MULTI-LAYER F0 MODELING FOR HMM-BASED SPEECH SYNTHESIS

Cheng-Cheng Wang, Zhen-Hua Ling, Bu-Fan Zhang, Li-Rong Dai

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

ABSTRACT

This paper proposes a two-layer fundamental frequency (F0) modeling method for HMM-based parametric speech synthesis. The F0 models are trained for each context-dependent phoneme in the conventional HMM-based speech synthesis system. Considering the super-segmental characteristics of F0 features, an explicit syllable-layer F0 model is introduced in this paper. At synthesis stage, the F0 contour is generated by maximizing the combined likelihood functions of the phone-layer and syllable-layer F0 models. The objective and subjective evaluation results in our experiments show that the proposed multi-layer F0 modeling method can improve the performance of F0 prediction for emotional speech synthesis.

Index Terms—Speech synthesis, hidden Markov model, fundamental frequency modeling

1. INTRODUCTION

The hidden Markov models (HMMs) have been successfully applied to speech synthesis systems. In the HMM-based speech synthesis system, phone is generally used as the basic unit for modeling and F0 HMMs are trained for each context-dependent phone. Before the HMM-based speech synthesis method has been proposed, HMMs had been widely applied to speech recognition systems, which focus on the modeling of spectrum more than F0. However, different from spectrum feature which is segmental, F0 is a super-segmental feature. People usually pay attentions to the characteristics of spectrum within a phone, while the contour of F0 beyond the phone layer. The F0 contour is associated with some high-level prosodic features with larger scope than a phone, such as tones of a syllable, stress in a phrase, prosodic boundary in a sentence, etc. Further more, emotional speech is influenced by not only the contextual information but also some paralinguistic features. It is difficult to study and model the paralinguistic variations from only the F0 features within a phone. Therefore, the construction of long-term F0 models is of important significance.

Furthermore, syllable is treated as the basic unit for F0 modeling by many prosodic models, such as target model [1], Fujisaki model [2] and SFC model [3]. In these models, the super-segmental characteristics of F0 are modeled explicitly. Take the SFC model for example, the F0 contour is assumed to be a superposition of several tiers, including clause tier, phrase tier, syllable tier and so on, and syllable is the smallest unit for modeling.

Based on the discussions above, a new F0 modeling method which combines syllable-layer model and phone-layer model for HMM-based speech synthesis is proposed. We can get much long-term information about F0 features by means of syllable model. Experimental results show that when an appropriate integration between the likelihood functions of phone model and syllable model in parameter generation is implemented we can achieve a good performance in accuracy of F0 parameter prediction. Three emotional speech databases — neutral, happiness and sadness are adopted in our experiments.

The rest of this paper is organized as follows: Section 2 describes the proposed method in respect of model training and parameter generation. In Section 3, experimental results of evaluation and comparison between the proposed method and the baseline system in three different emotional set are presented. Finally in Section 4, we will draw our conclusions and future plans.

2. METHOD

2.1. Model Training

The overall training and synthesis processes of HMM-based speech system is shown in Fig.1. In this paper, we only focus on the construction of F0 model. The spectrum part is the same as the conventional one [4]. Acoustic features are extracted from the speech waveforms of training database. STRAIGHT [5] as a high quality speech vocoder is adopted here to analyze the spectral envelop and F0 for each speech frame. In context-dependent phone model training, the feature for each frame consists of static, delta and delta-delta components of logarhythmized F0. The context-dependent HMMs are estimated according to the acoustic features and context labels under the maximum likelihood criterion [6]. F0 stream is modeled by a multi-space probability distribution (MSD) [7]. 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 phone HMM [4].
In syllable model training, the mean of logarithmized F0 of voiced frames extracted by STRAIGHT within each syllable is calculated as the syllable-layer F0 feature. The part of unvoiced frames is ignored. The boundaries of each syllable are given by the Viterbi alignment using the trained HMMs of spectrum and F0 for context-dependent phonemes. Instead of modeling in the unit of phone, context-dependent syllable HMMs are estimated according to the acoustic features and labels under the maximum likelihood criterion to give a better description of the super-segmental characteristics of F0. A decision tree-based model clustering method is also applied and the questions at phone layer are removed considering that syllable is the basic unit here. Single-state HMM structure is used for syllable-layer models and avoidance of over-fitting.

