMs, and it was modeled by single Gaussian distribution.

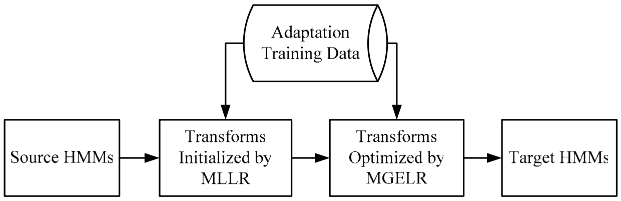


Figure 1: Model adaptation procedure

The whole model adaptation procedure, illustrated in Figure 1, is performed as follows:

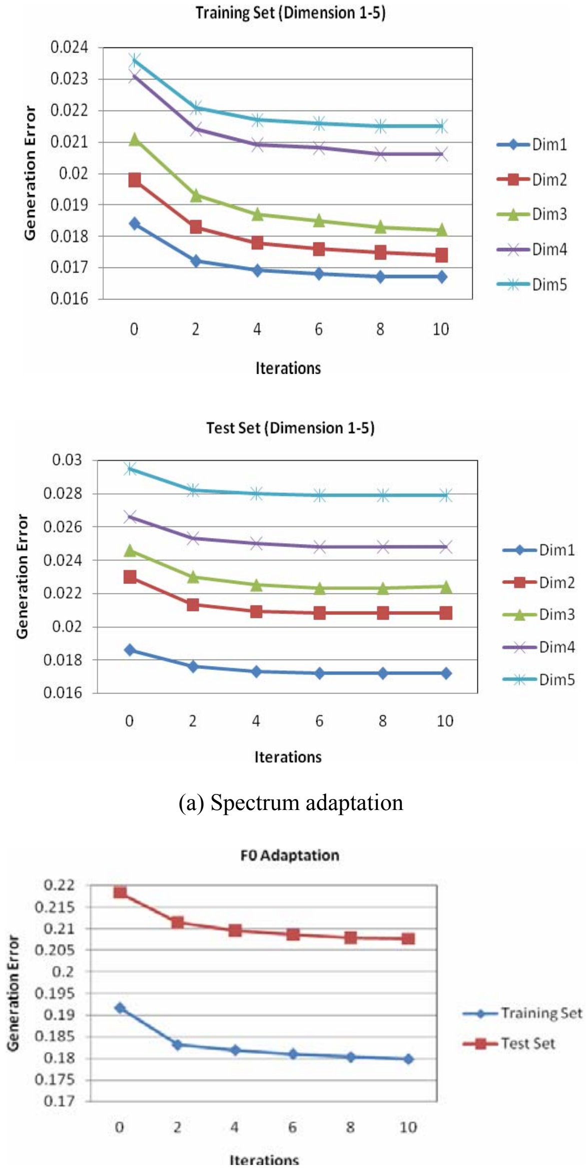
a. Firstly, the regression matrix for the mean vector of Gaussian distribution is initialized by the MLLR algorithm. And the optimal state path for all training data is obtained by the Viterbi algorithm and fixed in the later processes.
b. For each training data, the generation errors are calculated using Eq.(7), where the generated parameter sequence is estimated by Eq.(11).
c. The parameters of the related regression matrices are updated using Eq.(19).
d. The procedure (b) and (c) are performed by several iterations until the generation errors are converged.
e. Finally, from Eq.(10), by applying the updated regression matrices to the source models, we can get the adapted HMMs of the target speaker.

In this paper, both the spectral and F0 regression matrices were optimized using MGELR.

4.2. Experimental results

A female to male conversion was conducted using both the proposed MGELR-based model adaptation and the conventional MLLR method. The source model of the female speaker was trained using 1000 sentences. And 100 sentences randomly selected from the database of the male speaker were used for model adaptation training while the rest 100 sentences were kept as test set.

As the covariance matrix used here was diagonal, the generation errors were calculated independently for each dimension of LSP coefficients. Figure 2(a) shows the convergence property of MGELR-based model adaptation training for several representative dimensions. From the results of close and open test, the spectral MGELR training is converged after about 10 iterations. In the open test, the generation errors reduced about 5% for different dimensions of LSP coefficients after the MGELR-based model adaptation. The convergence property of F0 HMMs adaptation is illustrated in Figure 2(b), which presents the similar results as the spectrum adaptation.



