Minimum Generation Error Criterion for Tree-based Clustering of Context Dependent HMMs

Yi-Jian Wu, Wu Guo, Ren-Hua Wang.

iFly speech laboratory, Department of Electronic Engineering and Information Science
University of Science and Technology of China, Hefei, Anhui, China
jasonwu@mail.ustc.edu.cn, guowu@mail.ustc.edu.cn, rhw@ustc.edu.cn

Abstract

Due to the inconsistency between HMM training and synthesis application in HMM-based speech synthesis, the minimum generation error (MGE) criterion had been proposed for HMM training. This paper continues to apply the MGE criterion for tree-based clustering of context dependent HMMs. As directly applying the MGE criterion results in an unacceptable computational cost, the parameter updating rules of the MGE criterion are simplified to rapidly update the parameters of clustered models, and an appropriate strategy by combining the MGE criterion with the ML criterion is designed to select the optimal question for tree node splitting. From the experiment results, the quality of synthetic speech was improved after applying the MGE criterion for HMM clustering.

Index Terms: speech synthesis, HMM, minimum generation error

1. Introduction

In recent years, the HMM-based speech synthesis [1][2] had been proposed. Under its trainable framework, speech synthesis system can be constructed automatically and rapidly for the given speech data by a direct way [4] or using a model adaptation technique [3]. Moreover, its framework is general with less dependency for speakers, speaking styles, emotions, and even languages, which is quite suitable for current requirement of expressive speech synthesis. Due to these, HMM-based speech synthesis gradually becomes popular both in research and application [4][5].

Although current performance of HMM-based speech synthesis is quite good, there are two issues in the HMM training [6][7]. The first issue is related to the inconsistency between the Maximum Likelihood (ML) based HMM training and synthesis application of HMMs. Another issue is the ignorance of the constraints between static and dynamic features. Actually, after the feature extraction, the static and dynamic features are both used as "static" features in HMM training, whereas the constraints between static and dynamic features are considered in parameter generation. In order to solve these two problems, a new criterion, named Minimum Generation Error (MGE) [6], had been proposed for HMM training. In this training method, the parameter generation was incorporated into the training procedure, and the Generalized Probabilistic Descent (GPD) algorithm [10] was applied for parameter updating with the aim to minimize the generation errors between the training and generated data.

As the ML criterion was also used for tree-based clutering of context dependent HMMs, this paper followed up the previous work and applied the MGE criterion for HMM clustering. The simplest way is to directly replace the ML criterion by MGE criterion for calculation of splitting score and updating the parameters of HMM after splitting. However, the computational cost of this direct way is unacceptable even for the off-line training procedure. In order to reduce the computational cost, the parameter updating rules of MGE criterion was simplified, and an appropriate strategy by combing the MGE criterion with the ML criterion was designed to select the optimal question for tree node splitting, where the ML criterion was used to efficiently pre-select the optimal subset of questions, and then the MGE criterion was applied for selecting the optimal one from the subset.

This paper is organized as follows. In section 2, we briefly review the parameter generation algorithm and the MGE criterion for HMM training. In section 3, we present the MGE-based HMM clustering in detail, including the simplified parameter updating rules and the strategy by combing the MGE criterion with the ML criterion to select the optimal splitting question. Next, the experiments to evaluate the performance of the MGE-based HMM clustering are shown in Section 4. Finally, our conclusion and future work is given in Section 5.

2. MGE-based HMM training

In MGE-based HMM training [6], the parameter generation is incorporated into the training procedure for generation error calculation, and the parameters of the HMMs are optimized to minimize the generation error by using the GPD algorithm.

2.1. Parameter generation algorithm

For a given HMM λ and the state sequence Q, the parameter generation is to determine the speech parameter vector sequence $O = [o_1^{\mathrm{T}}, o_2^{\mathrm{T}}, ..., o_T^{\mathrm{T}}]^{\mathrm{T}}$ 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.

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

where c_t, Δc_t and $\Delta^2 c_t$ are the static, delta and delta-delta part of speech parameter vector, respectively. As the dynamic features can be calculated from the static features, 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 the limited space, the details of W is not given here, which can be found in [8].

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

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

where

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

and

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

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

are the mean and covariance matrix, respectively.