MDL (Minimum Description Length) criterion [8] can be adopted to determine the number of clusters automatically. In our training system, we modify MDL factor to control the size of the decision tree, in other words, to decide the number of leaf nodes. Appropriate number of nodes can achieve a balance between the accuracy of trained models and avoidance of over-fitting.

### 2.2. Parameter Generation

For an input text, the HMM sequences of the sentence are determined by the clustered HMMs and decision trees. Firstly, we should predict duration for each state of the sentence. Then spectral and F0 parameters are generated by maximum likelihood parameter generation algorithm [9]. Here, only the syllable model for F0 features is added, so we discuss the F0 parameter generation.

For a given phone HMM $\lambda_{pm}$, to generate speech parameters, we maximize $P(O|\lambda_{pm})$ with respect to the static speech parameter vector sequence

$$C = [c_1^T, c_2^T, ..., c_D^T].$$

$$C_{\text{max}} = \arg \max_C P(O|Q_{\text{max}}, \lambda_{pm}) = \arg \max_C P(W_{pm}C|Q_{\text{max}}, \lambda_{pm})$$

(1)

where

$$O = W_{pm}C$$

(2)

is the complete observation feature sequence containing static and dynamic components and $W_{pm}$ is a window function in order to calculate the dynamic parameters (e.g. delta and delta-delta parameter coefficients). We can get optimization of $C$ by the derivative of

$$\frac{\partial}{\partial C} \log P(O|Q_{\text{max}}, \lambda_{pm}) = 0$$

(3)

We can obtain that:

$$W_{pm}^TU_{pm}^{-1}W_{pm}C = W_{pm}^TU_{pm}^{-1}M_{pm}$$

(4)

where $M_{pm}$ denotes the F0 mean matrix and $U_{pm}$ is the F0 covariance matrix. $Q_{\text{max}}$ is the state and mixture sequence.

After the syllable model is introduced, the Eq. (1) is modified as the following:

$$C_{\text{max}} = \arg \max_C P(O|Q_{\text{max}}, \lambda_{pm})p^k(O|Q_{\text{max}}, \lambda_{sm}) = \arg \max_C P(W_{pm}C|Q_{\text{max}}, \lambda_{pm})p^k(W_{sm}C|Q_{\text{max}}, \lambda_{sm})$$

(5)

where $\lambda_{sm}$ denotes the parameter of syllable HMM model.

To control the contribution of the syllable model in the parameter generation, we add $k$ to be the weight coefficients. Different $k$ can lead to different contribution of the likelihood functions of phone model and the syllable model. An appropriate weight can lead to the best performance. By means of a series of derivation, we can obtain that:

$$\left(W_{pm}^TU_{pm}^{-1}W_{pm} + kW_{sm}^TU_{sm}^{-1}W_{sm}\right)C = W_{pm}^TU_{pm}^{-1}M_{pm} + kW_{sm}^TU_{sm}^{-1}M_{sm}$$

(6)

where $M_{pm}$ denotes the syllable F0 mean matrix, $U_{sm}$ is the syllable F0 covariance matrix. For we only consider the mean of a syllable to represent HMM model, $W_{sm}$ is a window function to calculate the mean of the log F0 of voiced frames within each syllable. It is an $N$-by-$D$ matrix, where $N$ is the number of syllables and $D$ is the number of frames in a sentence. The phone model and syllable model have the joint impact to the output $C$. Finally, we can get the final output $C$. 

