FULL COVARIANCE STATE DURATION MODELING FOR HMM-BASED SPEECH SYNTHESIS

Heng Lu† [1], *Yi-Jian Wu*‡ , *Keiichi Tokuda*‡ , *Li-Rong Dai*† , *Ren-Hua Wang*†

† iFlytek Speech Lab, University of Science and Technology of China, Hefei, Anhui, CHINA
‡ Department of Computer Science Nagoya Institute of Technology, Nagoya, 466-8555 JAPAN

ABSTRACT

This paper proposes a state duration modeling method using full covariance matrix for HMM-based speech synthesis. In this method, a full covariance matrix instead of the conventional diagonal covariance matrix is adopted in the multi-dimensional Gaussian distribution to model the state duration of each context-dependent phoneme. At synthesis stage, the state durations are predicted using the clustered context-dependent distributions with full covariance matrices. Experimental results show that the synthesized speech using full-covariance state duration models is more natural than the conventional method when we change the speaking rate of synthesized speech.

***Index Terms*—**full covariance, duration, HMM, speech synthesis

1. INTRODUCTION

Speech synthesis system had been under development for many years. One of the important research topics in speech synthesis is to model the duration in training and then predict it when synthesizing speech, which is not a trivial task since the duration is affected by many context information. Many methods had been proposed for duration modeling and prediction problem using statistical models such as linear regression [1], tree regression [2], and sum of product model [3]. In recent years, HMM-based speech synthesis method [4][5] [6] had been developed, in which a state duration model was proposed with the capabilities to control rhythm and tempo by investigating state duration densities [9]. In this method, state duration of each phoneme HMM is modeled by a multi-dimensional Gaussian distribution. A decision-tree-based context clustering [10] is used in this method to deal with the datasparseness problem and at the same time to avoid over-fitting to training data, and natural duration prediction of the generated synthesis speech is obtained.

However, in conventional HMM-based speech synthesis, state durations are treated as independent variables to each other, and the diagonal matrix is used in the multi-dimensional Gaussian distribution to model the state durations of each phoneme. In fact, the correlation between each state duration cannot be ignored. Therefore, we adopt full covariance matrices instead of diagonal covariance matrices in the Gaussian distribution of state duration models. The clustering of context dependent state duration models is re-formulated using the full covariance matrix. In addition, we investigate the effectiveness of the full-covariance state duration model by changing the speaking rate of synthesized speech.

The rest of paper is organized as follows. Section 2 briefly review the diagonal covariance state duration multi-dimensional Gaussian HMM modeling and prediction proposed in [9], which is used in the HTS tools [11] and conventional HMM-based speech synthesis [7][8]. Section 3 introduces our work using the full covariance matrices for state duration modeling, including full-context state duration multi-dimensional Gaussian distribution modeling, full covariance model clustering and state duration generation. Finally, experimental results and discussions are presented in section 4, and conclusions are given in section 5.

2. STATE DURATION MODELING WITH DIAGONAL COVARIANCE MATRIX

Figure 1 illustrates the flowchart of the state duration model training and speech synthesis procedure. In the training stage, the context dependent state duration models are initialized from the statistical information of training data, and then the tree-based state dependent clustering is applied to cluster the state duration model. In the synthesis stage, the constructed tree is used to determine the state duration sequence for an input context label sequence, and then the acoustic parameters are generated using the parameter generation algorithm [4].

In current framework, the context dependent state duration models are initialized based on the statistical information collected from the last iteration of forward-backward re-estimation of context dependent clustered HMMs. The mean μ_i and the variance σ_i^2 of duration density of state i are determined by

$$\mu(i) = \frac{\sum_{t_0=1}^{T}\sum_{t_1=t_0}^{T}\Gamma_{t_0,t_1}(i)(t_1 - t_0 + 1)}{\sum_{t_0=1}^{T}\sum_{t_1=t_0}^{T}\Gamma_{t_0,t_1}(i)}, \tag{1}$$

$$\sigma^2(i) = \frac{\sum_{t_0=1}^{T}\sum_{t_1=t_0}^{T}\Gamma_{t_0,t_1}(i)(t_1 - t_0 + 1)^2}{\sum_{t_0=1}^{T}\sum_{t_1=t_0}^{T}\Gamma_{t_0,t_1}(i)} - \mu^2(i), \tag{2}$$

[1] The work is partially done when the first writer visiting department of computer science, Nagoya Institute of Technology.

