

Integration of Intonation in F0 Trajectory prediction using MSD-HMMs

Xiaojun Zou, Xiao Bao and Lidong Luo

Speech and Hearing Research Center (SHRC)
Key Laboratory of Machine Perception (Minister of Education)
Peking University, Beijing, P. R. China
{zouxj, baox, luold}@cis.pku.edu.cn

Abstract

Present study in speech synthesis places more and more emphasis on the spectral continuities and diverse prosodic effects. The trainable HMM-based speech synthesis method tends to generate more continuous spectral structures than the traditional unit selection method. However, the F0 trajectory generated by HMM-based speech synthesis is often excessively smoothed and lacks prosodic variance. This paper proposed an approach to improve the effect of F0 trajectory prediction in mandarin speech synthesis in the framework of multi-space probability distribution HMMs (MSD-HMMs). In the proposed approach, the intonation, which is predicted by context-dependent decision trees, is integrated to the F0 trajectory generated by the MSD-HMMs as a weighted bias term. The experiments indicate that it has an encouraging improvement in the prosodic effectiveness of Mandarin speech synthesis.

Index Terms: Mandarin speech synthesis, MSD-HMMs, Prosody, Intonation, Tone, Register

1. Introduction

Prosody is an inherent supra-segmental feature [1] of human speech and is employed to convey many complex meanings of the speaker in daily speech communications, such as emotion, stress, preferences, assumptions, etc. In linguistics and language processing, the term prosody is employed to express the intonational and rhythmic aspects of language and has been studied by linguists, phoneticians, speech therapists for many decades [1-3]. In speech synthesis, prosody, which is related to pause, pitch, speech rate and loudness, plays an especially important role, because it greatly affects and determines the naturalness and intelligibility of the synthesized utterance [2]. Of all the features that represent and describe the prosodic structure, pitch or fundamental frequency (in this paper, we considers pitch and fundamental frequency (F0) are identical) is the most important one.

In contrast to English, a typical accent language, Mandarin is a tonal syllabic language, in which lexical tone plays an important role in determining word meaning. In Mandarin, lexical tone of local syllable and global intonation of a sentence, together with other hierarchical components have a complex interaction to compose the prosodic structure. The relationship between lexical tone of local syllable and global intonation of a sentence are vividly depicted as “the small ripples riding on top of large waves” [3].

Several models have been put forward to consider the intonation influence on the pitch contour generation. Fujisaki model [4] considers the changes on F0 contours are caused by phrase component and accent component. When modeling Chinese, the phrase component describes the intonation

information as in English and Japanese while the accent component with both positive and negative amplitudes suggests the lexical tone variations [5]. Though Fujisaki model performs excellently in analysis-synthesis process, the fine pitch structures of Chinese lexical tones are difficult to acquire by the combination of the positive and negative accent components in the model, because the combination is not syllable-synchronized and is not stable to be modeled.

The Target Approximation (TA) model [6] proposed by Yi Xu is syllable-synchronized and is focused on how lexical tones of Chinese were produced and perceived in continuous speech. In [7], a quantitative model of TA model is described. However, the TA model needs labels on the onset and offset of the pitch targets, and is difficult to implement on training speaker dependent prosodic styles. So the trainable HMM-based speech synthesis [8] stands out with its statistical method on large corpora and few manual labels.

In HMM-based speech synthesis, F0 values are regarded as the training parameters of a specified phoneme and the HMM training method is extended to MSD-HMMs method [9] to model both the voiced and unvoiced regions of the F0. To separate phonemes with different linguistic feature, a host of contextual information is considered to model phonemes in the HMM scheme. Though this method is useful when training data is extremely large, the data sparsity problem always exists. Therefore, to generate rich pitch variation, global intonation should be considered. In [10] four pitch level of each syllable is proposed as intonation information, but value of pitch level is added directly to the f0 contour generated by HMM scheme without smoothness. In [11] and [12], a multilevel HMM scheme is elaborated, in which contributions from multi-levels such as syllable-level, phrase-level are calculated in an EM iteration. However, this method does not solve the data sparsity problem and causes high complexity.

This paper proposed an approach to improve the effect of F0 trajectory prediction in Mandarin speech synthesis in the framework of MSD-HMMs. In this approach, the intonation, which is predicted by context-dependent decision trees, is integrated to the F0 trajectory generated by the MSD-HMMs as a bias term. The experiments indicate that it has an encouraging improvement in the prosodic effectiveness of Mandarin speech synthesis. The rest of this paper is organized as follows. In the second section, the modeling of intonation for the proposed pitch generation approach will be presented. The method for integrating intonation into the MSD-HMMs based pitch generation algorithm will be elaborated in Section 3. In the discussion section, we conclude the proposed method and discuss the further work.

2. Modeling of Intonation in Mandarin

Intonation is the melodic pattern of a language; it is universally regarded as the F0 contour of a sentence. Taking

English for example, it is falling for declarative sentences while rising for interrogative ones. But the Chinese intonation is rather complex than English and there are still various controversies, because there is a complex interaction between the intonation and lexical tone—on one hand, Chinese is a typical tone language, with each syllable having certain F0 contour which signals lexical distinction, while on the other hand, the local F0 contour must be modified to meet the requirement of intonation [13].

Cao (1998) proposed a theory to decompose the surface F0 into pitch contour and pitch register. In her definition, the pitch contour of each syllable or lexical item is observable from its spectrogram while the pitch register is the average pitch value. Further, the trend of pitch movement for each phrase or sentence can be defined as drawing through the register values of the first and last syllables in a certain phrase or sentence.