---

Fig. 1. Flowchart of HMM-based speech synthesis system
3. EXPERIMENTS

3.1. Experiment Condition

Speech synthesis is experimented on three different emotional Chinese speech databases of a female speaker, including 1253 happy sentences, 890 sad sentences and 1000 neutral sentences in total. Speech signals are sampled at 16 kHz and analyzed by STRAIGHT. The 5-state left-to-right phone HMM model and 1-state syllable HMM model with single-mixture are adopted. We preserve 10 sentences in each set for testing purpose; the others are for training. Objective measure — RMSE (Root of Mean Square Error) and subjective measure — Preference Score are selected to assess the performance.

3.2. Objective Experiment

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

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\log(f_0(i)) - \log(f_0^g(i)))^2}$$  \hspace{1cm} (7)

where $N$ is total frames in a sentence, $f_0(i)$ is original F0 parameter, $f_0^g(i)$ is estimated F0 parameter. Only the frames that are voiced in both natural and generated F0 sequences are used to calculate the RMSE according to Eq. (7).

Experiments are carried out in three emotional sets. From the view of objective measure, we get the RMSE of log F0 on the neutral set with the increase of the syllable weight as Fig.2, where $k=0$ means the baseline system.

Fig.2 shows that the RMSE firstly decreases then increases at the inflexion of $k=15$. So we decide 15 is best weight for this set. When we fix $k=15$, the RMSE varies with the number of nodes of the clustering decision-tree by setting the factor of MDL as shown in Fig.3. From Fig.3 we can find that the RMSE firstly falls then rises at the inflexion of MDL=0.8 and nodes=431. Finally, RMSE improves from 0.086073 to 0.082674 when we add syllable model to the original phone model. Same experiments of happiness and sadness emotions set are also carried out and the results are shown in Fig.4-7.
From Fig.4-7, we observe that RMSE improves from 0.126199 to 0.125433 at the inflexion of MDL factor=0.4 and nodes=606 when we add syllable model to the happy set. Also, RMSE improves from 0.10695 to 0.10555 at the inflexion of MDL factor=1.0 and nodes=190 for sadness.

3.3. Subjective Experiment

Preference Score measures which one is better between two similar sentences. The two sentences are usually synthesized from the same text. The better one gets the score of 1, the other gets 0. The preference score is defined as following:

\[
\text{Preference Score} = \frac{1}{Nm} \sum_{i=1}^{N} S(i)
\]

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

20 sentences are synthesized respectively using the baseline system where only phone-layer F0 model is used and the proposed system for each database. These sentences are evaluated by 4 speech expert listeners pair by pair. The final preference score of these two systems on the three databases are calculated as Eq. 8 and shown in Fig. 8.

3.4. Result Analysis

We can see from the figures above that the proposed model reduced the RMSE in three emotional sets. It has a good performance in sad set when measured with preference score rule. The proposed model has 68.75% preference score to 31.25% in sad set while the others have little different in the subjective evaluation. The syllable model smoothes some abnormal high pitch, which are unavoidable in estimation, to make the whole sentence leaning to sad emotion in perception, whereas the other two sets perform little in objective evaluation because of their original high level of F0. The result shows that syllable model plays a positive role in emotional representation. When we modify an appropriate weight and the number of clustering nodes, the integrative model can perform well.

4. CONCLUSION

In this paper, a multi-layer F0 modeling approach considering the super-segmental characters of F0 for HMM-based speech synthesis is presented. A syllable-layer F0 model is trained to predict the F0 contour together with the phone-layer model. In our experiments on three emotional speech databases, the proposed model reduces the RMSE of predicted log F0 by 3.95% at most compared with the baseline system. A subjective evaluation shows the proposed method can improve the naturalness of synthesized speech effectively on the sad emotion database. Future works include how to extract more informative features at syllable layer, such as the dynamic features among neighboring syllables, the slope of F0, and so on.

5. REFERENCES