FULL HMM TRAINING FOR MINIMIZING GENERATION ERROR IN SYNTHESIS

Yi-Jian Wu [1*] *Ren-Hua Wang* [2] *Frank Soong* [1]

[1] Microsoft Research Asia, Beijing, China
[2] University of Science and Technology of China, Hefei, China
yijwu@microsoft.com, rhw@ustc.edu.cn, frankkps@microsoft.com

ABSTRACT

In maximum-likelihood (ML) based HMM synthesis, the generated trajectory of a sentence in the training set is in general does not reproduce the trajectory of the original one. To overcome this shortcoming, a minimum generation error (MGE) criterion has been previously proposed. In this paper, a complete MGE-based HMM training is introduced, where the MGE criterion is applied to the entire training process, including context-dependent HMM training, context-dependent HMM clustering and clustered HMM training. In this procedure, the HMMs are trained to minimize the generation error of training data, which is in line with the HMM-based synthesis. From the experiments, the quality of synthesized speech is improved after applying the MGE criterion to the whole training process.

Index Terms— speech synthesis, HMM, maximum likelihood, minimum generation error

1. INTRODUCTION

The HMM-based speech synthesis method had been in existence for a decade [1], where spectrum, pitch and duration are modeled simultaneously in a unified framework [2]. In synthesis, all feature parameters are in the maximum likelihood sense generated from trained HMMs with dynamic feature constraints [3]. Under its statistical training framework, it can learn salient statistical properties of speakers, speaking styles, emotions, and etc., which is quite suitable for current requirement of expressive speech synthesis. Due to this, HMM-based speech synthesis gradually becomes popular in research and successful in application [4][5][6].

Although current performance of HMM-based speech synthesis is quite good, there are two issues related to the HMM training [7][8]. The first issue is that the HMM training based on ML criterion is to maximize the likelihood of training data, whereas HMM-based synthesis aims to generate the data as faithful as possible, i.e. the HMMs should be trained to re-generate the training data with minimum errors. Another issue is the ignorance of the constraints between static and dynamic features. In order to solve these two problems, a new criterion, named minimum generation error (MGE), had been proposed for HMM parameter optimization [8]. Furthermore, the MGE criterion was applied to tree-based clustering of context-dependent HMMs [9], where the parameter updating rules are simplified to reduce the computational cost to an acceptable degree.

In this paper, we design a new HMM training procedure by applying the MGE criterion to the whole training process. In this new training procedure, the ML criterion is only used for monophone HMM training and context-dependent HMM initialization. All subsequent training steps, including context-dependent HMM training, context-dependent HMM clustering and clustered HMM training, are based on the MGE criterion. Therefore, the whole HMM training is to minimize the generation error, which is line with the HMM-based synthesis.

The rest of paper is organized as follows. In section 2, the minimum generation error criterion, including MGE-based HMM parameter optimization and MGE-based context-dependent HMM clustering, are introduced. In section 3, we present the MGE-based training procedure. In section 4, the experiments to evaluate the effectiveness of new training procedure are described, and results are presented. Finally, our conclusion and future work are given in section 5.

2. MINIMUM GENERATION ERROR CRITERION

2.1. MGE-based HMM parameter optimization

In MGE-based HMM parameter optimization, the parameter generation is incorporated into the training procedure, and the parameters of HMMs are optimized to minimize the generation error between the synthesized and original data in the training set.

2.1.1. Parameter generation algorithm

For a given HMM λ and the state sequence Q, the parameter generation is to determine the speech parameter vector sequence $O=[o_1^{\mathrm{T}},o_2^{\mathrm{T}},\ldots,o_T^{\mathrm{T}}]^{\mathrm{T}}$ to maximize $P(O\mid\lambda,Q)$. In order to keep the smooth property of the generated parameter sequence, the dynamic features including delta and delta-delta coefficients are used, i.e.

