MINIMUM GENERATION ERROR TRAINING FOR HMM-BASED SPEECH SYNTHESIS

Yi-Jian Wu *Ren-Hua Wang*

iFly speech laboratory, University of Science and Technology of China

E-mail: jasonwu@mail.ustc.edu.cn, rhw@ustc.edu.cn

ABSTRACT

In HMM-based speech synthesis, there are two issues critical related to the MLE-based HMM training: the inconsistency between training and synthesis, and the lack of mutual constraints between static and dynamic features. In this paper, we propose minimum generation error (MGE) based HMM training method to solve these two issues. In this method, an appropriate generation error is defined, and the HMM parameters are optimized by using the generalized probabilistic descent (GPD) algorithm, with the aims to minimize the generation errors. From the experimental results, the generation errors were reduced after the MGE-based HMM training, and the quality of synthetic speech is improved.

1. INTRODUCTION

The Hidden Markov Model (HMM) had been popularly used for speech recognition, and made a significant progress. In the last decade, the HMM has been applied for speech synthesis application [1][2][3], and HMM-based speech synthesis was proposed [4]. In this method, spectrum, pitch and duration are modeled simultaneously in a unified framework of HMMs [5], and the parameters are generated from HMMs by using the dynamic features [4]. In order to synthesize speech with various voice characteristics, the MLLR adaptation algorithm had been applied to transform HMM parameters with limited target speech data [6], e.g. 5 sentences.

Although the current performance of HMM-based speech synthesis is quite good, there are two issues in the HMM training. The first issue is related to the inconsistency between the training and application of the HMM. In general, the aim of HMM-based speech synthesis is to generate the speech (acoustic parameters) as close to the nature speech as possible. However, the conventional HMM training method is adopted from speech recognition [7], which is based on Maximum Likelihood Estimation (MLE) criteria, i.e. it is not designed for speech synthesis application. 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 the "static" features in HMM training, whereas the constraints between static and dynamic features are considered in parameter generation.

In order to resolve above two issues, a trajectory model had been introduced into HMM-based speech synthesis [8][9], in which the HMM training is performed under the constraints between static and dynamic features. Although the new training criterion implied the minimization of the error between training and generated data, the HMM training is still under the MLE framework, which cannot actually resolve the first issue.

In this paper, a new HMM training criterion, named Minimum Generation Error (MGE), was proposed to train the HMM. By incorporating the parameter generation into the training procedure, the inconsistency between training and generation was eliminated, and the constraints between static and dynamic features are considered in HMM training. With the definition of generation error between training and generated data, the Generalized Probabilistic Descent (GPD) algorithm [10] was applied for parameter updating with the aim to minimize the generation error.

This paper is organized as follows. In Section 2, we briefly review the HMM-based speech synthesis framework and the parameter generation algorithm. In Section 3, we present the MGE-based HMM training method in detail, including the definition of generation error and the parameters updating schedule. Next, the experiments to evaluate the performance of the MGE-based HMM training are shown in Section 4. Finally, our conclusion and future work is given in Section 5.

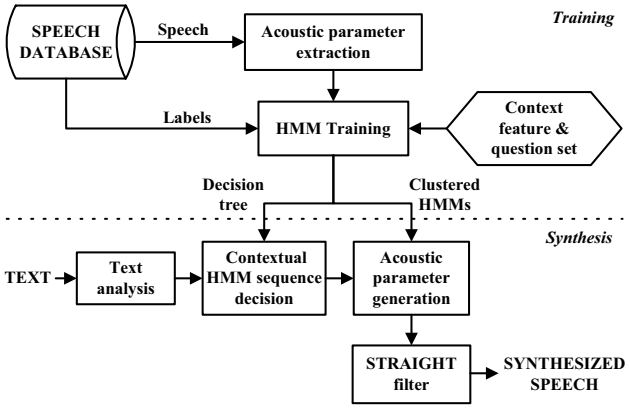


Fig. 1 HMM-based speech synthesis system

2. HMM-BASED SPEECH SYNTHEISIS SYSTEM

Figure 1 shows an overview of the HMM-based speech synthesis system, which consists of two stages, the training and synthesis stage.