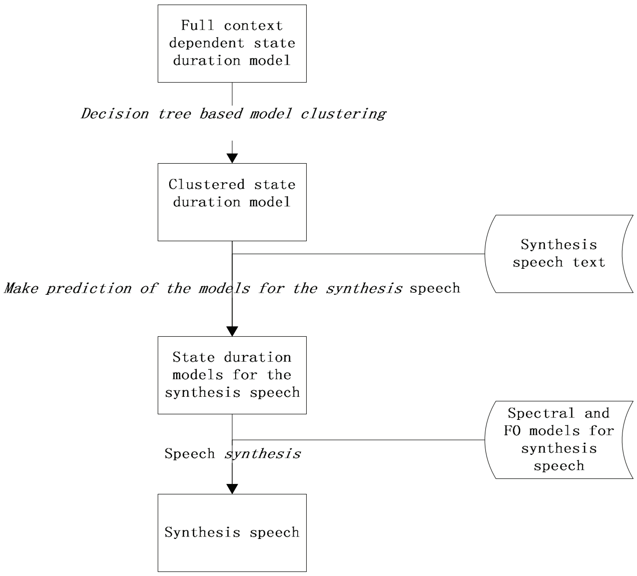
Full context dependent state duration model
Decision tree based model clustering
Clustered state duration model
Synthesis speech text
Make prediction of the models for the synthesis speech
State duration models for the synthesis speech
Spectral and F0 models for synthesis speech
Speech synthesis
Synthesis speech

Figure 1: Flowchart of state duration model clustering and waveform synthesis

respectively. $\Gamma_{t_0,t_1}(i)$ is the probability that the observations from time t_0 to t_1 was occupied by the state model i, which can be calculated as

$$\Gamma_{t_0,t_1}(i)=(1-\gamma_{t_0-1}(i))\cdot\prod_{t=t_0}^{t_1}\gamma_t(i)\cdot(1-\gamma_{t_1+1}(i)) \tag{3}$$

where $\gamma_t(i)$ is the probability that the observation at time t is produced by the state model i. Note that $\gamma_{-1}(i)=\gamma_{T+1}(i)=0$ in Eq. (3).

In the clustering of context dependent duration model, a minimum description length (MDL) [12][13] criterion is used for conducting the decision tree based context dependent HMM model clustering. The splitting condition of each tree node is

$$K\log W\leq \frac{1}{2}(\Gamma_m\log|\Sigma_m|-\Gamma_{mqy}\log|\Sigma_{mqy}|-\Gamma_{mqn}\log|\Sigma_{mqn}|) \tag{4}$$

where K is the dimension of the state duration feature vector, and it equals to the state number in a phone HMM model. W is the total number of training samples, which is a constant number during the process of tree node splitting. Γ_m, Γ_{mqy} and Γ_{mqn} are the occupation count of training data in the current splitting node, the left and right child tree nodes after split, and Σ_m, Σ_{mqy} and Σ_{mqn} are the corresponding covariance matrix of the splitting node and the left and right child tree nodes after split, respectively. In previous duration modeling, the diagonal covariance matrices are used.

For a given synthesis sentence with total length T, state durations $\{d_{n,k}, n=1,2,...,N; k=1,2,...,K\}$ which maximize the log likelihood to the state model is given by

$$d_{n,k}=\mu_{n,k}+\sigma_{n,k}^2\cdot\rho \tag{5}$$

$$\rho=\frac{(T-\sum_{n=1}^{N}\sum_{k=1}^{K}\mu_{n,k})}{\sum_{n=1}^{N}\sum_{k=1}^{K}\sigma_{n,k}^2} \tag{6}$$

where $d_{n,k}$ is the duration of state k of phone n, N is the total number of phones in the sentence and K is the state number in a phone HMM . $\mu_{n,k}$ and $\sigma_{n,k}^2$ are the mean and variance of the state duration model of state k of phone n respectively.