[*] The work of this paper was done at iFly speech laboratrory in University of Science and Technology of China.

$$o_t = [c_t^{\mathrm{T}}, \Delta c_t^{\mathrm{T}}, \Delta^2 c_t^{\mathrm{T}}]^{\mathrm{T}}, \quad (1)$$

where c_t, Δc_t and $\Delta^2 c_t$ are the static, delta and delta-delta part of speech parameter vector, respectively. Then the speech parameter vector sequence O can be rewritten as

$$O = WC, \quad (2)$$

where $C = [c_1^{\mathrm{T}}, c_2^{\mathrm{T}}, ..., c_T^{\mathrm{T}}]^{\mathrm{T}}$. Due to limited space, here the details of W, which can be found in [2][8], is not given.

Under the condition (2), maximizing $P(O \mid \lambda, Q)$ respect to O is equivalent to that with respect to C. By setting $\frac{\partial}{\partial C} \log P(O \mid Q, \lambda) = 0$, we obtain

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

where

$$R = W^{\mathrm{T}} U^{-1} W, \; r = W^{\mathrm{T}} U^{-1} \mu, \quad (4)$$

and μ and U^{-1} are the mean and covariance matrix of Q, respectively.

2.1.2. HMM parameter optimization

In MGE criterion, the first important thing is to define the generation error. For a given speech parameter vector sequence $O = WC$, the optimal state sequence Q_{opt} obtained by the Viterbi algorithm was used for parameter generation, and then the generation error $\ell(C, \lambda)$ is defined as the distance between the original vector sequence C and the generated one $\tilde{C}$. For simplification, the Euclidean distance was adopted, i.e.

$$\ell(C, \lambda) = D(C, \tilde{C}) = \left\| C - \tilde{C} \right\|^2 = \sum_{t=1}^{T} \left\| c_t - \tilde{c}_t \right\|^2. \quad (5)$$

Under the definition of generation error, the parameters of the HMMs were optimized to minimize the generation error by using the Generalized Probabilistic Descent (GPD) algorithm [12]. For a sample C_n in the training set, the updating procedure for the HMM parameters is

$$\lambda(n+1) = \lambda(n) - \varepsilon_n \frac{\partial \ell(C_n; \lambda)}{\partial \lambda}\Big|_{\lambda = \lambda(n)}, \quad (6)$$

where ε_n is the step size for parameter update.

From Eq. (3) and Eq. (5), the updating rule for the mean parameter can be formulated as

$$\frac{\partial \ell(C, \lambda)}{\partial \mu_{i,j}} = 2 \cdot \left(\tilde{C} - C\right)^{\mathrm{T}} \frac{\partial \tilde{C}}{\partial \mu_{i,j}}, \quad (7)$$

where

$$\frac{\partial \tilde{C}}{\partial \mu_{i,j}} = R^{-1} W^{\mathrm{T}} U^{-1} Z_{\mu}. \quad (8)$$

Finally,

$$\mu_{i,j}(n+1) = \mu_{i,j}(n) - 2\varepsilon_n (\tilde{C}_n - C_n)^{\mathrm{T}} R^{-1} W^{\mathrm{T}} U^{-1} Z_{\mu}, \quad (9)$$

where $\mu_{i,j}$ is the jth dimension of the mean vector of the state model related to the i-th frame, and $Z_{\mu} = [0, ..., 0, 1_{i \times M + j}, 0, 0, ...0]^{\mathrm{T}}$, where M is the dimension of the speech parameter vectors.

Similarly, the updating rule for the covariance parameter can be formulated as

$$v_{i,j}(n+1) = v_{i,j}(n) - 2\varepsilon_n \left(\tilde{C} - C\right)^{\mathrm{T}} R^{-1} W^{\mathrm{T}} Z_v \left(\mu - W\tilde{C}\right), \quad (10)$$

where $v_{i,j} = 1/\sigma_{i,j}^2$ is the covariance parameter corresponding to $\mu_{i,j}$, and $Z_v = diag[0, ..., 0, 1_{i \times M + j}, 0, 0, ...0]$.