In the training stage, the output vector of the HMM consists of spectrum part and F0 part. In our system, the spectrum part consists of Line Spectral Pair (LSP), their delta and delta-delta coefficients. The F0 part consists of a logarithm of F0, its delta and delta-delta coefficients. The spectrum part is modeled by continuous distribution HMMs and the F0 part is modeled by multi-space probability distribution HMMs [11]. In order to capture the variations caused by different contextual features, the contextual dependent HMM are used, and the tree-based clustering technique is applied for spectrum, F0 and duration to improve the robustness.

In the synthesis stage, the input text is firstly converted to a context-dependent label sequence, and the decision trees generated in the training stage are used to choose the appropriate clustered state HMMs for each label. Then the parameter generation algorithm is used to generate the acoustic parameter sequence, including spectrum and F0. Finally, the speech is synthesized from the generated spectrum and F0 data using the STRAIGHT filter.

3. MINIMUM GENERATION ERROR TRAINING

In this section, we first review the parameter generation algorithm [4], and then introduce the Minimum Generation Error (MGE) based HMM training method with a generation error definition. In this method, the parameter generation is incorporated into the HMM training procedure for generation error calculation, and the parameters of the HMMs are optimized to minimize the generation error by using the GPD algorithm.

3.1. 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, which are calculated as

$$\Delta c_t = \sum_{\tau=-L_-^{(1)}}^{L_+^{(1)}} w^{(1)}(\tau) c_{t+\tau} , \tag{1}$$

$$\Delta^2 c_t = \sum_{\tau=-L_-^{(2)}}^{L_+^{(2)}} w^{(2)}(\tau) c_{t+\tau} . \tag{2}$$

Then the speech parameter vector can be rewritten as

$$o_t = [c_t^{\mathrm{T}}, \Delta c_t^{\mathrm{T}}, \Delta^2 c_t^{\mathrm{T}}]^{\mathrm{T}} , \tag{3}$$

and

$$O = WC , \tag{4}$$

where

$$C = [c_1^{\mathrm{T}}, c_2^{\mathrm{T}}, ..., c_T^{\mathrm{T}}]^{\mathrm{T}} , \tag{5}$$

$$W = [w_1, w_2, ..., w_T]^{\mathrm{T}} , \tag{6}$$

$$w_t = [w_t^{(0)}, w_t^{(1)}, w_t^{(2)}] , \tag{7}$$

$$\begin{aligned} w_t^{(n)} = [& \underset{1st}{0_{M\times M}}, ..., 0_{M\times M}, \underset{(t-L_-^{(n)})\text{-}th}{w^{(n)}(-L_-^{(n)}) I_{M\times M}}, ..., \\ & \underset{t\text{-}th}{w^{(n)}(0) I_{M\times M}}, ..., w^{(n)}(L_+^{(n)}) I_{M\times M}, \underset{(t+L_+^{(n)})\text{-}th}{0_{M\times M}}, \\ & ..., \underset{T\text{-}th}{0_{M\times M}}]^T, \quad n = 0,1,2 \end{aligned} \tag{8}$$

M is the dimension of speech parameter vector c_t.

By setting $\dfrac{\partial}{\partial C} \log P(O \mid Q, \lambda) = 0$, we obtain

$$W^{\mathrm{T}} U^{-1} W C = W^{\mathrm{T}} U^{-1} \mu, \tag{9}$$

where

$$\mu = [\mu_{q_1}^{\mathrm{T}}, \mu_{q_2}^{\mathrm{T}}, ..., \mu_{q_T}^{\mathrm{T}}]^{\mathrm{T}} , \tag{10}$$

$$U^{-1} = diag[U_{q_1}^{-1}, U_{q_2}^{-1}, ..., U_{q_T}^{-1}]^{\mathrm{T}} , \tag{11}$$

are the mean and covariance matrix, respectively.

3.2. Generation error definition

In general, the aim of HMM-based speech synthesis is to generate the speech (acoustic parameters) as close to the nature speech as possible, i.e. the generation error is as small as possible. From this point, the first important thing is to define an appropriate objective measure for generation error.

For a state sequence Q of a given speech parameter vector sequence $O = WC$, the generated vector sequence $\tilde{C}(\lambda, Q)$ can be calculated by equation (9). We assume the distance between original and generated data as $D(C, \tilde{C}(\lambda, Q))$. Without loss of generality, we denote $\tilde{C}(\lambda, Q)$ as $\tilde{C}$. Here the Euclidean distance was adopted to calculate $D(C, \tilde{C})$, i.e.

$$D(C, \tilde{C}) = \left\| C - \tilde{C} \right\|^2 = \sum_{t=1}^{T} \left\| c_t - \tilde{c}_t \right\|^2 \tag{12}$$

It should be noted that the distance measure can be replaced by other measure which is more suitable for the real application. The following equations can be reformulated accordingly.