2.2. Minimum generation error training

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 here to calculate the distance, i.e.

$$\ell(C,\lambda) = D(C,\tilde{C}) = \left\|C-\tilde{C}\right\|^2 = \sum_{t=1}^{T}\left\|c_t-\tilde{c}_t\right\|^2\,. \tag{7}$$

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 [10]. For a sample C_n in the training set, the updating rule of the HMM parameters is

$$\lambda(n+1) = \lambda(n) - \varepsilon_n\frac{\partial\ell(C_n;\lambda)}{\partial\lambda}\Big|_{\lambda=\lambda(n)}\,, \tag{8}$$

where ε_n is the step size for parameter updating.

From Eq. (3) and Eq. (7), the updating rule for the mean parameter can be formulated as

$$\frac{\partial\ell(C,\lambda)}{\partial\mu_{i,j}} = 2\cdot\left(\tilde{C}-C\right)^{\mathrm{T}}\frac{\partial\tilde{C}}{\partial\mu_{i,j}}\,, \tag{9}$$

where

$$\frac{\partial\tilde{C}}{\partial\mu_{i,j}} = R^{-1}W^{\mathrm{T}}U^{-1}Z_{\mu}\,. \tag{10}$$

Finally,

$$\mu_{i,j}(n+1) = \mu_{i,j}(n) - 2\varepsilon_n(\tilde{C}_n - C_n)^{\mathrm{T}}R^{-1}W^{\mathrm{T}}U^{-1}Z_{\mu}\,, \tag{11}$$

where $\mu_{i,j}$ is the j th 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

$$v_{i,j}(n+1) = v_{i,j}(n) - 2\varepsilon_n\left(\tilde{C}-C\right)^{\mathrm{T}}R^{-1}W^{\mathrm{T}}Z_v\left(\mu - W\tilde{C}\right), \tag{12}$$

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

3. MGE-based HMM clustering

Comparing the enormous possible combination of context features, which is used to characterize the contextual variation of acoustic features in HMM-based speech synthesis, the training data is spare for context dependent HMM training. Therefore, the tree-based clustering [11] for the context dependent HMMs was applied to improve the robustness of HMM training. In current framework, the ML criterion was applied for the tree-based clustering, where the question with 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.

Due to the inconsistency between the ML criterion and the synthesis application, we followed up the MGE-based HMM training and continued to apply the MGE criterion for the tree-based clustering of context dependent HMMs. The simplest way is to directly replace the ML criterion by MGE criterion for calculation of splitting score and updating the parameters of HMM after splitting. However, the computational cost of this direct way is unacceptable, even the HMM clustering is an off-line task. In this section, we simplified the MGE criterion for HMM clustering and designed appropriate strategy by combining the MGE criterion with the ML criterion to select the optimal question for node splitting.

3.1. Simplified parameter updating rules

From Eq. (11) and Eq. (12), the parameter updating rules of MGE criterion is time-cosuming due to the sample-by-sample updating manner and the calculation of R^{-1} . To solve the first problem, we considered all training samples in the same time and used the following updating rule

$$\lambda_{update} = \lambda_{old} - \varepsilon\sum_{n=1}^{N}\frac{\partial\ell(C_n;\lambda)}{\partial\lambda}\Big|_{\lambda=\lambda_{old}}\,, \tag{13}$$

where is N the total number of the training sample related to λ . Then the updating iteration for convergence can be reduced.

From the parameter updating rules in Eq. (11) and Eq. (12), the most computational cost is related to the calculation of R^{-1} . Although we have approximated R^{-1} to a quasi-diagonal matrix with a diagonal bandwidth B , which can largely reduce the computation cost [6], it was still unacceptable for HMM clustering, where the parameter updating should be performed for each splitting attempt. Due to this, we need to simplify the parameter updating rules to avoid the calculation of R^{-1} .

Considering that WW^T is a quasi diagonal matrix and the diagonal elements are larger than other elements, we made an approximation as

$$WW^T \approx aI\,, \tag{14}$$

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

$$\frac{\partial\tilde{C}}{\partial\mu} = R^{-1}W^TU^{-1}Z_{\mu} \approx R^{-1}W^TU^{-1}WW^TZ_{\mu} = W^TZ_{\mu}\,. \tag{15}$$

and

$$\frac{\partial \ell(C_n;\lambda)}{\partial \mu} \approx 2\cdot\left(\tilde{C}_n - C_n\right)^T W^T Z_\mu = 2\cdot\left(\tilde{O}_n - O_n\right)^T Z_\mu = \sum_{t=S_n}^{E_n} 2\cdot\left(\tilde{o}_{n,t} - o_{n,t}\right) \quad (16)$$

where S_n and E_n are the start and end time of the sample O_n. Then the updating rule in Eq. (13) can be simplified as

$$\mu_{update} = \mu_{old} - 2\varepsilon\sum_{n=1}^{N}\sum_{t=S_n}^{E_n}\left(\tilde{o}_{n,t} - o_{n,t}\right), \quad (17)$$