2.2. MGE-based context-dependent HMM clustering

Since ML criterion is used for tree-based clutering of context-dependent HMMs, where the question which yields the largest likelihood increase is selected to split the tree node, and the parameters of the clustered model after splitting was estimated by ML criterion. Therefore, we apply the MGE criterion for the tree-based clustering of context-dependent HMMs. As the computational cost by directly applying the MGE criterion for HMM clustering is very expensive [9], in order to regulate the computational cost, the MGE criterion for HMM clustering is simplified and appropriate strategy is designed by combining the MGE criterion with the ML criterion to select the optimal question for node splitting in decision tree growing.

2.2.1. Simplified parameter updating rules

From Eq. (9), the parameter updating rules of MGE criterion is time-cosuming due to the sequential update mod and calculation of R^{-1}. To solve the first problem, we considered all training samples in a batch mode and used the following updating rule

$$\lambda_{update} = \lambda_{old} - \varepsilon \sum_{n=1}^{N} \frac{\partial \ell(C_n; \lambda)}{\partial \lambda}\Big|_{\lambda = \lambda_{old}}, \quad (11)$$

where is N the total number of the training samples.

In order to avoid calculation of R^{-1}, we need to simplify the updating rules in Eq. (9). Considering that WW^T is a quasi-diagonal and diagonal dominant matrix, we made an approximation as

$$WW^T \approx aI, \quad (12)$$

where I is an unit matrix, and a is a constant number for normalization. Without loss of generality, we assume $a = 1$.

For the mean vector, we apply this approximation to the updating rule and obtained

$$\frac{\partial \tilde{C}}{\partial \mu} = R^{-1} W^T U^{-1} Z_{\mu} \approx R^{-1} W^T U^{-1} W W^T Z_{\mu} = W^T Z_{\mu}. \quad (13)$$

and

$$\frac{\partial \ell(C_n; \lambda)}{\partial \mu} \approx 2 \cdot \left(\tilde{C}_n - C_n\right)^T W^T Z_{\mu} = \sum_{t=S_n}^{E_n} 2 \cdot \left(\tilde{o}_{n,t} - o_{n,t}\right). \quad (14)$$

where S_n and E_n are the start and end time of O_n.

By setting the step size to $\varepsilon = \frac{1}{2N_{total}}$, where $N_{total} = \sum_{n=1}^{N} \sum_{t=S_n}^{E_n} 1$ is the total frames of the training sample related to current updated model, the updating rule in Eq. (11) can be simplified as

$$\begin{aligned}\mu_{update} &= \mu_{old} - \frac{1}{N_{total}} \sum_{n=1}^{N} \sum_{t=S_n}^{E_n} \left(\tilde{o}_{n,t} - o_{n,t} \right), \\ &= \mu_{old} - \left(\mu_{gen} - \mu_{orig} \right)\end{aligned} \tag{15}$$

where $\mu_{gen} = \sum_{n=1}^{N} \sum_{t=S_n}^{E_n} \tilde{o}_{n,t}$ and $\mu_{orig} = \sum_{n=1}^{N} \sum_{t=S_n}^{E_n} o_{n,t}$. From this equation, the mean parameters of HMMs are updated by the difference between the mean of generated speech parameters and the mean of original speech parameters.

2.2.2. Strategy for splitting question selection

Since direct incorporation of MGE criterion for question selection is computationally expensive, we adopt an efficient strategy by combining the MGE criterion with the ML criterion to select the optimal question for node splitting. In this procedure, the ML criterion was used to pre-select a subset of the questions, and then the simplified MGE criterion was applied to select the best splitting question from the subset.