The posterior probability $P(Q \mid \lambda, O)$ can be used to "weight" the generation error of C to define a corresponding lose function for all possible paths Q:

$$\ell(C, \lambda) = \sum_{All\ Q} P(Q \mid \lambda, O) D(C, \tilde{C}) \tag{13}$$

If we directly calculate the generation error using equation (13), the computational cost is excessive large. In practice, we can use the representative N-best path to calculate the generation error, and equation (13) can be rewritten as

$$\ell(C, \lambda) = \frac{1}{K} \cdot \sum_{n=1}^{N_{best}} P(Q_n \mid \lambda, O) D(C, \tilde{C}_n) \tag{14}$$

where K is the constant number for normalization. For simplification, here we only used the optimal state sequence (1-best path) obtained by the Viterbi algorithm. Then the definition of generation error can be simplified as

$$\ell(C, \lambda) = D(C, \tilde{C}(\lambda, Q_{opt})) \tag{15}$$

where Q_{opt} is the optimal state sequence for O.

3.3. Minimum generation error criterion

Under the definition of generation error, we incorporated the parameter generation into the HMM training procedure for generation error calculation. In order to minimize the generation errors, the GPD algorithm is applied and shown below in detail. This new training method, aiming at minimizing the generation error, is called Minimum Generation Error (MGE) based HMM training.

For the given definition of generation error $\ell(C, \lambda)$, the GPD algorithm is to minimize the empirical generation error

$$L(\lambda) = \frac{1}{N} \sum_{i=1}^{N} \ell(C_i, \lambda) = \int \ell(C, \lambda) p_N(C) dC \tag{16}$$

according to an iterative procedure

$$\lambda_{n+1} = \lambda_n - \varepsilon_n S_n \nabla \ell(C_n, \lambda) \Big|_{\lambda=\lambda_n} . \tag{17}$$

where S_n is a positive definite matrix, C_n is the n th training sample used in the sequential training process, and ε_n is a sequence of positive numbers that satisfies the conditions:

$$i) \sum_{n=1}^{\infty} \varepsilon_n \to \infty , \ ii) \sum_{n=1}^{\infty} \varepsilon_n^2 < \infty . \tag{18}$$

A more detailed introduction and discussion of GPD algorithm can be found in the literature [10].

For a sample C_n in the training set, the updating rule of the HMM parameters is

$$\lambda(n+1) = \lambda(n) - \varepsilon_n \frac{\partial \ell(C_n;\lambda)}{\partial \lambda}\Big|_{\lambda=\lambda(n)} . \quad (19)$$

Under the definition of generation error in equation (15), we obtain

$$\frac{\partial \ell(C,\lambda)}{\partial \lambda} = 2\cdot\left(\tilde{C}-C\right)^{\mathrm{T}} \frac{\partial \tilde{C}}{\partial \lambda} \quad (20)$$

From equation (9), the generated vector sequence is written as

$$\tilde{C} = \left(W^{\mathrm{T}}U^{-1}W\right)^{-1} W^{\mathrm{T}}U^{-1}\mu = R^{-1}r \quad (21)$$

where

$$R = W^{\mathrm{T}}U^{-1}W\ ,\ r = W^{\mathrm{T}}U^{-1}\mu \quad (22)$$

For the mean parameter $\mu_{i,j}$, i.e. the j th dimension of the mean vector of the state model related to the i th frame, equation (20) can be written as

$$\frac{\partial \ell(C,\lambda)}{\partial \mu_{i,j}} = 2\cdot\left(\tilde{C}-C\right)^{\mathrm{T}} \frac{\partial \tilde{C}}{\partial \mu_{i,j}}, \quad (23)$$

where

$$\frac{\partial \tilde{C}}{\partial \mu_{i,j}} = R^{-1}W^{\mathrm{T}}U^{-1}Z_{\mu} . \quad (24)$$

where $Z_{\mu} = [0,...,0,1_{i\times M+j},0,0,...0]^{\mathrm{T}}$. Finally, the updating rule for the mean vector is

$$\mu_{i,j}(n+1) = \mu_{i,j}(n) - 2\varepsilon_n(\tilde{C}_n - C_n)^{\mathrm{T}}R^{-1}W^{\mathrm{T}}U^{-1} . \quad (25)$$