In order to make the meaning of the updating rule clearer, let's denote $N_{total} = \sum_{n=1}^{N}\sum_{t=S_n}^{E_n} 1$ as the total frames of the training sample related to current updated model, and set the step size $\varepsilon = \frac{1}{2N_{total}}$. Then Eq. (17) can be rewritten as

$$\mu_{update} = \mu_{old} - \frac{1}{N_{total}}\sum_{n=1}^{N}\sum_{t=S_n}^{E_n}\left(\tilde{o}_{n,t} - o_{n,t}\right) = \mu_{old} - (\mu_{gen} - \mu_{orig}), \quad (18)$$

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 the models are updated by the difference between the mean of generated speech parameters and the mean of original speech parameters.

Similarly, the updating rule for covariance parameters can be simplified as

$$\sigma^2_{update} = \sigma^2_{old} - \frac{2\varepsilon}{\sigma^2_{old}}\sum_{n=1}^{N}\sum_{t=S_n}^{E_n}\left(\tilde{o}_{n,t} - o_{n,t}\right)\left(\tilde{o}_{n,t} - \mu_{n,t}\right), \quad (19)$$

In fact, the above simplification is equivalent to approximate Eq. (3) in parameter generation to

$$\tilde{C} = \left(W^{\mathrm{T}}U^{-1}W\right)^{-1}W^{\mathrm{T}}U^{-1}\mu \approx W^{\mathrm{T}}\mu. \quad (20)$$

From this equation, we can see the meaning of the approximation is to loosen the influence of adjacent model, and the parameter of each frame is generated by using the static and dynamic features of current model. It should be noted the parameter generation used in the MGE-clustering is still under Eq. (3), whereas Eq. (20) is only implied in the updating rules.

3.2. Efficient strategy for splitting question selection

When splitting one node in the tree-based clustering, we attempt each question in the question set, to split the node and calculate the splitting score. Then the optimal question with the best splitting score was selected to split this node. Under the MGE criterion, the generation error reduction was used as the splitting score. For each splitting attempt, the generation error was calculated by re-generating all training sample related to the node, as the model parameters of the split node were updated. That means it needs to generate N_{qs} times for the training sample related to the node, where N_{qs} is the size of the question set. In order to characterize all possible contextual rules for acoustic variation, the size of the question set used in current framework is larger than 1000. Due to this, the procedure of tree-based clustering is computationally very expensive.

Considering that the selection of splitting question based on ML criterion can be performed efficiently, and the results of preliminary experiment showed the best question selected by MGE criterion also has large likelihood increase, we designed an appropriate strategy for selecting of the splitting question by combining the MGE with ML criterion, which is show in Fig. 1. 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 of questions. As the subset size of pre-selected questions reflects the lowest likelihood rank of the best question selected by MGE criterion, and the times to generate the training samples, it should be carefully designed to balance the accuracy and the efficiency.

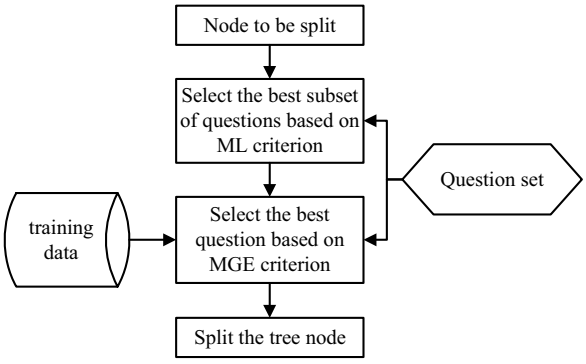

Figure 1 *Procedure for selecting the best splitting question*

4. Experiments and Discussions

4.1. Experimental conditions

The training data consists of 1000 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, including F0 and 24-order LSP coefficients, were obtained by STRAIGHT [9] filter with a 5ms shift. Feature vector consists of F0 and spectrum parameter vector. Spectrum parameter vector consists of 25 LSP coefficients with the gain, delta and delta-delta coefficients. F0 parameter vector consists of a logarithm of F0, its delta and delta-delta coefficients. The 5-state left-to-right with no skip HMM structure was used. Regarding to the Chinese characteristics, the context feature and question set were designed for contextual HMM modeling and tree-based clustering.