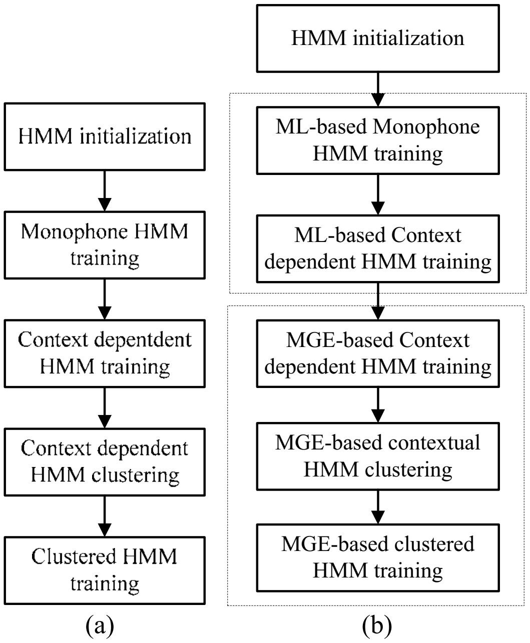
HMM initialization
Monophone HMM training
Context depentdent HMM training
Context dependent HMM clustering
Clustered HMM training
(a)
HMM initialization
ML-based Monophone HMM training
ML-based Context dependent HMM training
MGE-based Context dependent HMM training
MGE-based contextual HMM clustering
MGE-based clustered HMM training
(b)

Fig. 1 HMM training procedure
(a) ML-based HMM training procedure
(b) MGE-based HMM training procedure

3. HMM TRAINING PROCEDURE

3.1. ML-based HMM training procedure

Fig 1(a) shows the original ML-based HMM training procedure. In this procedure, all the training steps are based on the ML criterion, including monophone HMM training, context-dependent HMM training and clustering, and clustered HMM training.

3.2. MGE-based HMM training procedure

Due to the issues related to the ML-based HMM training, the MGE criterion had been proposed. In our previous work [8][9], we only applied the MGE criterion to the corresponding training step and other training parts are still based on the ML criterion. Here, we design a new training procedure by applying the MGE criterion to the whole training procedure, which is shown in Fig. 1(b).

In this new training procedure, only the monophone HMM training and the first iteration of context-dependent HMM training are based on the ML criterion, which can be regarded as the initialization of context-dependent HMMs. The subsequent training steps, including context-dependent HMM training, context-dependent HMM clustering and clustered HMM training, are all based on the MGE criterion. Therefore, the HMMs are trained to minimize the generation error of training data, which is in line with the HMM-based synthesis application.

4. EXPERIMENTS

4.1. Experimental conditions

The training data consists of 1,000 phonetically balanced Chinese sentences, including 25,096 syllable initials and 29,942 syllable finals. Speech signal were sampled at a rate of 16kHz. The acoustic features, consists of F0 and 24th-order LSP coefficients, were obtained by STRAIGHT [10] filter shifted every 5ms. Feature vector consists of F0 and spectrum parameter vector. Spectrum parameter vector consists of 40 LSP coefficients with the gain, delta and delta-delta coefficients. F0 parameter vector consists of logarithm of F0, its delta and delta-delta coefficients. A 5-state left-to-right, no skip HMM was used. Regarding to the Chinese characteristics, the context feature and question set were designed for context-dependent HMM modeling and tree-based clustering.

We evaluated the effect of MGE-based training procdure by comparing the performance of the HMMs trained by three different procedures, which are

1) MLT: ML-based training procedure in Fig. 1(a),
2) PMGT: the modified training procedure based on Fig. 1(a), only clustered HMM training based on MGE criterion, and other training steps are based on ML criterion.
3) MGET: MGE-based training procedure in Fig. 1(b),

It should be noted that the stopping threshold for context-dependent HMM clustering had been carefully designed to make the number of the clustered models obtained by MGE criterion comparable to that obtained by ML criterion. Here, the MGE criterion is applied to both spectrum and F0 parameters.

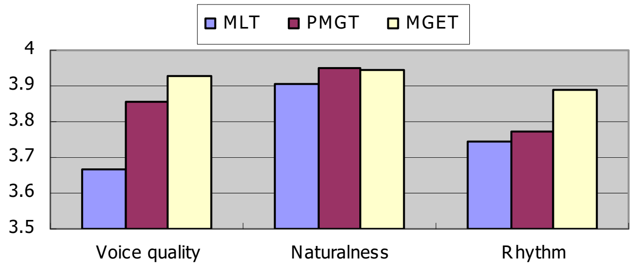

Fig. 2 Evaluation results of three training procedure

4.2. Results

In the informal subjective listening test, we compared synthesized speech from the HMMs trained by the three different procedures, and found that not only voice quality, but also intonation and rhythm of synthesized speech, are different. In order to understand the effect of the MGE criterion in depth, we need to evaluate these differences separately.