Similarly, we denote $v_{i,j} = 1/\sigma_{i,j}^2$, where $\sigma_{i,j}^2$ is the covariance parameter corresponding to $\mu_{i,j}$. We derivate $R\cdot\tilde{C} = r$ respect to $v_{i,j}$, and obtain

$$\frac{\partial \tilde{C}}{\partial v_{i,j}} = R^{-1}\left(\frac{\partial r}{\partial v_{i,j}} - \frac{\partial R}{\partial v_{i,j}}\tilde{C}\right) = R^{-1}W^{\mathrm{T}}Z_v\left(\mu - W\tilde{C}\right). \quad (26)$$

where $Z_v = diag[0,...,0,1_{i\times M+j},0,0,...0] = Z_{\mu}Z_{\mu}^{\mathrm{T}}$. Finally, the updating rule for the covariance parameter is

$$v_{i,j}(n+1) = v_{i,j}(n) - 2\varepsilon_n\left(\tilde{C}-C\right)^{\mathrm{T}}R^{-1}W^{\mathrm{T}}Z_v\left(\mu - W\tilde{C}\right). \quad (27)$$

3.5. Discussion

As a probability measure, the HMM generally has some original constraints, e.g. $\sigma > 0$. In order to maintain the constraints and normalize the step size during parameter updating, we should take some parameter transformations as follows:

$$U \to \hat{U} = \log(U), \quad (28)$$

$$\mu \to \hat{\mu} = \mu U^{-1} . \quad (29)$$

The similar updating rules can be formulated correspondingly.

It should be noted that the computational cost of the MGE-based training, in which the most computation cost is associated to the calculation of R^{-1} in equation (25) and (27). To calculate R^{-1} directly, we need $O(T^3M^3)$, and it becomes $O(T^3M)$ when U_t are diagonal. If we consider that R is a quasi-diagonal matrix, R^{-1} can also be approximated to a quasi-diagonal matrix with a diagonal bandwidth B, which can be regarded as the influence range of the current state. Usually, 50 ~ 100 is large enough for B. Finally, the computational complexity reduces to $O(T^2MB)$, which is acceptable as the HMM training is an offline task.

As we have indicated, there are two issues related to the conventional HMM training, including the inconsistency between the training and application of the HMM and the ignorance of the constraints between static and dynamic features. In the MGE-based HMM training, these two issues are both resolved. By using the minimum generation error criterion, the HMM training aims to minimize the generation error, where the inconsistency between the training and application of the HMM is eliminated. Furthermore, as the constraints between static and dynamic features are considered in the parameter generation, they are also considered in the HMM training by incorporating the parameter generation into the training procedure.

4. EXPERIMENTS

4.1. Experimental conditions

The training data consists of 1000 phonetically balanced Chinese sentences, including 25,096 initials and 29,942 finals. The test data consists of 800 sentences, including 17,860 initials and 21,389 finals. Regarding to the Chinese characteristics, the context feature and question set were designed for contextual HMM modeling and tree-based clustering.

Speech signal were sampled at a rate of 16KHz. The acoustic features, including F0 and 24-order LSP coefficients, were obtained by STRAIGHT [12] 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.

We evaluate the effect of MGE-based training by comparing the performance of the HMMs trained by MLE and MGE criterion. The MGE-based training is performed as follows:

a. Firstly, the HMMs were initialized by the results of MLE-based training, and the optimal state path for all data were obtained by the Viterbi algorithm and fixed in the later processes.
b. For each training data, the generation errors were calculated and the related HMM parameters were updated using equation (25) and (27).
c. The procedure (b) were performed by several iterations until the generation errors are converged.

It should be noted that the clustered HMMs are used to initialize the HMMs for MGE-based training, i.e. the MGE criterion are only applied to the clustered HMM training. Furthermore, only spectrum parameters are updated in current training procedure. In future work, F0 parameters will be updated, and we will apply the MGE criterion to the context-dependent HMM training.

4.2. Experimental results

As the covariance matrix is diagonal, the generation errors were calculated independently for each dimension of LSP coefficients. Figure 2 shows convergence property of MGE-based HMM training for several representative dimensions. From the results of close and open test, the MGE-based HMM training are converged

after 10~20 iterations. In the open test, the generation errors reduced about 8~15% for different dimensions of LSP coefficient after the MGE-based HMM training.

