Asynchronous F0 and Spectrum Modeling for HMM-Based Speech Synthesis

Cheng-Cheng Wang, Zhen-Hua Ling, Li-Rong Dai

USTC iFlytek Speech Lab, University of Science and Technology of China, Hefei,China
ccwang23@mail.ustc.edu.cn, zhling@iflytek.com, lrdai@ustc.edu.cn

10.21437/Interspeech.2009-136

Abstract

This paper proposes an asynchronous model structure for fundamental frequency(F0) and spectrum modeling in HMM-based parametric speech synthesis to improve the performance of F0 prediction. F0 and spectrum features are considered to be synchronous in the conventional system. Considering that the production of these two features is decided by the movement of different speech organs, an explicitly asynchronous model structure is introduced. At training stage, F0 models are training asynchronously with spectrum models. At synthesis stage, the two features are generated respectively. The objective and subjective evaluation results show the proposed method can effectively improve the accuracy of F0 prediction.

Index Terms: speech synthesis, hidden Markov model, asynchronous modeling, fundamental frequency

1. Introduction

The hidden Markov models (HMMs) have been successfully applied to speech synthesis. In the conventional HMM-based speech synthesis system, a synchronous multi-stream HMM is trained for each context-dependent phone to model the spectrum and F0 features simultaneously. That means the spectrum and F0 parameters share a same state sequence. According to the speech production mechanism, the characteristics of F0 and spectrum features are decided by the movement of different speech organs. Different from the spectrum features which reflect the shape of the vocal tract, the F0 features are generated the vibration of the vocal cords when people are pronouncing [1]. Therefore, there exists asynchrony between these two feature sequences which conflicts with the strictly synchronous model structure adopted in the conventional HMM-based speech synthesis.

If we presume the synchronous relationship between the F0 and spectrum features, what affection will bring to the HMM-based speech synthesis system? As we known, at the HMM training stage, the parameters of HMM models, including the state transition probability and the output probability distribution, are updated by Baum-Welch algorithm to make the observed feature sequences in the training database have the maximum probability. Under the synchrony assumption, a unique state occupancy probability is used in the model re-estimation formulas for both F0 and spectrum features at time t as follows:

$$\gamma_t(i) = P(q_t = S_i \mid \boldsymbol{O}_1, \boldsymbol{O}_2, \lambda) \tag{1}$$

where $\gamma_t(i)$ is the occupancy probability of state i at time t. $\boldsymbol{O}_1, \boldsymbol{O}_2$ denote the observation of spectrum and F0 parameters respectively. λ is the parameter of HMM. The dimension of spectrum parameter is much higher than that of F0 parameter. Therefore, the spectrum parameter dominates the calculation of state occupancy probability. This degrades the model accuracy of F0 features which usually has low dimensions.

Based on the discussions above, an asynchronous F0 and spectrum model structure for HMM-based speech synthesis is proposed. At model training stage, two state occupancy probabilities are calculated for F0 features and spectrum features respectively at each frame, so F0 models are training asynchronously with spectrum models. At synthesis stage, the parameters are generated by maximizing the likelihood functions of the F0 models and spectrum models respectively. Experimental results show that when the asynchronous modeling method is adopted we can improve the accuracy of F0 parameter prediction effectively in both objective and subjective views.

The rest of this paper is organized as follows: Section 2 describes the proposed method in respect of model training and parameter generation comparing with the baseline which uses synchronous model structure. Experimental results are presented in Section 3. Finally in Section 4, we draw the conclusions.

2. Method

2.1. Baseline System

2.1.1. *Model Training*

The overall training and synthesis processes of HMM-based speech system are shown in Fig.1. Acoustic features are extracted from the speech waveforms of training database. STRAIGHT [2] as a high quality speech vocoder is adopted here to analyze the spectral envelop and F0 for each speech frame. F0 and spectrum models trained by acoustic features are considered to be synchronous in the conventional HMM-based speech synthesis system. In context-dependent phone model training, the acoustic features for each frame consist of static, delta and delta-delta components of logarithmized F0 and the spectral envelop. The context-dependent HMMs are estimated according to the acoustic features and context labels under the maximum likelihood criterion [3]. F0 stream is modeled by a multi-space probability distribution (MSD) [4]. In our experiments, we use two-space probability distribution, which represent the either of voiced and unvoiced frames. A decision tree based model clustering method is applied to deal with data-sparsity problems and predict the context-dependent models outside the training set. In training period, the mean and variance parameters of F0 and spectrum features are estimated for each context-dependent HMM [5].