3. STATE DURATION MODELING WITH FULL COVARIANCE MATRIX

In previous state duration modeling with diagonal covariance matrix, the state durations are treated as independent variables of each other. In fact, the correlation between each state duration cannot be ignored. Therefore, we adopt full covariance matrices instead of diagonal covariance matrices in the multi-dimensional Gaussian distribution of state duration models.

3.1. Initialization of context dependent duration model

In the initialization of context dependent state duration model with full covariance matrix, due to the innumerous combination of the context information, almost all of the context dependent duration models have only one sample. We thus use the diagonal covariance matrices of duration models obtained from equation (1) and (2) as the initial full covariance matrix for context dependent models and set the value of all the non-diagonal elements to zero.

3.2. Full-covariance state duration model clustering

A tree-based clustering is conducted to cluster the context dependent state duration model, and a MDL-based stoping criterion is used to control the size of tree. The updating equations of the mean and covariance matrices are as follows

$$\mu_m=\frac{\sum_{l=1}^{L_m}\Gamma_l^m\mu_l^m}{\sum_{l=1}^{L_m}\Gamma_l^m} \tag{7}$$

$$\Sigma_m=\frac{\sum_{l=1}^{L_m}\Gamma_l^m(\Sigma_l^m+(\mu_l^m)(\mu_l^m)')}{\sum_{l=1}^{L_m}\Gamma_l^m}-(\mu_m)(\mu_m)' \tag{8}$$

where μ_m is the mean of node m and $\sum_m$ is the full covariance matrix of node HMM m. μ_l^m, Σ_l^m and Γ_l^m are the mean, full covariance matrix and the occupation probability of the l 'th context dependent model which belongs to the m 'th tree node, respectively. L_m is the number of context dependent model which belongs to the m 'th tree node.

As full covariance matrices are used here, the number of free parameters of each tree node m becomes

$\{(\frac{K^2-K}{2})+K+K\}/2=\frac{K^2+3K}{4}$. Therefore, the description length (DL) I of all the tree leaf nodes becomes

$$\begin{aligned} l(U) &= \sum_{m=1}^{M}\frac{1}{2}\Gamma_m(K+K\log(2\pi)+\log|\boldsymbol{\Sigma}_m|) \\ &+(\frac{K^2+3K}{4})M\log W+C \end{aligned} \qquad (9)$$

where M is the total number of leaf nodes in the decision tree. If node m is split, the variation of DL is

$$\begin{aligned} &\Delta_m(q)=l(U')-l(U) \\ &=\frac{1}{2}(\Gamma_{mqy}\log|\boldsymbol{\Sigma}_{mqy}|+\Gamma_{mqn}\log|\boldsymbol{\Sigma}_{mqn}|-\Gamma_m\log|\boldsymbol{\Sigma}_m|) \\ &+\frac{(K^2+3K)}{4}\log W \end{aligned} \qquad (10)$$

where q is the question set. If $\Delta_m(q)<0$, the node splitting is conducted, and the splitting condition becomes

$$\begin{aligned} &\frac{(K^2+3K)}{4}\log W \\ &\le\frac{1}{2}(\Gamma_m\log|\boldsymbol{\Sigma}_m|-\Gamma_{mqy}\log|\boldsymbol{\Sigma}_{mqy}|-\Gamma_{mqn}\log|\boldsymbol{\Sigma}_{mqn}|) \end{aligned} \qquad (11)$$

3.3. State duration prediction

When the total length T of the utterance is given in synthesis stage, the state durations are determined by maximizing the following log likelihood

$$\begin{aligned} &\log P(\boldsymbol{d}\,|\,\lambda,T) \\ &\boldsymbol{d}_n=(d_{n,1},d_{n,2},...,d_{n,K})' \\ &\boldsymbol{d}=(\boldsymbol{d}_1',\boldsymbol{d}_2',...,\boldsymbol{d}_N')' \end{aligned} \qquad (12)$$

under the constraint

$$T=\sum_{n=1}^{N}\sum_{k=1}^{K}d_{n,k}$$

As the multi-dimensional Gaussian distribution is used here, we use Lagrange Multiplier to solve the problem, and the equation is rewritten as

