COMBINING GAUSSIAN MIXTURE MODEL WITH GLOBAL VARIANCE TERM TO IMPROVE THE QUALITY OF AN HMM-BASED POLYGLOT SPEECH SYNTHESIZER

Javier Latorre, Koji Iwano, Sadaoki Furui*

Tokyo Institute of Technology, Department of Computer Science
2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan
E-mail: {latorre,iwano,furui}@furui.cs.titech.ac.jp

ABSTRACT

This paper proposes a new method to calculate the cepstral coefficients for an HMM-based synthesizer. It consists in a direct maximization of the log-likelihood function of a Gaussian mixture model using a gradient ascent algorithm. The method permits to integrate efficiently the Global Variance factor with a Gaussian mixture acoustic model.The perceptual experiments confirmed that these two factors produce significant improvements on the speech quality, which are independent from each other. By using the proposed method, it is possible to get the benefits of both factors. This paper also proposes a 2-class model for the Global Variance that discriminates between consonants and vowels. Such 2-class Global Variance model produces more stable cepstral coefficients than the single-class one.

Index Terms— HMM-based speech synthesis, Gaussian mixture , Global Variance, polyglot.

1. INTRODUCTION

The advantages of the HMM-based synthesis algorithm against concatenative synthesis methods are the simplicity of its implementation, the smoothness of the voice it produces and its flexibility to generate new voices. On the other hand, its main disadvantage has always been its audio quality, equivalent to that of a vocoder. Several modifications have been proposed to improve the quality of the HMM-based synthesizer. Roughly, they can be divided into those that improve the source excitation model and those that improve the modeling of the F0 and vocal tract coefficients, such as Hidden semi-Markov Models and Trajectory Models. The parameter generation algorithm however, remained unaltered. It consisted in maximizing the total log-likelihood of a sequence of HMMs with respect to the observation vector, considering the constraints defined by the relationship between the dynamic and static features of that observation vector. Recently, a modification to this algorithm was proposed by Toda and Tokuda [5]. They added to the parameter generation algorithm a new constraint that takes in consideration the variance of the static part of the observation vector along the whole utterance. The addition of this factor improves greatly the quality of the synthetic speech. However, since it introduces a quadratic term, the maximization cannot be solved efficiently using the standard synthesis algorithms [6] for models using Gaussian mixtures.

*Presently with the Corporate Research & Development Center, Toshiba Corporation.

2. STANDARD ALGORITHM

The standard HMM-based speech synthesis algorithm consist in finding the observation vector $\mathbf{O}$ and the sequence of states $\mathbf{q}$ that maximize the output probability of a hidden Markov model λ, which represents the phones of the sentence to be synthesized. This probability can be decomposed into two terms

$$P[\mathbf{q}, \mathbf{O}|\lambda] = P[\mathbf{q}|\lambda] \cdot P[\mathbf{O}|\mathbf{q}, \lambda] \tag{1}$$

Therefore, its maximization can be done in two steps, first with respect to the states sequence $\mathbf{q}$, and second with respect to the observation vector $\mathbf{O}$, conditioned on the already calculated $\mathbf{q}$.

2.1. Determination of the sequence of states

The typical structure of the HMMs used for speech synthesis is left-to-right with no skip. Consequently, determining the states sequence is equivalent to finding the duration of each state d_s, i.e., the number of frames assigned to each state. The state sequence $\mathbf{q} = \{q_1, q_2, \ldots, q_S\}$ is the one that maximizes

$$\log P(\mathbf{q}|\lambda, T) = \sum_{s=1}^{S} \log(P_s(d_s)) \tag{2}$$

for the model λ of the sentence to be synthesized, given the constraint

$$T = \sum_{s=1}^{S} d_s \tag{3}$$

where T is the total number of frames assigned to λ, S the total number of states, d_s the number of frames assigned to

the state s and $P_s(d_s)$ is the probability of staying in state q_s during d_s frames. The duration probability can be defined by self-transition probabilities P_s or by density functions. In our implementation, we used the first option together with an external module that estimates the duration T_p assigned to each phone p of the sentence [1]. For this purpose, each submodel λ_p that corresponds to a phone p is considered as an independent model with total duration T_p. If the proportions among the durations d_s of each state $q_s \in \lambda_p$ are assumed to be independent of T_p, they can be calculated as:

$$d_s = T_p \frac{1/(1-P_s)}{\sum_{q_s \in \lambda_p} 1/(1-P_s)} \tag{4}$$

2.2. Estimation of the observation vector

For a fixed sequence of states $\mathbf{q}$, the probability $P[\mathbf{O}|\mathbf{q},\lambda]$ depends only on the static features vector $\mathbf{c}$ as:

$$\log P[\mathbf{O}|\mathbf{q},\lambda] = \sum_{t=1}^{T} \log b_{s_t}(\mathbf{o}_t) + Const \tag{5}$$

where T is the total number of frames, $b_{s_t}(\mathbf{o}_t)$ is the output probability of the state s at frame t for the observation vector $\mathbf{o}_t = [\mathbf{c}_t^\top, \mathbf{\Delta c}_t^\top]^\top$ and $Const$ is a constant that represents the log-likelihood of the state transition and initial state probabilities. In order to obtain a smooth transition of the synthesized parameters from one frame to another, the relationship between static $\mathbf{c}$ and dynamic features $\Delta\mathbf{c}$:

$$\Delta\mathbf{c}_t = \sum_{l=-L}^{L} \omega(l)\mathbf{c}_{t+l} \tag{6}$$

have to be considered, with ω the weighting vector. If the output probabilities $b_{s_t}(\mathbf{o}_t)$ are modeled by single Gaussians, the derivate of Eq.(5) with respect to $\mathbf{c}$ considering the constraints of Eq.(6) produces the linear equation:

$$\frac{\partial \log P[\mathbf{O}|\mathbf{q},\lambda]}{\partial \mathbf{c}} = 2\mathbf{W}^\top \mathbf{\Sigma}^{-1}(\mathbf{W}\mathbf{c} - \boldsymbol{\mu}) \tag{7}$$

where $\mathbf{\Sigma}$ is a matrix formed by the covariance matrices of the states assigned to each frame t, $\boldsymbol{\mu}$ is the vector of mean values of the states assigned to each frame, and $\mathbf{W}$ is a transformation matrix that summarizes Eq.(6) so that $\mathbf{O} = \mathbf{W}\mathbf{c}$.

In the case that the output probabilities $b_{s_t}(\mathbf{o}_t)$ are modeled by a mixture of M_{s_t} multivariate Gaussians distributions, it is common to use the approximation:

$$\begin{aligned} \log b_s(\mathbf{o}) = & \quad \log\left(\textstyle\sum_{m=1}^{M_s} \kappa_s^m \mathcal{N}_s^m(\mathbf{o})\right) \\ \simeq & \quad \log\max(\kappa_s^1 \mathcal{N}_s^1(\mathbf{o}), \ldots, \kappa_s^{M_s} \mathcal{N}_s^{M_s}(\mathbf{o})) \end{aligned} \tag{8}$$

In this way, each Gaussian mixture is equivalent to a substate with the mixture weight as transition probability. For this approximation, Tokuda et al. proposed several algorithms that maximize Eq.(5) in a time-recursive manner [6].

3. GLOBAL VARIANCE

One of the problems of HMM-based speech synthesis is that the variances of the generated parameters are much lower than the variances of the original ones. As a result, the synthetic speech sounds usually over-smooth. Recently, an effective method to alleviate this problem was proposed [5]. It consists in considering not only the constraints between dynamic and static features, but also the constraints given by the probability of the Global Variance (GV) of the static features. After adding the GV constraint to Eq.(5), the function to be maximized becomes:

$$\log P[\mathbf{O}|\mathbf{q},\lambda,\lambda_{GV}] = \sum_{t=1}^{T} \log b_{s_t}(\mathbf{o}_t) + \log P[\mathbf{v}(\mathbf{c}),\lambda_{GV}] \tag{9}$$

where λ_{GV} is the probability model of the Global Variance $\mathbf{v}(\mathbf{c})$ which is defined as

$$\mathbf{v}(\mathbf{c}) = [v^1, \cdots, v^d, \cdots, v^D]^\top \tag{10}$$

$$v^d = \frac{1}{T}\sum_{t=1}^{T}(c_t^d)^2 - \left(\frac{\sum_{t=1}^{T} c_t^d}{T}\right)^2 \tag{11}$$