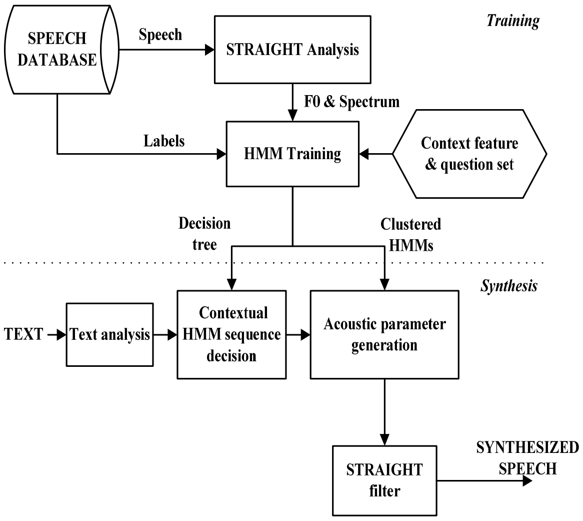

Fig.1. *Flowchart of the baseline HMM-based speech synthesis system*

2.1.2. *Parameter Generation*

For an input text, the HMM sequences of the sentence are determined by the clustered HMMs and decision trees. The spectral and F0 parameters are generated by maximum likelihood parameter generation algorithm [6].

For a given sentence HMM λ, to generate speech parameters, we maximize $P(\boldsymbol{O}\,|\,\lambda)$ with respect to the speech parameter vector sequence $\boldsymbol{O}$ as:

$$\boldsymbol{O}_{\max} = \arg\max_{\boldsymbol{O}} P(\boldsymbol{O}|\boldsymbol{q}_{\max}, \lambda) \qquad (2)$$

where $\boldsymbol{q}_{\max}$ is the state sequence predicted by combining phone duration models and state duration models in the baseline system.

In the baseline system, a state duration model is trained to predict the duration of every state in the utterance for synthesis. d_i is the duration of one state can be predicted as:

$$d_i^* = \max_{d_i}[\sum_i \log P_i(d_i\,|\,\lambda_i) + w * \log P(d\,|\,\lambda)] \qquad (3)$$

where $P_i = N(d_i\,|\,m_i, \sigma_i^2)$ presents the state duration model and $P = N(d\,|\,m, \sigma^2)$ presents the phone duration model. w is set as the weight between these two models. So we can get optimization of d_i by the derivative of the Eq. 3 [7]:

$$d_i^* = m_i + \rho.\sigma_i^2 \qquad (4)$$

where, $$\rho = \frac{w(m - \sum_i m_i)}{\sigma^2 + w\sum_i \sigma_i^2} \qquad (5)$$

2.2. Proposed Method

2.2.1. *Introduction of Asynchronous Model Structure*

As the baseline system introduced in section 2.1, F0 and spectrum models which are synchronous and independent are shown in Fig.2 (a). Based on the presentation in the section of Introduction, an asynchronous and independent model is proposed as shown in Fig.2 (b). The state sequence $\boldsymbol{q}$ is shared by the F0 feature sequence $\boldsymbol{O}_1$ and spectrum feature sequence $\boldsymbol{O}_2$. Whereas, in the proposed asynchronous modeling structure, the two feature steams correspond to two different state sequences. When this model is introduced into our HMM-based speech synthesis systems, it plays an asynchronous role in both the training and synthesis stages. In training stage, two different optimized state sequences of spectrum and F0 is extracted respectively by training data. In synthesis stage, two state sequences for F0 and spectral features are predicted respectively

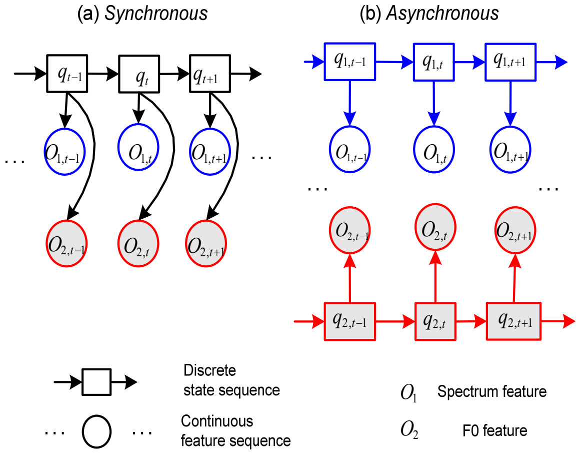

Fig.2. *Feature production Models of the synchronous and asynchronous modeling method*