(a) Spectrum adaptation

(b) F0 adaptation

Figure 2: Convergence of MGELR-based model adaptation

To evaluate the effectiveness of the MGELR-based model adaptation, formal subjective listening test was conducted. We compared both the quality and the discrimination of the synthesized speech generated from the adapted HMMs using MGELR and the conventional MLLR method. 10 test sentences, which were not contained in the adaptation training data, were synthesized from the target HMMs estimated by these two methods, respectively. Subjects, including 10 persons, were presented a pair of synthesized speech from different methods in the random order, and then asked which speech sound more natural and which speech sound more like the target speaker. The preference scores are shown in Figure 3. It can be seen that the system performance, especially the quality of the synthetic speech, was improved after applying the MGELR-based model adaptation.

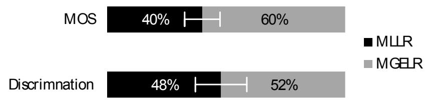


Figure 3: Preference scores

5. DISCUSSION

From Figure 2(a), we can find that the generation error for each dimension of the LSP coefficient is converged after about 10 iterations. In fact, the convergence property for different dimension of the LSP coefficient is not always the same. But we can set different step size for different dimension to make them converge at the same time during the iteration calculation. In addition, as illustrated in Figure 2, the generation errors of the LSP coefficient and F0 converge faster in the open test than those in the close test.

Figure 3 shows that the performance improvement on MOS is much more remarkably than that on the discrimination when applying the MGELR-based model adaptation. But when less adaptation training data, such as 50 sentences, is used, the improvements on MOS and on the discrimination are nearly the same. The synthetic speech using MGELR wins about 55% of the whole test. When only 5 sentences of the target speaker's speech are available, although the generation errors of LSP coefficients and F0 still decline and converge, the informal listening test for the synthesized speech using MLLR and MGELR indicates that the difference between these two methods can be ignored.

6. CONCLUSION

In this paper, we applied the MGE criterion to model adaptation for HMM-based speech synthesis. In order to ensure that all HMMs can be effectively adapted, the linear regression based model adaptation was employed. In the model adaptation training, the regression matrix estimated by MLLR was selected as the initial transform matrix, and then the parameters of regression matrix were re-estimated under the MGE criterion to minimize the generation errors using the PD algorithm. The results of the subjective and objective tests indicate that the system performance including the quality and the discrimination of the synthetic speech is improved with the proposed MGELR-based model adaptation.

Future work is to apply the MGELR-based model adaptation not only to the mean vector but also to the covariance matrix of Gaussian distribution.

7. REFERENCES

[1] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Speech synthesis from HMMs using dynamic features," *Proc. ICASSP-1996*, 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," *Proc. Europspeech-1999*, vol. 5, pp. 2347-2350, Mar. 1999.

[3] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features," *Proc. ICASSP-1995*, pp. 660-663, 1995.

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

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

[6] J. Yamagishi, M. Tachibana, T. Masuko, and T. Kobayashi, "Speaking style adaptation using context clustering decision tree for HMM-based speech synthesis," *Proc. ICASSP-2004*, vol. 1, pp. 5-8, 2004.

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

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

[9] Y.J. Wu, and R.H. Wang, "Minimum generation error training for HMM-based speech synthesis," *Proc. ICASSP-2006*, vol. 1, pp. 89-92, 2006.

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

[11] Y.J. Wu, W. Guo, and R.H. Wang, "Minimum generation error criterion for tree-based clustering of context dependent HMMs," *Proc. ICSLP-2006*, pp. 2046-2049, 2006.

[12] Y.J. Wu, R.H. Wang, and F. Soong, "Full HMM training for minimizing generation error in synthesis," *Proc. ICASSP-2007*, vol. IV, pp. 517-520, 2007.

[13] H. Kawahara, "Restructuring speech representations using a pitch-adaptive time frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sound", *Speech Communication 27*, pp. 187-207, 1999.