4. GRADIENT ASCENDENT BASED ALGORITHM

If the Global Variance probability is modeled by a single Gaussian distribution with mean value $\boldsymbol{\mu}_v$ and covariance $\mathbf{\Sigma}_{vv}$, the derivate of its log-likelihood with respect to the d dimension of the static feature vector at time t, c_t^d, is

$$\begin{aligned} \frac{\partial \log P[\mathbf{v}(\mathbf{c}),\lambda_{GV}]}{\partial c_t^d} = & \quad -\tfrac{2}{T}\textstyle\sum_{r=1}^{D} s_v(r,d)(v^d - \mu_v(d)) \\ & \cdot\left(c_t^d - \tfrac{1}{T}\textstyle\sum_{\tau=1}^{T} c_\tau^d\right) \end{aligned} \tag{12}$$

where $s_v(r,d)$ is the $(r,d)^{th}$ element of the inverse of the covariance matrix $\mathbf{\Sigma}_{vv}$. This derivate is non-linear. Therefore, the maximization of Eq.(9) has to be solved by means of a gradient ascent algorithm. For the EM algorithm proposed in [6], this non-linearity implies that the auxiliary Q-function has to be maximized with a gradient descent loop at each M-step, reducing thus the efficiency of this algorithm.

However, if the sequence of states $\mathbf{q}$ is determined independently of the observation vector $\mathbf{O}$, as described in section 2.1, the approximation of Eq. (8) is not needed. This permits to obtain the vector of static features by directly maximizing Eq. (9) in a single loop. A direct maximization of Eq. (5) was first proposed in [4]. However, in that approach the log-likelihood of each frame was maximized independently of the neighbor frames. For this reason, although the global log-likelihood improved after each iteration, the final convergence was not guaranteed.

In this paper we propose a direct maximization of Eq. (9) based on the gradient ascent algorithm. The convergence to a local maxima is usually achieved after 20-25 iterations.

The derivate of Eq.(5) with respect to c_t^d is

$$\begin{aligned}\frac{\partial \log P[\mathbf{O}|\mathbf{q},\lambda]}{\partial c_t^d} &= \sum_{t=1}^{T} \frac{1}{b_{s_t}(\mathbf{o}_t)} \frac{\partial b_{s_t}(\mathbf{o}_t)}{\partial c_t^d} \\ &= \sum_{l=-L}^{L} \frac{1}{b_{s_{(t+l)}}(\mathbf{o}_{(t+l)})} \frac{\partial b_{s_{(t+l)}}(\mathbf{o}_{(t+l)})}{\partial c_t^d}\end{aligned} \tag{13}$$

where $2L$ is the number of frames considered to calculate the dynamic features. For diagonal covariance matrices and considering the constraints given by Eq.(6), the derivate with respect to c_t^d of the output observation probability $b_{s_{(t+l)}}(\mathbf{o}_{(t+l)})$ modeled by a mixture of Gaussians is

$$\begin{aligned}\frac{\partial b_{s_{(t+l)}}(\mathbf{o}_{(t+l)})}{\partial c_t^d} = \sum_{m=1}^{M_{s_{(t+l)}}} \Big(\kappa_{s_{(t+l)}}^m \frac{\exp(\mathcal{F}_{s_{(t+l)}}^m(\mathbf{o}_{(t+l)}))}{\sqrt{(2\pi)^D |\mathbf{\Sigma}_{s_{(t+l)}}^m|}} \\ \cdot \frac{\partial \mathcal{F}_{s_{(t+l)}}^m(\mathbf{o}_{(t+l)})}{\partial c_t^d} \Big)\end{aligned} \tag{14}$$

where

$$\mathcal{F}_{s_t}^m(\mathbf{o}_t) = -\frac{1}{2}(\mathbf{o}_t - \boldsymbol{\mu}_{s_t}^m)^\top (\mathbf{\Sigma}_{s_t}^m)^{-1}(\mathbf{o}_t - \boldsymbol{\mu}_{s_t}^m) \tag{15}$$