Finally, the formal subjective listening test was conducted to evaluate the performance of MGE-based training procedure. In the tests, 50 test sentences were synthesized from the HMMs trained by the three training procedure. Six subjects were presented synthesized speech samples generated from the three different models, and asked to give the score in two aspects, including voice quality and naturalness. The scoring scale is 1(bad) to 5(good). The evaluation results are shown in Fig. 1.

As seen in Fig. 1, the synthesized speech from the MGET models is better than that from the MLT and PMGT models in voice quality aspect. From the aspect of naturalness, the results of the MGET and PMGT models are similar and better than the MLT models. By analyzing the synthesized samples from the MGET and PMGT models, we found that the rhythm of synthesized speech was improved after applying MGE criterion to the whole training process. However, at the same time it introduced some tonal errors in the synthesized speech, which due to the rough measurement of the generation error for F0 parameters in MGE-based clustering [11]. Therefore, in order to see the positive impact of the MGE-based training procedure, we ignored the test sentences with the tonal-error syllables, and re-evaluated the synthesized speech in the aspect of rhythm. From the results in Fig. 1, the rhythm of synthesized speech from the MGET is better than other two models, which indicates that the rhythm of synthesized speech are improved by applying the MGE criterion to the whole training procedure.

5. CONCLUSIONS

In this paper, we present the MGE training criterion for HMM parameter optimization and context-dependent HMM clustering. Furthermore, a new HMM training procedure based on MGE criterion is proposed, where the MGE criterion is applied to the whole training process, including context-dependent HMM training, context-dependent HMM clustering and clustered HMM training. Experimental results of MGE training show that the quality of synthesized speech is improved over original ML based training.

Future work is to improve the MGE-based training procedure for F0 parameters of MSD-HMM.

6. 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 Eurospeech, pp. 2347-2350, 1999.

[3] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, “Speech parameter generation algorithms for HMM-based speech synthesis,” in Proc. of ICASSP, pp. 1315-1318, 2000.

[4] K. Tokuda, H. Zen, and A. W. Black, “An HMM-based speech synthesis system applied to English,” 2002 IEEE Speech Synthesis Workshop, California, Sep. 11-13, 2002.

[5] H. Zen, and T. Toda, “An Overview of Nitech HMM-based Speech Synthesis System for Blizzard Challenge 2005”, in Proc. of Eurospeech, pp. 93-96, 2005.

[6] Z. H. Ling, Y. J. Wu, Y. P. Wang, L. Qin and R. H. Wang, “USTC System for Blizzard Challenge 2006 – an Improved HMM-based Speech Synthesis Method,” in Interspeech 2006 satellite meeting, Blizzard Challenge 2006

[7] K. Tokuda, H. Zen and T. Kitamura, “Trajectory modeling based on HMMs with the explicit relationship between static and dynamic features,” in Proc. of Eurospeech, 2003, pp. 865-868

[8] Y. J. Wu and R. H. Wang, “Minimum generation error training for HMM-based speech synthesis,” in Proc. of ICASSP 2006, vol. 1, pp. 89-92, May. 2006.

[9] Y. J. Wu, W. Guo and R. H. Wang, “Minimum generation error criterion for tree-based clustering of context dependent HMMs,” in Proc. of Interspeech 2006, pp., Sep. 2006.

[10] H. Kawahara, I. Masuda-Katsuse and A. deCheveigne, “Restructuring speech representations using 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

[11] Y. J. Wu, “Research on HMM-based speech synthesis”, Ph. D. thesis (in Chinese), University of Science and Technology of China, 2006

[12] J. R. Blum, “Multidimensional stochastic approximation methods,” Ann. Math. Stat, vol. 25, pp.737-744, 1954