A Novel Hybrid Approach for Mandarin Speech Synthesis

Shifeng Pan, Meng Zhang, Jianhua Tao

[1] National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences
{sfpan, mzhang, jhtao}@nlpr.ia.ac.cn

Abstract

The paper investigates a new method to solve concatenation problems of Mandarin speech synthesis which is based on the hybrid approach of HMM-based speech synthesis and unit selection. Unlike other works which use only boundary F0 errors as concatenation cost, a CART based F0 dependency model which considers much context information is trained to measure smoothness of F0. Instead of phoneme-sized units, the basic units of our HUS system are syllables, which has been proved to be better for the prosody stability in Mandarin. The experiments show that the proposed method achieves better performance than conventional hybrid system and unit selection system.

Index Terms: Speech synthesis, hidden Markov model, unit selection, hybrid

1. Introduction

Recently, the hybrid approach of speech synthesis which combines the advantages of both statistical parametric speech synthesis and unit selection speech synthesis, e.g. HMM-based unit selection, has been investigated by many researchers[1]-[6]. The synthesized speech by hybrid approach has more stable prosody performance than unit selection and higher voice quality than statistical parameter approach.

For HMM-based unit selection (HUS), the prosody and acoustic parameters generated by HMMs are used to carry out the selection of unit. During the unit selection stage, target cost and concatenation cost are always two basic criteria. For instance, [1]–[4] directly use spectrum parameters, F0 and durations generated from HMMs as the references to get target costs for the speech candidates, nevertheless, [5]-[7] use HMM likelihoods for the target costs. The concatenation cost is considered in [1]-[5]. In [1]-[4], only the discontinuity of prosody and spectrum between two joint units is calculated. In [5], context dependant and decision tree based clustering models are trained to model the transition of F0 and spectrum at phone boundaries. However, the comment on this model is brief, and the features used to build the decision tree are not mentioned.

Mandarin is a tonal language. The intonation variation of Mandarin speech is much wider than most of the western languages. The pitch is not always smooth due to several reasons, e.g. pause duration, stresses, tone patterns and etc. Thus, the pitch patterns (smoothing or jumping) across the prosody boundaries are very complicated within different context information. The simple bias calculation at the splices of the speech concatenation parts is not enough to ensure the naturalness of the Mandarin speech synthesis.

In this paper, we propose a new concatenation method which is based on a F0 dependency model for Mandarin HUS system. The F0 dependency model is based on CART (classification and regression tree) and considers more context information. It gets a better prosody relationship between joint speech units than other simple computation of prosody bias.

The model has been proved to be effective for Mandarin prosody prediction in our previous unit selection system [8]. In this work, the parameters generated from HMMs are used as the references of target cost while the F0 dependency model is used for the concatenation cost. As most of the HTS systems only use phonemes, the basic units of our HUS system are syllables, which has been proved to be better for the prosody stability in Mandarin. The experiments show that the proposed method improves the naturalness of HMM-based unit selection system. It achieves comparable performance to the conventional unit selection system.

The rest of this paper is organized as follows. In section 2, target cost and concatenation cost of the proposed HMM-based unit selection approach are introduced. The basic unit in acoustic level is also discussed in this part. Section 3 introduces the system framework of the proposed approach. In section 4, evaluation results are presented. The summary of this paper is presented in section 5.

2. HMM-based Unit Selection

2.1. Calculation of Target Cost

In HMM-based speech synthesis system, the prosody and acoustic parameters are generated by HMMs. The parameters include F0, spectrum, power and duration. Generally speaking, there are two basic methods to calculate the target costs. One is measuring the likelihood of the candidate units' parameters to the probability distribution of the target models predicted by HMMs. The other is measuring the distance between the parameters of candidate units and those generated by HMMs. Here the parameters generated by HMMs are the trajectory of F0 and spectrum of target unit, power and duration of target unit.

For the first method, the likelihood is increasing while F0 or spectrum of candidate units is approaching to the mean values of each model. That is to say, this method is to find the candidate unit whose parameter's are nearest to the static values output by each target state models with max probability. For Mandarin speech synthesis, The F0 trajectory is important to both naturalness and intelligibility of synthesized syllable. The latter method takes smooth F0 trajectory generated by HMMs as reference, hence it has the ability to controls the F0 trajectory of synthesized speech directly. Therefore, the latter method is adopted to calculate the target cost.