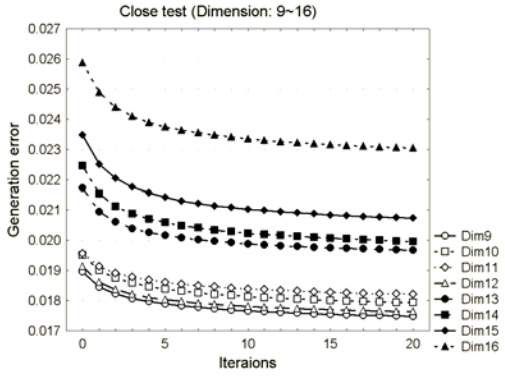
From the informal perception experiment, the synthetic speech becomes clearer and the unnaturalness is alleviated after the MGE-based HMM training. To evaluate the effectiveness of the MGE-based HMM training, formal subjective listening test was conducted. We compared the quality of synthetic speech generated from the HMMs trained with MLE and MGE criterion. In the tests, 50 test sentences, which were not contained in the training data, were synthesized from the HMMs trained by MLE and MGE criterion, respectively. Subjects, including 6 persons, were presented a pair of synthesized speech from different models in random order, and asked which speech sound more natural.

Figure 3 shows the preference scores. It can be seen that the quality of synthetic speech are improved after applying the MGE-based HMM training.

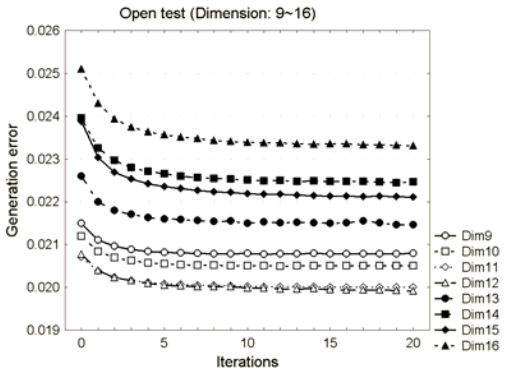
5. CONCLUSION & FUTRUE WORK

In this paper, we proposed minimum generation error (MGE) based HMM training method for HMM-based speech synthesis. In this method, an appropriate generation error is defined, and the HMM parameters are optimized by using the GPD algorithm, with the aims to minimize the generation errors. In the MGE-based HMM training, two issues in the MLE-based HMM training, including the inconsistency between the training and application of the HMM and the ignorance of the constraints between static and dynamic features, had been resolved. From the experimental results, the generation errors were reduced after the MGE-based HMM training, and the quality of synthetic speech are improved.

Future work is to apply the MGE criterion to MSD-HMM for F0 parameter updating. Furthermore, we will apply it to the contextual dependent HMM training, and design a similar criterion for the tree-based clustering.



(a) close test



(b) open test

Fig. 2 Convergence of MGE-based HMM training

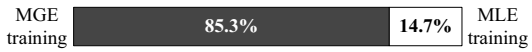


Fig. 3 Preference score

ACKNOWLEDGEMENT

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

REFERENCES

[1] A. Ljolje, J. Hirschberg, and J.P.H. van Santen, "Automatic speech segmentation for concatenative inventory selection," in Progress in speech synthesis, J.P.H van Santen, R.W. Spout, J.P. Olive, and J. Hirshberg, Eds. Springer-Verlag, 1997

[2] R.E. Donovan and E.M. Eide, "The IBM trainable speech synthesis system," in Proc. of ICSLP, 1998, vol. 5, pp. 1703-1706

[3] X. Huang, A. Acero, H. Hon, Y. Ju, J. Liu, S. Merdith, and M. Plumpe, "Recent improvements on Microsoft's trainable text-to-speech system – Whistler," in Proc. of ICASSP, 1997, pp. 959-962

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

[5] 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

[6] 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

[7] L.R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proc. IEEE, 1989, vol. 77, pp. 257-286.

[8] 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

[9] H. Zen, K. Tokuda, and T. Kitamura, "An introduction of trajectory model into HMM-based speech synthesis," 5th ISCA Speech Synthesis Workshop, 2004, pp. 191-196

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

[11] 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, 1999, pp. 229-232

[12] 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," Speech Communication, vol. 27, pp. 187-207, 1999