$$\begin{aligned} &F=(\boldsymbol{d}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{d}-\boldsymbol{\mu})+\lambda(\sum_{n=1}^{N}\sum_{k=1}^{K}d_{n,k}-T) \\ &\boldsymbol{\mu}_n=(\mu_{n,1},\mu_{n,2},...,\mu_{n,K})' \\ &\boldsymbol{\mu}=(\boldsymbol{\mu}_1',\boldsymbol{\mu}_2',...,\boldsymbol{\mu}_N')' \end{aligned} \qquad (13)$$

and $\boldsymbol{\Sigma}^{-1}=diag(\boldsymbol{\Sigma}_1^{-1},\boldsymbol{\Sigma}_2^{-1},...,\boldsymbol{\Sigma}_N^{-1})$.

By setting $\frac{\partial F}{\partial d_{n,k}}=0,n=1,2,...,N;k=1,2,...,K$ and $\frac{\partial F}{\partial\lambda}=0$, we get a set of linear equations, and solve them as

$$\boldsymbol{d}=\boldsymbol{\mu}+\Sigma\cdot\boldsymbol{\rho} \qquad (14)$$

$$\boldsymbol{\rho}=\frac{(T-\boldsymbol{I}'\cdot\boldsymbol{\mu})}{\boldsymbol{I}'\cdot\Sigma\cdot\boldsymbol{I}}\cdot\boldsymbol{I} \qquad (15)$$

where $\boldsymbol{I}=\{[1\quad 1\quad ...\quad 1]_{1\times NK}\}'$. By comparing the above equations to Equ. (5)-(6), it can be seen f that the above solution using full-covariance duration model is a more general form of the solution using the diagonal covariance duration model.

4. EXPERIMENTS

We used a Mandarin database containing 1,000 phonetically balanced sentences from a female speaker for training. Speech signals were sampled at 16kHz. We calculated 40 static mel-cepstral coefficients and 1-dimensional log f0 feature with their first and second derivatives as the acoustic feature. 5-state left-to-right HMMs were used in the experiment. Single Gaussian distribution with diagonal covariance matrix was used for the spectral modeling and the multi-space distribution (MSD) for F0 modeling [6]. In our experiments, we investigate the effectiveness of state duration model with full covariance matrix by comparing it with the state duration model with diagonal covariance matrix.

The context features used in the experiment include:

- Left phone : phone before the current phone
- Current phone : the phone that we are modeling
- Right phone : phone after the current phone
- Left tone : tone of the syllable before the current syllable
- Current tone : the tone of the current syllable
- Right tone : tone of the syllable after the current syllable
- Part-of-speech: nature of the current word
- Relative positions of current syllable, word, phrase, sentence, sentence group
- Absolute positions from head and tail of current syllable, word, phrase, sentence, sentence group

After the tree-based clustering for state duration models, 465 nodes were generated in the baseline system using diagonal covariance matrix, and 260 nodes were generated for the state duration models with full covariance matrices.

In the synthesis stage, we set the length T of whole utterance to alpha times of the sum of the means of all state duration models, and imitate the stretch and the compression of the utterance length, which is

$$T=\alpha\cdot\sum_{n=1}^{N}\sum_{k=1}^{K}\mu_{n,k} \qquad (16)$$

In the experiment, we set $\alpha=1.5$ and used the clustered state duration models with the diagonal covariance matrices in the baseline and the full covariance matrices in our proposed method to generate state durations using equations in section 2 and section 3, respectively. And the same spectral and f0 models are used to generate spectral and f0 parameters.

We generated 30 sentences using these two state duration models, and conducted a subjective listening test to evaluate the performances of the duration models. 5 listeners participated in the listening test. Each of them listened to the 30 sentences pair by pair in a random order, and was asked to select the sentence which is more natural. The result is shown in figure 2, where the preference scores with 95% confidence interval are given.