2.2.2. *Model Training*

A semi-asynchronous model structure is adopted in our implementation. The reason why we choose semi-asynchronous modeling instead of full-asynchronous one is that if the spectrum and F0 features are modeled without constraining in the same phone boundary, some unvoiced utterances will cause the inaccuracy of the phone boundary of F0 feature. Further more, the alignment between the generated spectral and F0 features will be more difficult in synthesis stage for full-asynchronous modeling. The phone boundaries are determined by Viterbi alignment using the trained models of baseline system. That is an important constrain in asynchronous system. The spectrum and F0 features have the same phone boundary as the baseline. So actually we only model spectrum and F0 respectively inside the unit of phone.

The asynchronous model training can be divided into several steps as Fig.3. Firstly, the process is the same as the baseline system until getting the clustered HMMs. Fully context-dependent models can be produced by F0 and spectrum parameters. Then initial clustered HMM model is trained from full context model by decision trees. Secondly, Viterbi alignment gives the synchronous phone boundary of spectrum and F0 features. These two steps are the same as conventional synchronous training. Thirdly, the spectrum and F0 models are trained separately by Baum-Welch algorithm based on the given phone boundaries. Here two different state occupancy probability of each frame are calculated for spectrum and F0 features. Finally, we can get state segmentation in each phone for spectrum and F0 respectively to train phone duration, spectrum state duration and F0 state duration model. These two steps belong to asynchronous training stage.

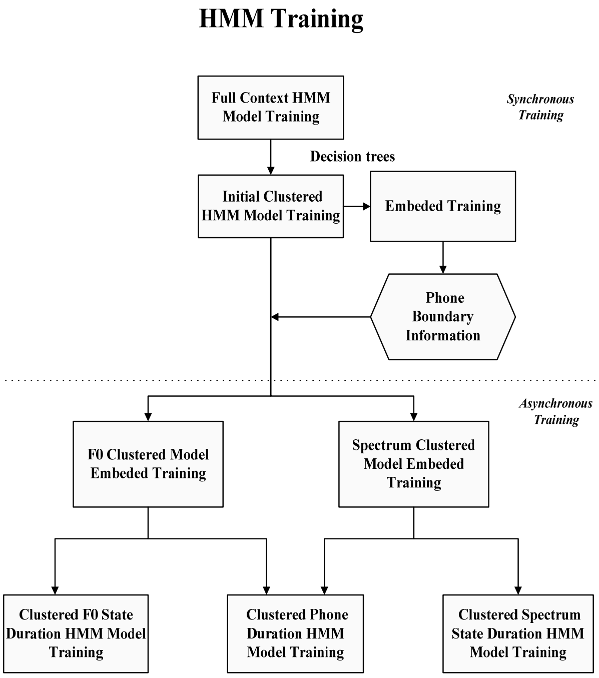
HMM Training
Full Context HMM Model Training
Synchronous Training
Decision trees
Initial Clustered HMM Model Training
Embeded Training
Phone Boundary Information
Asynchronous Training
F0 Clustered Model Embeded Training
Spectrum Clustered Model Embeded Training
Clustered F0 State Duration HMM Model Training
Clustered Phone Duration HMM Model Training
Clustered Spectrum State Duration HMM Model Training

Fig.3.*Flowchart of the proposed semi-asynchronous model training method*

2.2.3. Parameter Generation

The maximum likelihood parameter generation algorithm [6] is also adopted here to generate spectral and F0 parameters at synthesis stage. For a given HMM λ , to generate speech parameters, we maximize $P(\boldsymbol{O}_1,\boldsymbol{O}_2|\lambda)$ with respect to the spectrum and F0 speech parameter vector sequence $\boldsymbol{O}_1$, $\boldsymbol{O}_2$.

$$P(\boldsymbol{O}_1,\boldsymbol{O}_2|\lambda)=\sum_{q_1,q_2}P(\boldsymbol{O}_1,\boldsymbol{O}_2|\lambda,\boldsymbol{q}_1,\boldsymbol{q}_2)P(\boldsymbol{q}_1,\boldsymbol{q}_2|\lambda)$$
$$=\sum_{q_1,q_2}P(\boldsymbol{O}_1|\lambda,\boldsymbol{q}_1)P(\boldsymbol{O}_2|\lambda,\boldsymbol{q}_2)P(\boldsymbol{q}_1,\boldsymbol{q}_2|\lambda) \qquad (6)$$