The target costs are calculated as follows,

$$C_{tgt} = w_1 D_{F_0} + w_2 D_{mcep} + w_3 D_{pow} + w_4 D_{dur} \quad (1)$$

$$D_{F_0} = \sqrt{\frac{1}{N_{F_0}} \sum_{i=1}^{N_{F_0}} \left(F_{0unit}(i) - F_{0tgt}(i) \right)^2} \quad (2)$$

$$D_{mcep} = \sqrt{\frac{1}{N_{F_0} N_{mcep}} \sum_{i=1}^{N_{F_0}} \sum_{j=1}^{N_{mcep}} \left(mcep_{unit}(j,i) - mcep_{tgt}(j,i) \right)^2} \quad (3)$$

$$D_{pow} = | pow_{unit} - pow_{tgt} | \quad (4)$$

$$D_{dur} = \frac{1}{N_{F_0}} | N_{unit} - N_{F_0} | \quad (5)$$

where C_{tgt} denotes the target cost of a candidate unit, D_{F_0} and D_{mcep} denote the F_0 and spectrum distance between candidate unit and target unit respectively, D_{pow} and D_{dur} denote the power and duration deviance between candidate unit and target unit respectively, N_{F_0} and N_{mcep} denote the frame number of target unit and orders of MCEP (mel-cepstrum) respectively, $F_{0unit}(i)$ and $F_{0tgt}(i)$ denote the F_0 value of *i*-th frame of candidate unit and target unit respectively, $mcep_{unit}(j,i)$ and $mcep_{tgt}(j,i)$ are the *j*-th order MCEP at frame *i* of candidate unit and target unit respectively, pow_{unit} and pow_{tgt} denote the power of candidate unit and target unit respectively, N_{unit} denotes the frame number of candidate unit, and w_1 - w_4 are weights. Note that the duration of each candidate unit may be not equal to that of target unit. Therefore the F_0 sequence and MCEP sequences of candidate units need to be normalized to the same length to target unit. Here the linear transformation is used to conduct the normalization.

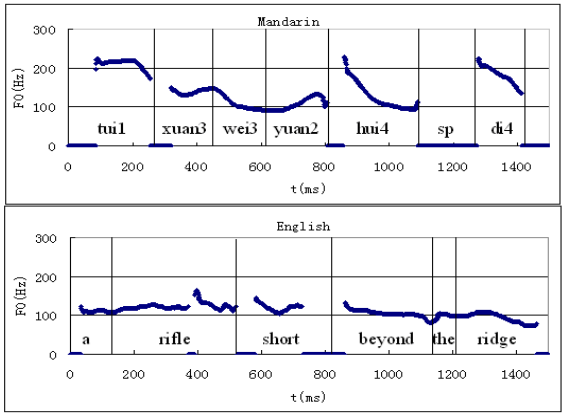

Figure 1: *Examples of F0 contour of Mandarin and English speech. (The numbers after each Mandarin syllable are tone patterns)*

2.2. Basic Unit in Acoustic Level

As we know, Mandarin is a mono syllable and tonal language with wide pitch range and high pitch changing rate. Each syllable has a [Consonant]+Vowel+[Nasal] structure where [·] indicates optional parts, which can be split into initial part ([Consonant]) and final part (Vowel+[Nasal]). Figure 1 shows the F0 contour of a Mandarin speech segment read by a male speaker with medium speed. To compare, the F0 contour of an English speech segment read by a male speaker with medium speed is also given. (Note: These two sentences are chosen randomly from database.)

Generally speaking, the smaller the unit size is, the more flexible the selection of unit is and easier the control of target cost is. In [3] and [9] frame is used as basic unit. However, small unit may be not appropriate for Mandarin speech. The variation range inside a syllable is wide. It is difficult to make the synthesized syllable possess such natural pitch contour as that of natural speech. Too many concatenation points will also make the discontinuity problem of F0 contour inside syllable become prominent. On the other hand, the pitch variation across syllable boundary is determined by the context of the whole syllable such as tone of syllable, or half syllable such as initial part and final part. If we choose syllable as basic unit, the prosody and acoustic parameters inside the syllable are stable and natural. In addition, the context information of target unit (syllable) can be used to calculate the concatenation cost, which can be used to perform the selection of candidate syllables directly. This simplifies the incorporating of concatenation cost in total cost function and the selection of unit. Therefore, syllable is chosen as basic unit in this paper.