In this figure, the average preference percentage of proposed method and the baseline is 58.67% and 41.33% with the 95%

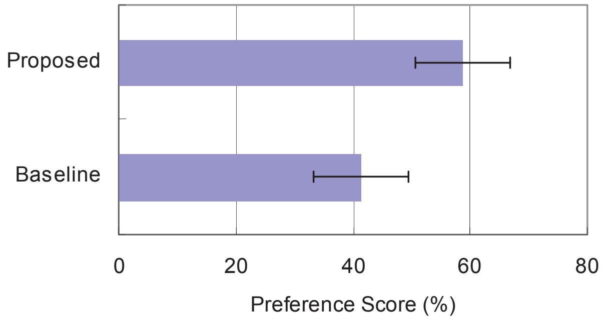

Figure 2: Preference score for diagonal and full covariance state duration model with 95% confidence interval.

confidence interval of ±8.16%. From the result, we can see that, the proposed full covariance matrix state duration modeling method outperforms the diagonal covariance matrix state duration modeling with fixed sentence time T . The result is reasonable for the full covariance matrix state duration modeling models state durations without losing their correlation, and in the synthesis stage, state durations are predicted with regard of the interconnections with each other, which cause the timing of the synthesis wave more natural.

However, if we set the stretch time $\alpha = 1$, and maximum likelihood criterion is used, the state durations will be equal to the means of the state duration models. In this case, the covariance matrices will not be used and thus not affect the result of generated durations in the speech synthesis process.

5. CONCLUSION

In this paper, we propose a state duration modeling method with full covariance matrix Gaussian distributions. Compared with the traditional diagonal covariance duration modeling, the proposed method has the capability to capture the correlation between each state duration. In the synthesis stage, full covariance Gaussion distributions are used to generate state durations forsynthesized speech. Subjective listening result shows that, with a given sentence time T , the proposed method outperforms the conventional diagonal state duration modeling in the overall naturalness. More natural timing is obtained with the proposed method.

In the future work, we plan to study more on the full covariance state duration modeling and do more experiments on state duration generation using full covariance state duration models together with phone duration models.

6. ACKNOWLEDGEMENTS

This work was partially supported by National Natural Science Foundation of China (Grant No.: 60610298).

7. REFERENCES

[1] N. Kaiki, K. Takeda, Y. Sagisaka: “Linguistic properties in the control of segmental duration for speech synthesis,” *Talking Machines: Theories, Models, and Designs,* Elsevier Science Publishers, pp.255–263, 1992.

[2] M. Riley: “Tree-based modelling of segmental duration,” *Talking Machines: Theories, Models, and Designs,* Elsevier Science Publishers, pp.265–273, 1992.

[3] J. P. H. van Santen, C. Shih, B. Mobius, E. Tzoukermann and M. Tanenblatt: “Multi-lingual duration modeling,” *Proc. EUROSPEECH-97*, vol5, pp.2651–2654, 1997.

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

[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] 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

[7] 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.

[8] H. Zen, T. Toda, M. Nakamura, and K. Tokuda, “Details of the Nitech HMM-based speech synthesis for Blizzard Challenge 2005,” *IEICE Trans. on Inf. and Systems*, vol. E90-D, no. 1, 2007.

[9] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura: “Duration Modeling in HMM-based Speech Synthesis System,” *Proc. of ICSLP*, v01.2, pp.29-32,1998.

[10] S.J. Young, J.J. Odell, and P.C. Woodland: “Tree Based State Tying for High Accuracy Modeling”, *ARPA Workshop on Human Language Technology*, Morgan Kaufmann Publishers, Princeton, NJ, March 1994.

[11] K. Tokuda, H. Zen, S. Sako, T. Yoshimura, J. Yamagishi, M. Tamura, and T. Masuko: “The HMM-based speech synthesis software toolkit,” http://hts.ics.nitech.ac.jp

[12] K Shinoda, T Watanabe: “MDL-based context-dependent subword modeling for speech recognition” Acoustical Science and Technology”, *J-STAGE Page 1. J. Acoust. Soc. Jpn. (E)* 21, 2 (2000)

[13] K Shinoda, T Watanabe: “Acoustic modeling based on the MDL criterion for speech recognition,” *EuroSpeech97* 1.99-102 (1997)