IEICE TRANS. INF. & SYST., VOL.E90–D, NO.3 MARCH 2007

LETTER

State Duration Modeling for HMM-Based Speech Synthesis

Heiga ZEN[†a)], ***Nonmember***, **Takashi MASUKO**[††*], **Keiichi TOKUDA**[†b)], ***Members***, **Takayoshi YOSHIMURA**[†**], ***Nonmember***, **Takao KOBAYASIH**[††c)], ***and*** **Tadashi KITAMURA**[†d)], ***Members***

SUMMARY This paper describes the explicit modeling of a state duration's probability density function in HMM-based speech synthesis. We redefine, in a statistically correct manner, the probability of staying in a state for a time interval used to obtain the state duration PDF and demonstrate improvements in the duration of synthesized speech.

key words: *duration modeling, speech synthesis, hidden Markov model*

1. Introduction

In the HMM-based speech synthesis, each state duration's probability density function (PDF) is modeled by a multivariate Gaussian distribution. In the previous formulation [1], [2], the state duration PDF was obtained based on the *a posteriori* probability of staying in a state for a time interval given an observation sequence and HMM, which is calculated from the state occupancy probabilities. However, the previous formulation excludes state transitions, resulting in an inconsistency between model structure and obtained state duration PDFs. For example, the probability of staying in a state for more than two frames is not zero even if the state's self transition probability is zero. To resolve this problem, we redefine the probability of staying in a state.

2. Probability of Staying in a State

In [1], [2], we defined $\chi_{t_0,t_1}(i)$, the probability of staying in the i-th state from time t_0 to t_1 given an observation sequence $\boldsymbol{o} = \{\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T\}$ of length T, as

$$\chi_{t_0,t_1}(i) = \left(1 - \gamma_{t_0-1}(i)\right) \cdot \prod_{t=t_0}^{t_1} \gamma_t(i) \cdot \left(1 - \gamma_{t_1+1}(i)\right), \qquad (1)$$

where $\gamma_t(i)$ is the probability of being in state i at time t, and we defined $\gamma_{-1}(i) = \gamma_{T+1}(i) = 0$. Based on $\chi_{t_0,t_1}(i)$, the mean $\xi(i)$ and the variance $\sigma^2(i)$ of the duration PDF of state i is obtained as

$$\xi(i) = \frac{\sum_{t_0=1}^{T} \sum_{t_1=t_0}^{T} \chi_{t_0,t_1}(i)\,(t_1 - t_0 + 1)}{\sum_{t_0=1}^{T} \sum_{t_1=t_0}^{T} \chi_{t_0,t_1}(i)}, \qquad (2)$$

$$\sigma^2(i) = \frac{\sum_{t_0=1}^{T} \sum_{t_1=t_0}^{T} \chi_{t_0,t_1}(i)\,(t_1 - t_0 + 1)^2}{\sum_{t_0=1}^{T} \sum_{t_1=t_0}^{T} \chi_{t_0,t_1}(i)} - \xi^2(i). \qquad (3)$$

However, the previous definition of $\chi_{t_0,t_1}(i)$ is statistically incorrect because the state transitions were not taken into account.

In this paper, we redefine $\chi_{t_0,t_1}(i)$ in a statistically correct manner as

$$\begin{aligned}\chi_{t_0,t_1}(i) &= P\left(q_{t_0-1} \neq i, q_{t_0} = i, \ldots, q_{t_1} = i, q_{t_1+1} \neq i \mid \boldsymbol{o}, \lambda\right) \\ &= \frac{1}{P\left(\boldsymbol{o} \mid \lambda\right)} \left(\sum_{j \neq i} \alpha_{t_0-1}(j) a_{ji}\right) \cdot a_{ii}^{t_1-t_0} \\ &\quad \cdot \prod_{t=t_0}^{t_1} b_i(\boldsymbol{o}_t) \cdot \left(\sum_{k \neq i} a_{ik} b_k(\boldsymbol{o}_{t_1+1}) \beta_{t_1+1}(k)\right), \end{aligned} \qquad (4)$$

where q_t denotes the state at time t, λ denotes the parameter set of the HMM, $\alpha_t(j)$ and $\beta_t(k)$ denote the forward and backward variables, and a_{ij} and $b_i(\cdot)$ denote the state transition probability from the i-th state to the j-th state, and the state output PDF of the i-th state, respectively.

3. Experiments

We obtained the state duration's PDFs using Eqs. (2), (3), and (4), and compared them with the previous definition Eq. (1).

From the ATR Japanese speech database B-set, 503 phonetically balanced sentences uttered by female speakers FTK and FYM and male speakers MHT and MYI were used for training and testing. Speech signals were sampled at 16 kHz and windowed by a 25-ms Blackman window with a 5-ms shift. Mel-cepstral coefficients were obtained by 25-th

Manuscript received July 27, 2006.

[†]The authors are with the Department of Computer Science and Engineering, Nagoya Institute of Technology, Nagoya-shi, 466–8555 Japan.

[††]The authors are with the Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Yokohama-shi, 226–8502 Japan.

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

[**]Presently, with Toyota Central R&D Labs., Inc.

a) E-mail: zen@ics.nitech.ac.jp

b) E-mail: tokuda@ics.nitech.ac.jp

c) E-mail: takao.kobayashi@ip.titech.ac.jp

d) E-mail: kitamura@nitech.ac.jp

DOI: 10.1093/ietisy/e90–d.3.692

Table 1 Total number of frames of training data and generated training sentences using sets of state duration PDFs estimated based on the previous definition Eq. (1) and new definition Eq. (4).

Speaker	Training data	Generated training sentences	
		Previous definition	New definition
FTK	473,309	465,202	473,309
FYM	451,571	441,814	451,571
MHT	425,688	413,686	425,688
MYI	501,382	480,541	501,382

order mel-cepstral analysis. The feature vectors consisted of 25 mel-cepstral coefficients including the zeroth coefficient, logarithm of fundamental frequency, and their delta and delta-delta coefficients. Context dependent phoneme labels were constructed based on phoneme labels included in the database. Each context dependent phoneme was modeled by a five-state left-to-right HMM with single Gaussian distributions and multi-space probability distributions as the state output PDFs of spectral and F_0 parts, respectively, and a five-dimensional Gaussian distribution as the state duration's PDF. A speaker-dependent HMM-based speech synthesis system was trained for each speaker using the first 450 sentences. In this experiment, we estimated two sets of state duration PDFs: one was based on the previous definition Eq. (1), and the other was based on the new definition Eq. (4).

For each speaker, we generated the 450 training sentences from the constructed HMM-based speech synthesis system. Table 1 shows the total number of frames of training data and generated training sentences. It can be seen from the table that there was inconsistency in the total number of frames between the training data and generated training sentences when the set of state duration PDFs estimated by the previous definition was used. However, this inconsistency was removed when the proposed formulation was used. This indicates that the proposed formulation is proper in the sense of statistical modeling and resolves the inconsistency in the previous definition.

4. Conclusion

In this paper, the probability of staying in a state for a time interval, which is used to obtain the state duration's PDFs in the HMM-based speech synthesis, was redefined in a statistically correct manner. It was shown that inconsistency in duration between training data and synthesized speech was eliminated using the proposed formulation.

References

[1] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," IEICE Trans. Inf. & Syst. (Japanese Edition), vol.J83-D-II, no.11, pp.2099–2107, Nov. 2000.

[2] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Duration modeling for HMM-based speech synthesis," Proc. ICSLP, pp.29–32, 1998.