2.3. F0 Dependency Model

The naturalness of synthesized speech depends on the F0 contour greatly. When using syllable as basic unit, the continuity of F0 contour inside of unit can be preserved. Then the matching of the F0 contour boundary of the adjacent units is critical to the naturalness of synthesized speech.

Since we use HMMs to model the prosody and acoustic parameters, it seems a good idea to model the joint probability together. Taylor introduced the frame sequence probability algorithm, a means of replacing the traditional concatenation costs by a purely statistical method [10]. This joint model seeks to provide genuine probability that one section of speech follows another. When using this model together with the likelihood based target cost, we will get a complete probability based unit selection.

However, a simple statistics on the occurrence of joint units is not always useful. For Mandarin speech, there are many factors that affect the boundary of syllable's F0 contour, such as the tone of the two syllables, the duration of the voiceless initial part of the latter syllable, etc. Figure 2 shows samples of F0 contour of two adjacent syllables. In subgraph (a) of Fig. 2, the F0 contour is continuous when crossing the syllable boundary, while in other two subgraphs the boundary F0 values have large variations.

For the two adjacent syllables, it seems that these four boundary parameters, the F0 ending value (F_{0_E}) and F0 ending derivative ($F_{0_{ED}}$) of the former syllable and the F0 starting value (F_{0_S}) and F0 starting derivative ($F_{0_{SD}}$) of the latter syllable, have some strong relationship. Even in the situation where there are pitch jumps across the syllable boundary, it seems that F0 contour is virtually connected across the silence and voiceless initial, as Fig. 2 shows. Fig. 3 shows these four parameters in one syllable. In this paper, the CART based F0 dependency model is adopted to predict these boundary parameters of F0 contour [8]. Parameters involved in the concatenation cost include F_{0_E}, $F_{0_{ED}}$, F_{0_S} and $F_{0_{SD}}$. When we predict these four parameters, features listed in Table 1 are used.

2.4. Calculation of Concatenation Cost

The predicted value by CART can be considered as the expected boundary value by adjacent syllables. Therefore the difference between these predicted values and actual values can be used to measure the concatenation cost of F0. For

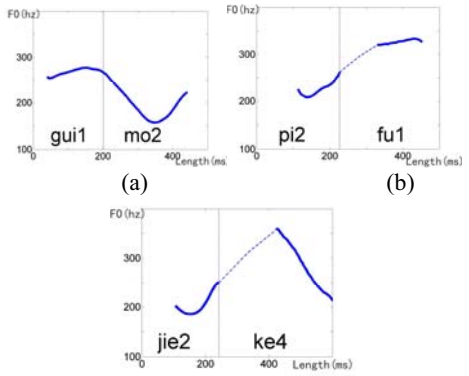

Fig 2: *Examples showing the impact of adjacent syllables' F0 contours on the current one [8].*

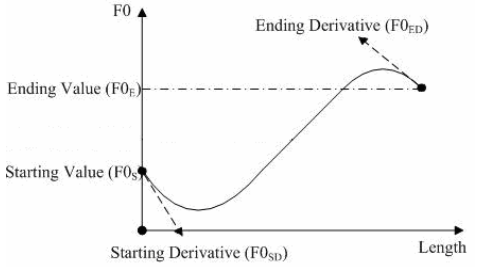

Fig 3: *Four boundary parameters used in calculating concatenation costs [8].*

Table 1. *Features used in predicting* F_{0_E}, $F_{0_{ED}}$, F_{0_S} *and* $F_{0_{SD}}$

Features in predicting F_{0_S} and $F_{0_{SD}}$	Features in predicting F_{0_E} and $F_{0_{ED}}$
frequently used text information (tone, initial/final identity, prosody structure, etc)	frequently used text information (tone, initial/final identity, prosody structure, etc)
previous syllable's F_{0_E} and $F_{0_{ED}}$	following syllable's F_{0_S} and $F_{0_{SD}}$
pause length before current syllable	pause length after current syllable
current syllable's final part length	following syllable's initial part length

spectrum, the continuity of spectrum across the concatenation point can be used to measure the concatenation cost of spectrum. Therefore, the total concatenation cost can be calculated as Eq. (6) demonstrates.

$$C_{con} = w_5 * DF_{0_S} + w_6 * DF_{0_E} + w_7 * DF_{0_{SD}} + w_8 * DF_{0_{ED}} + w_9 * D_{spec} \quad (6)$$

where DF_{0_S} denotes the difference between the predicted and actual F_{0_S}, it is the same for DF_{0_E}, $DF_{0_{ED}}$ and $DF_{0_{SD}}$, D_{spec} denotes the distance between the boundary spectrum of the two candidate units to be concatenated, and w_5 - w_9 are weights.