In conventional synchronous system, state sequence $\boldsymbol{q}_1=\boldsymbol{q}_2=\boldsymbol{q}$; while, in proposed asynchronous system, $\boldsymbol{q}_1\neq\boldsymbol{q}_2$. So the parameter generation algorithm approximately can be divided into two maximum steps:

$$[\boldsymbol{q}_1^*,\boldsymbol{q}_2^*]=\arg\max_{q_1,q_2}P(\boldsymbol{q}_1|\lambda)P(\boldsymbol{q}_2|\lambda) \qquad (7)$$

$$[\boldsymbol{O}_1^*,\boldsymbol{O}_2^*]=\arg\max_{O_1,O_2}P(\boldsymbol{O}_1|\lambda,\boldsymbol{q}_1^*)P(\boldsymbol{O}_2|\lambda,\boldsymbol{q}_2^*) \qquad (8)$$

Prediction of duration as Eq. 7 relates to the weighting of phone duration, spectrum state duration and F0 state duration. The prediction process equals maximum the following likelihood function F:

$$\mathrm{F}=\sum_i\log P_i(d_i|\lambda_i)+w_1\sum_i\log P_i'(d_i'|\lambda_i')+w_2\log P(d|\lambda) \quad (9)$$

$$\text{S. T.}: \quad \sum_i d_i=d \qquad (10)$$

$$\sum_i d_i'=d \qquad (11)$$

where, the first part of the likelihood function is the likelihood of spectrum state duration; the second part is the likelihood of F0 state duration; the third is the likelihood of phone duration. The three parts combine with the weight coefficients w_1 and w_2 . $P_i=N(d_i|m_i,\sigma_i^2)$ presents the spectrum state duration model, $P_i'=N(d_i'|m_i',\sigma_i'^2)$ presents the F0 state duration model, $P=N(d|m,\sigma^2)$ presents the phone duration model. We can get optimization of d_i and d_i' by the derivative of the likelihood function as the baseline method [7]:

$$d_i^*=m_i+\rho_1.\sigma_i^2 \qquad (12)$$

$$d_i'^*=m_i'+\rho_2.\sigma_i'^2 \qquad (13)$$

where, $$\rho_1=\frac{d-\sum_i m_i}{\sum_i\sigma_i^2} \qquad (14)$$

$$\rho_2=\frac{w_1(d-\sum_i m_i')}{\sum_i\sigma_i'^2} \qquad (15)$$

Once the state sequences are given, Eq. 8 can be solved following the same maximum likelihood parameter generation method used in the baseline system [6].

3. EXPERIMENTS

3.1. Experiment Condition

Speech synthesis is experimented on Chinese speech databases of a female speaker, including 1000 sentences. Speech signals are sampled at 16 kHz and analyzed by STRAIGHT vocoder. The 5-state left-to-right phone HMM model with single-mixture are adopted. We selected 50 sentences as test set; the others were used for training. Objective measure — RMSE (Root of Mean Square Error) and subjective measure — Preference Score are used to assess the performance.

3.2. Objective Experiment

The average log probability of estimated HMMs on the training data set is used here as an objective measure to evaluate the accuracy of trained models and the results are shown in Table. 1. From this table, we can see that the average log probability of the proposed asynchronous system is higher than that of the baseline system, which means the asynchronous model structure can describe the F0 and spectral features in the training data set better than the synchronous assumption.

Table.1. *Average log probability of the two systems*

System	Average log probability per frame	
Baseline	4.390736e+02	
Proposed	4.393973e+02	
	F0 Stream	Spectrum Stream
	5.004088	4.343933e+02

Comparing the state alignments of training sentences using the baseline system and proposed system, we find that the results are almost the same for spectral features but are quite different for F0 features. This coincides with the analysis in section 1 and 2. The following measurements focus on F0 prediction instead of spectrum because of its variety.

RMSE is commonly used to evaluate the mean error between generated parameter and original parameter. We define it as following function:

$$RMSE = \sqrt{\sum_{i}^{N} (\log(f_0(i))\text{-}\log(f_e(i)))^2 / N} \qquad (16)$$

where N is total frames in a sentence, $f_0(i)$ is original F0 parameter, $f_e(i)$ is estimated F0 parameter. Only the frames that are voiced in both natural and estimative generated F0 sequences are used to calculate the RMSE according to Eq. 16. In order to calculate the RMSE we use the natural state duration given by Viterbi alignment using trained HMM here to achieve the same number of frames with natural F0 parameter. Table.2 shows the average RMSE of 20 sentences in the test set. The proposed system reduces the RMSE of predicted log F0 by 23.20% compared with the baseline system.