We evaluated the effect of MGE criterion by comparing the performance of the HMMs clustered by MLE and MGE criterion. The training procedure is performed as follows. The context dependent HMMs were firstly initialized by the results of the MLE-based training, and then the ML-based or MGE-based clustering was applied. Finally, the ML-based embedded training was performed for re-estimation of the clustered HMMs. It should be noted that the stopping threshold for splitting had been carefully designed to make the number of the clustered

models obtained by MGE criterion is comparable to that obtained by ML criterion. Here, the MGE-based clustering was only applied for spectrum parameters.

4.2. Results and discussions

Table 1 shows the results of the ML-based and MGE-based clustering. Under the equivalent number of the clustered models, the generation errors of training data after the MGE-based clustering is smaller than that obtained after the ML-based clustering, which is coincident with the aim of the MGE criterion. Also, it is reasonable that the ML-based clustering is outperformed the MGE-based clustering in likelihood scale.

Table 1. *Comparison of the results of MGE-based and ML-based HMM clustering*

	Clustered model Num.	Generation errors	Likelihood
ML	3508	$2.293*10^5$	$7.580*10^8$
MGE	3370	$1.900*10^5$	$7.504*10^8$

From the informal perception experiment, the improvement on the synthetic speech is not distinct after using MGE criterion for HMM clustering. By analyzing the training procedure, the ML-based embedded training was performed for re-estimation of the clustered HMMs, which maybe weaken the effect of the MGE-based clustering. Therefore, we tried to directly use the clustered HMMs before re-estimation to synthesize the speech, and found the synthetic speech is distinctly better than that synthesized by the ML-trained HMMs.

Finally, formal subjective listening test was conducted to evaluate the effectiveness of the MGE-based HMM clustering. In the tests, 50 test sentences, which were not contained in the training data, were synthesized from the HMMs trained by three different procedures, which are

1) ML: the ML-based training procedure
2) NMGE: the new procedure with the MGE-based HMM clustering, introduced in Section 4.1
3) PMGE: the procedure similar to 2), without the ML-based embedded training for the clustered HMMs

The subjects, including 6 persons, were presented a set of synthesized speech from the above three different models, and rank them according to the synthetic quality. Table 2 shows the average rank for each model. It can be seen the HMMs obtained by MGE-clustering without ML-based re-estimation had the best rank, and the rank value shows the speech synthesized by this model is significant better than the speech synthesized by other two models. From this point of view, the MGE-based HMM clustering is outperform the ML-based HMM clustering in the synthetic quality.

Table 2. *Average rank of three models*

	PMGE	NMGE	ML
Average rank	1.113	2.264	2.623

5. Conclusions and future works

In this paper, we applied the MGE criterion for tree-based clustering of context dependent HMMs. In order to reduce the computational cost of MGE-based clustering, the parameter updating rules of MGE criterion are simplified to rapidly update the HMM parameters of split node, and an appropriate strategy by combing the MGE with ML criterion is designed to select the optimal question for splitting the tree node. From the experiment results, the quality of synthetic speech was improved after applying the MGE criterion for HMM clustering, even without the re-estimation of the clustered HMMs by MGE-based training.

Future work is to apply the MGE-based clustering to MSD-HMM for F0 parameter updating. Furthermore, we will design a MGE-based training procedure by combining the MGE-based HMM training with MGE-based HMM clustering.

6. Acknowledgements

The authors would like to thank Dr. Frank K. Soong for helpful discussion and suggestion. This work was partially supported by the National Natural Science Foundation of China under grant number 60475015.

7. References

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

[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, 1999, vol. 5, pp. 2347-2350

[3] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, “Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR,” in Proc. of ICASSP, May 2001, pp. 805-808

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

[5] Y. J. Wu, and R. H. Wang, “HMM-based trainable speech synthesis for Chinese”, (in Chinese), accepted by Journal of Chinese Information Processing, 2006

[6] 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

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

[9] 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

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

[11] J. Odell, “The use of context in large vocabulary speech recognition,” Dissertation of doctor degree, 1995