As we know, it is difficult to automatically compute the F0 boundary value accurately near both sides of unvoiced section. In other words, the automatically computed F0 boundary values are usually not accurate, which will degrade the accuracy and robustness of the model. To overcome this problem, we use F0 values of five boundary frames to fit a second-order polynomials to represent the true boundary F0 contour. Then $F_{0_{SD}}$ and $F_{0_{ED}}$ can be derived from the fitted polynomials.

3. System Framework of HMM-based Unit Selection

The framework of the proposed HMM-based unit selection system is shown in Fig. 4. During the training stage, the context-dependant HMMs, duration models and CART based F0 dependency model are trained. During the synthesis stage, firstly, the front-end analyzes the input text and context information is derived. Then the sentence HMMs is decided according to the context information. Sentence HMMs are used to generate the prosody and acoustic parameters, including trajectory of F0 contour and spectrum contour, duration and power. The boundary F0 parameters are derived according to the context information of input text and the features of each candidate unit. Then the target cost of each candidate unit and the concatenation costs between each pair of adjacent candidate units can be calculated. The optimal candidate units are selected by Viterbi search. Finally the synthesized speech is generated by concatenating the selected units. For saving the calculating cost, a KLD based pre-selection can be performed before the unit selection stage [6]. The KLD defined as the divergence between the state model which the candidate frame is tied to and the state model predicted by HMMs. For that syllable is chosen as the basic unit, the KLD of candidate syllable is the sum of state-level KLD that belong to the candidate syllable.

4. Experiments

4.1. Experiment Conditions

To carry out the experiment, the proposed HMM-based unit

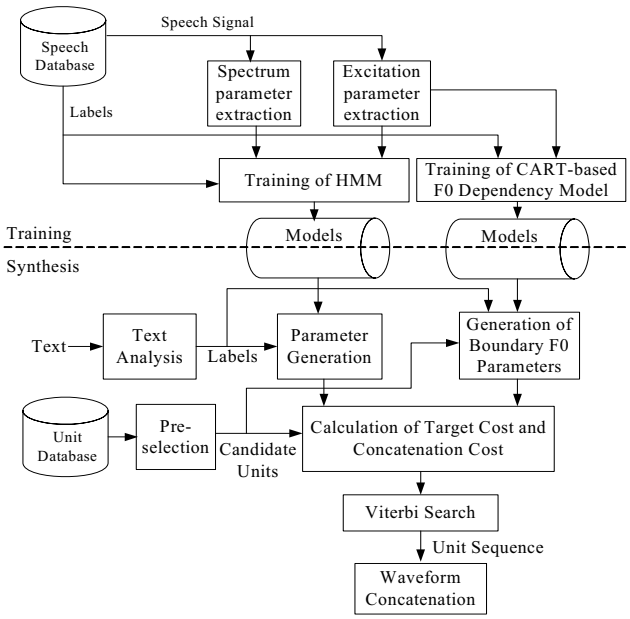

Fig 4: *Framework of the proposed HMM-based speech synthesis system*

selection speech synthesis system was built. The whole corpus used to build the system consisted of 10,000 phonetically balanced Chinese sentences. The database used for HMM training consists of 2000 phonetically balanced Chinese sentences chosen from the corpus (our computer could not conduct HMM training on larger database). The whole corpus was used to train CART based F0 dependency model. In HMM training stage speech signal was windowed by a 25ms hanning window with a 5 ms shift, and the mel-cepstrum order was 25 (including 0-order). 5-state left to right with no skip HMM structure was adopted to model each phoneme of Mandarin. The F0 trajectory and spectrum trajectory were generated by MSD-HMM [11]. The weights w_1 - w_9 for calculating target costs and concatenation costs are set by empirical values.

For comparison purpose, two other systems were also built. One system was quite the same to the above system except that the concatenation cost was ignored. The same 2000 sentences were used to train HMMs. Another one was the conventional unit selection system—WISTON speech synthesis system [12], where the CART based F0 dependency model was also used for calculating of concatenation cost. The same 10,000-sentences database was used to train the system and build the unit inventory.