Table.2. *F0 prediction RMSE of the two systems*

System	Baseline	Proposed
RMSE (test set)（logHz）	0.0932140	0.0715883

In order to have more direct perception, we select a segment of voiced speech to see the accuracy of F0 prediction as shown in Fig.4. From this figure we can see that the proposed system works much better than the baseline. It gets more close to the natural F0 contour.

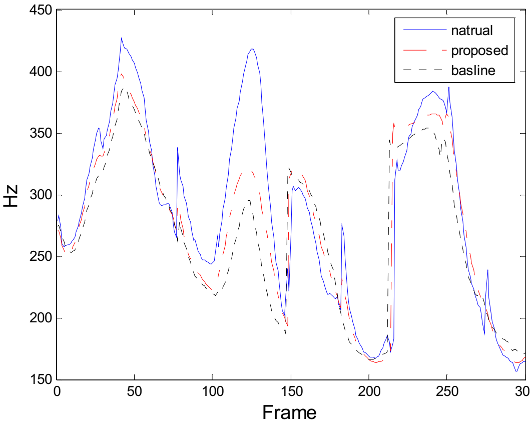

Fig.4. *The F0 prediction results using the baseline system and proposed system*

3.3. Subjective Experiment

Preference Score measures which one is better between two similar sentences. The two sentences are synthesized from the same text in the test set. The one with more natural pitch contour gets the score of 1, the other gets 0. The preference score is defined as following:

$$\text{Preference Score} = \frac{\sum_i S(i)}{N * m} \qquad (17)$$

where N is the number of listeners, m is the number of sentences in each set. *S(i)* is the score of each set.

20 sentences which are the same as the sentences used in the objective experiment are synthesized respectively using the baseline system and the proposed system for each database. These sentences are evaluated by 5 speech expert listeners pair by pair. The final preference score of these two systems are calculated as Eq. 17 and shown in Fig. 5.

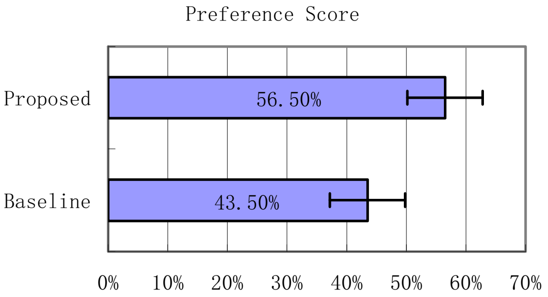

Fig.5. *The evaluation results of preference score*

The proposed system also has a good performance in the subjective listening test. The proposed model has 56.5% preference score to 43.5% in the subjective evaluation. The 95% confidence interval result in our experiments shows that the proposed asynchronous F0 modeling method can improve the performance of F0 prediction for speech synthesis significantly.

4. Conclusions

In this paper, an asynchronous HMM structure is proposed to present the asynchronous relationship between F0 and spectrum features for HMM-based speech synthesis is presented. Compared to the baseline which the two feature streams share a same state sequence, here the state sequences of the two features are independent in the proposed method. The two state sequences should also be predicted simultaneously at synthesis stage. In our experiments, the proposed model reduces the RMSE of predicted log F0 by 23.20% compared with the baseline system. A subjective evaluation shows the proposed method can improve the naturalness of synthesized speech effectively.

5. References

[1] X.Huang, A.Acero, and H.Hou, “Spoken Language Processing: A Guide to Theory, Algorithm, and System Development,” *Prentice Hall*, pp. 19-33, 2001.

[2] H.Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, “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.

[3] 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*, pp. 2347-2350, 1999.

[4] 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,* pp. 229-232, 1999.

[5] Zhenhua Ling, Long Qin, Heng Lu, Yu Gao, Lirong Dai, and Renhua Wang, “The USTC and iFlytek Speech Synthesis System for Blizzard Challenge 2007,” *in ICSLP Satellite Workshop, Blizzard Challenge,* 2007.

[6] 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, June 2000.

[7] T.Yoshimura, K.Tokuda, T.Masuko, T.Kobayashi and T.Kitamura, “Duration Modeling For HMM-Based Speech Synthesis,” *In ICSLP-1998,* paper 0939. 1998.