Consequently

$$\begin{aligned}\frac{\partial \mathcal{F}_{s_{(t+l)}}^m(\mathbf{o}_{(t+l)})}{\partial c_t^d} = -\Big(\chi(l) \frac{(c_t^d - \mu_{s_{(t+l)}}^m(d))}{\sigma_{s_{(t+l)}}^m(d)} \\ + \omega(-l) \frac{(\Delta c_{(t+l)}^d - \mu_{\Delta s_{(t+l)}}^m(d))}{\sigma_{\Delta s_{(t+l)}}^m(d)} \Big)\end{aligned} \tag{16}$$

where $\kappa_{s_t}^m$, $\boldsymbol{\mu}_{s_t}^m$ and $\mathbf{\Sigma}_{s_t}^m$ are the weight, mean and variance of the m-th Gaussian component $N_{s_t}^m$ of the state s at frame t; $\mu_{s_t}^m(d)$ and $\sigma_{s_t}^m(d)$ are the mean and variance for the d dimension of the static features of the m-th Gaussian component at state s_t; $\mu_{\Delta s_t}^m(d)$ and $\sigma_{\Delta s_t}^m(d)$ are the mean and variance for the d dimension of the dynamic features at state s_t, and $\chi(l) = 1$ for $l = 0$ and 0 otherwise.

5. EXPERIMENT

In order to test our approach, we have performed two subjective evaluations. The first one analyzed the effect of increasing the number of Gaussians in the model, and the second one the effect of using the GV factor. For the first evaluation, stimuli generated from speaker-adapted acoustic models with 1, 4 and 16 Gaussian mixtures using the GV term were compared with each other. In the second experiment, stimuli synthesized with and without the GV term from speaker-adapted models with 4 and 16 Gaussians mixtures were compared. Both evaluations were conducted using pair tests, where subjects were asked to select the stimuli with overall better speech quality. For each test, 10 native Japanese subjects evaluated a set of 15 pairs of stimuli in Japanese, presented in random order. All the subjects evaluated the same set of pairs.

Since the evaluation was conducted in the framework of an HMM-based speaker adaptable polyglot synthesizer [2], average voice polyglot models were trained first with data from 50 speakers, 10 for each one of the training languages: Spanish, German, French, Russian and Japanese. These average models were then adapted to one speaker for each language to create the models used in the evaluation. The feature vector consists of 25 mel-cepstral coefficients and their delta, calculated from a 16 ms Blackman window with a 5 ms shift. The training database was GlobalPhone [3].

5.1. Prosody estimation and source excitation

The prosody of the stimuli was generated using the quantification method proposed in [1].

The source excitation model is based on the mixed-excitation algorithm [7]. The source-excitation was trained as a second stream of an HMM model, with the first stream being the Mel-Cepstral coefficients, so that both streams were synchronized. The parameters of the mixed-excitation stream consist of the noise gain, its delta, the voicing strength of the bands 0-1KHz, 1-2KHz, 2-4KHz and 4-6KHz and their deltas. After the training of context-dependent models, each stream was clustered independently. The 2-streams tied models were then retrained using a 2-Gaussian mixture for each stream. Finally, the streams were detached into two independent single stream tied models. This 2-Gaussian model is the one used later to synthesize the mixed-excitation parameters. The models with the Mel-Cepstral stream were further refined incrementing the number of Gaussians to create the average voice models mentioned in the previous section. The synthesis of the mixed-excitation parameters from the HMM was done in the same way as for the Mel-cepstral parameters, but without the GV factor. In the 6-8KHz band, a voicing strength value of half of the one obtained for the 4-6KHz band was assigned.

5.2. Multi-class Global Variance

As Fig.1 shows, the average GVs of consonants and vowels are different, especially for low cepstral coefficients, and they are also different from the total average GVs obtained for a single-class model. In some informal experiments, it was found that when the a single-class GV model was applied to all the phones, the synthesized voice was often unstable, i.e. screeches were often produced and the voiced seemed to tremble. By creating a separated GV model for consonants and vowels, these problems almost completely disappeared. In general, the application of the GV factor to vowels was found to produce a stronger improvement of the speech quality than when applied to consonants. In our implementation, no GV term was added to the silences.

In some preliminary tests, the GV term was also modeled by a mixture of Gaussians. However, it did not produce any noticeable improvement over the single Gaussian 2-class model described above.