4.2. Subject Evaluation

In the subject evaluation, AB test was adopted. The first experiment was conducted between the proposed HMM-based unit selection speech synthesis system (system A) and the same HMM-based unit selection system without concatenation cost (system B). The purpose of conducting this experiment was to investigate the significance of incorporating concatenation cost into HMM-based unit selection system. The second experiment was conducted between the proposed HMM-based unit selection system and the conventional unit selection system (system C). The purpose of conducting this experiment was to investigate the performance of proposed system versus the state of art unit selection system. 20 sentences were synthesized by each of the three systems. 11 speech experts were asked to participant the evaluation, which are all under-graduate university students. Each listener was required to make a decision from each pair of sentences which one sounds more natural. The results of the evaluation are shown in table 2.

As seen from Table 2, system A exceeds system B greatly, which means the concatenation cost is very important to the performance of HMM-based speech synthesis. Comparison of system A with system C shows that the former exceeds the latter a little. Analyzing the synthesized voices and evaluation results of these two system shows that synthesized voice of system C is more expressive than system A when there are very appropriate units in unit inventory, which is usually preferred to by listeners. However, the voice of system A is more steady and smooth, and the occurrence of inappropriate unit is lower than that of system C, which is also preferred to by listeners. The above two factors mean that none of the two systems is completely better than the other. However, the overall results show that the synthesized voice of proposed system is slightly better than conventional unit selection system.

Table 2. *AB test results*

(A: HMM-based unit selection system. B: HMM-based unit selection system without concatenation cost. C: convention unit selection system.)

System	System A	System B
Preference(%)	81	19
System	System A	System C
Preference(%)	53	47

5. Conclusion

In this paper, a hybrid approach of HMM-based speech synthesis and unit selection for Mandarin speech synthesis is investigated. The analysis of F0 contour of Mandarin speech indicates that choosing syllable as basic unit is more preferable than smaller size. During unit selection stage, the CART based concatenation cost and acoustic and prosody distance based target cost calculation were adopted. Subject evaluation results show that the concatenation cost is important for HMM-based unit selection. The results also show that the proposed HMM-based unit selection method achieves better performance than conventional unit selection method.

6. Acknowledgment

The work was supported by the National Science Foundation of China (No. 60873160), 863 Programs (No. 2009AA01Z320) and China-Singapore Institute of Digital Media (CSIDM).

7. References

[1] H. Kawai, T. Toda, J. Ni, M. Tsuzaki, and K. Tokuda, "XIMERA: A new TTS from ATR based on corpus-based technologies," ISCA SSW5, 2004.

[2] S. Rouibia and Rosec, "Unit selection for speech synthesis based on a new acoustic target cost," Interspeech, 2005, pp. 2565–2568.

[3] T. Hirai and S. Tenpaku, "Using 5 ms segments in concatenative speech synthesis," ISCA SSW5, 2004.

[4] J.-H. Yang, Z.-W. Zhao, Y. Jiang, G.-P. Hu, and X.-R. Wu, "Multitier non-uniform unit selection for corpus-based speech synthesis," Blizzard Challenge Workshop, 2006.

[5] Z.-H. Lin, R.-H. Wang, "HMM-based hierarchical unit selection combining Kullback-Leibler divergence with likelihood criterion," ICASSP, 2007, pp. 1245-1248.

[6] Z.-H. Lin, R.-H. Wang, "HMM-based hierarchical unit selection combining Kullback-Leibler divergence with likelihood criterion," ICASSP, 2007, pp. 1245-1248.

[7] Z.-H. Lin, R.-H. Wang, "Minimum unit selection error training for HMM-based unit selection speech synthesis system," ICASSP, 2008, pp. 3949-1352.

[8] J. Yu and J.-H. Tao, "A novel prosody adaptation method for mandarin concatenation-based text-to-speech system," Acoust. Sci. & Tech., 2009, pp. 33–41.

[9] Z. Ling and R. Wang, "HMM-based unit selection using frame sized speech segments," Interspeech, 2006, pp. 2034–2038.

[10] P. Taylor, "Unifying unit selection and hidden Markov model speech synthesis," Interspeech, 2006, pp. 1758–1761.7.

[11] K. Tokuda, T. Masuko, N. Miyazaki,and T. Kobayashi, "Multi-space probability distribution HMM," IEICE Trans. Inf. & Syst., vol. E85-D, no. 3, pp. 455–464, Mar. 2002.

[12] J.-H. Tao, J. Yu, etc , "The WISTON text to speech system for blizzard 2008," Blizzard Challenge Workshop, 2008.