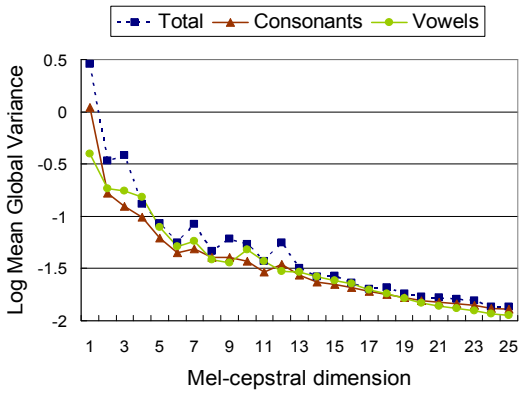

Fig. 1. Logarithm of the mean Global Variance vector: consonants, vowels and total

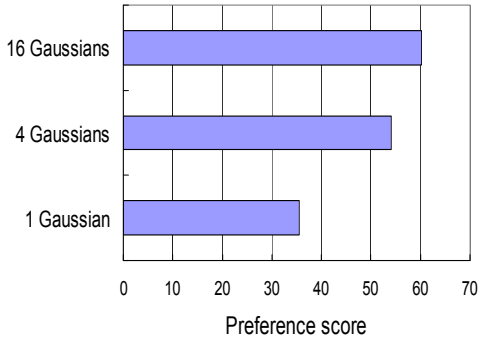

Fig. 2. General preferences for models with different number of mixtures

6. RESULTS

The general preferences of the fist subjective evaluation are shown in Fig.2. As expected, the more Gaussians are used the more accurate the representation of the vocal tract parameters and therefore, the higher total preference. However, the improvement produced by a higher number of Gaussians tends to saturate, therefore, the difference between the 16 and the 4 Gaussians model were not significant.

The preferences for models with and without the GV factor are shown in Fig.3. It can be seen that regardless of the number of Gaussians in the model, the stimuli synthesized using the GV term were clearly preferred.

7. CONCLUSIONS

A direct maximization of the log-likelihood function for the HMM-based synthesis algorithm using a gradient ascent method was proposed. This approach permits to integrate efficiently

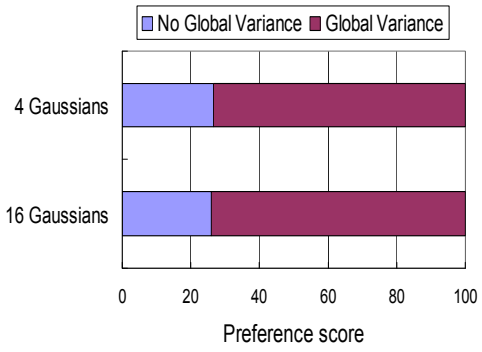

Fig. 3. Preference scores: Standard vs. Global Variance

the Global Variance factor with the usage of Gaussian mixture models, and to obtain thus the improvements of the speech quality due to these two factors.

8. ACKNOWLEDGEMENTS

This work was partially funded by the 21^{st} COE-Large-scale Knowledge Resources Program.

9. REFERENCES

[1] Iwano, K., Yamada, M., Togawa, T. and Furui, S., Prosody Control for HMM-based Japanese TTS, Text to Speech Synthesis. *New Paradigms and Advances*, Shrikanth Narayanan and Abeer Alwan (Eds.), Prentice Hall PTR, pp.155-173, July 2004.

[2] Latorre, J., Iwano, K. and Sadaoki, F., New approach to the polyglot speech generation by means of an HMM-based speaker adaptable synthesizer, *Speech Communication*, Vol. 48, Issue 10, pp. 1227-1242, October 2006.

[3] Schultz, T., Globalphone: a multilingual speech and text database developed at Karlsruhe University. *Proc. ICSLP*, pp. 345-348, September 2002.

[4] Tachiwa, W. and Furui, S., A study of speech synthesis using HMMs. *Proc. of Spring Meeting of the Acoustical Society of Japan (ASJ)*, Vol. 1, pp.239-240, March 1999. (In Japanese)

[5] Toda, T. and Tokuda, K., Speech parameter generation algorithm considering global variance for HMM-based speech synthesis. *Proc. Interspeech*, pp. 2801-2804, September 2005.

[6] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., and Kitamura, T., Speech parameter generation algorithm for HMM-based speech synthesis. *Proc. ICASSP*, pp. 1315-1318, June 2000.

[7] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T. and Kitamura, T., Mixed excitation for HMM-based speech synthesis. *Proc. Eurospeech*, pp 2263-2266, September 2003.