In order to achieve a better capture of prosodic structure, Heiga Zen etc. also proposed a context-dependent multi-layer log F0 model for HMM-based speech synthesis; however, it is computationally expensive and does not work well as expected. In the evaluation, the model takes all the context-dependent layers into consideration performs even worse than the baseline; on the contrary, the model’s simplified vision (just considers syllable- and HMM-state-level contexts) achieves an encouraging result.

Inspired by Cao’s theory and Zen’s work above, we define syllable register as the mean of pitch value at the corresponding syllable. Note that it is the same as Cao’s pitch register in definition, but slightly different in computation. We use the fluctuating trends of the syllable register to represent the global intonation of the sentence; as the register fluctuation determines the pitch movements of a sentence or phrase, it can be used as the global prosodic information, or exactly the intonation of the sentence or phrase, in Mandarin speech synthesis. Then many elaborate statistical machine learning algorithm can be adopted to learn and predict the syllable register at any given context. In the proposed method, we adopt classification and regression tree (CART) [14] based approach to predict the syllable register, or the intonation of the sentence, considering the value to be predicted for each syllable is continuous quantity other than discrete tags.

3. Integration of Intonation in MSD-HMMs

In HMM-based speech synthesis, the spectrum, pitch, and duration are simultaneously modeled by the corresponding HMMs. For a given text, speech parameters are generated from the trained HMMs in such a way that the output probability is maximized and this approach tends to generate more continuous spectral structures than the traditional unit selection method. However, the F0 trajectory generated by HMM-based speech synthesis is often excessively smoothed and lacks prosodic variance. Tomoki Toda *et al.* proposed a parameter generation algorithm which considers global variance (GV) of the generated parameters to remedy the excessive smoothness of the F0 trajectory. [15] Though this approach improves the naturalness of the synthesized speech in some way, it fails to consider the phonetic part that accounts for the muffledness or unnaturalness of the speech.

As mentioned in [16], in MSD-HMMs based speech synthesis, it seems essential to train context dependent HMMs with feature vector which consists of spectrum, pitch and their dynamic features. In this case, the F0 pattern in MSD-HMMs is modeled in the same scale of spectrum—both of them are

modeled in the scale of phonetic-level HMM states. It then fails to model the F0 patterns in a higher-level scale for that matter, especially, the global intonation of the sentence is not used, because the higher levels in prosodic structure cannot easily be modeled by the phonetic-level HMM states.

In this paper, we proposed an improvement in F0 trajectory prediction in the MSD-HMMs framework by integrating the global intonation of sentence. As is mentioned in section 2, we use the fluctuating trends of the syllable register to represent the global intonation of the sentence. The schematic diagram is shown in Figure 1.

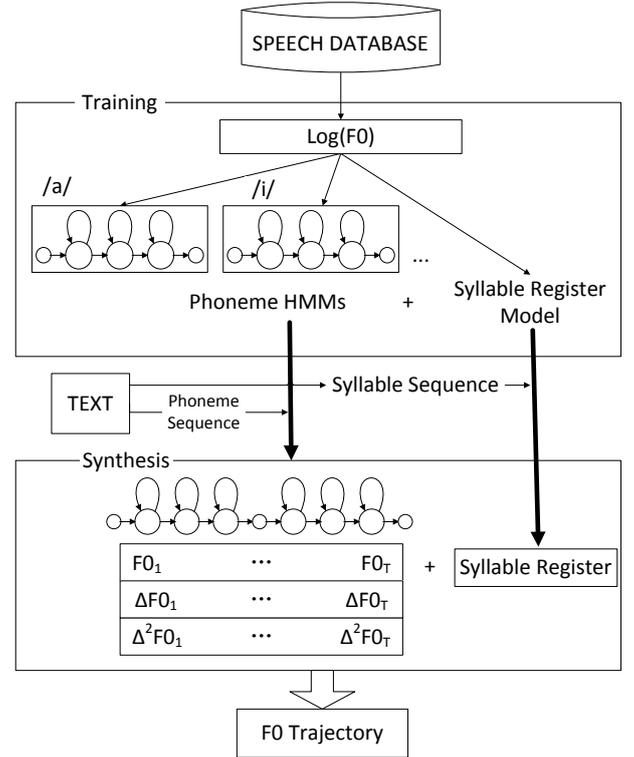


Figure 1: Schematic diagram for the integration of intonation in the framework of MSD-HMMs

First, both the MSD-HMMs and syllable register model are trained using the speech database and the manually labeled text. Then the parameters of F0 trajectory, containing both static and dynamic features (the delta and delta-delta features) are generated by the trained MSD-HMMs model; at the same time, the syllable register, which represents the intonation of the sentence, is predicted by the trained syllable register model. Finally, the syllable register, or the intonation of the sentence, is added to the mean value of the Gaussian distribution of the F0 features as a weighted bias term and the dynamic features of the F0 trajectory keep the original values. The improved approach of generating the F0 trajectory can be illustrated as

$$F0_{new} = w \times register + (1 - w) \times mean_{MSD-HMM}, \quad (1)$$

where *register* is predicted in the syllable register model, $mean_{MSD-HMM}$ denotes the mean value of the Gaussian distribution generated in MSD-HMMs model, and w is a weight between 0 and 1.

The new pitch pattern generated in this approach encompasses both the phone-level information (note that the states in the HMM is modeled in the phone-level) and global

intonation of the sentence, and as the dynamic features are not changed, the F0 trajectory would keep smooth at the joint of adjacent phones. Thus, the new F0 trajectory will be much closer to the original speech than that generated by the conventional MSD-HMMs, and the naturalness of the speech will also be notably improved.

4. Experiments & Results

The experiments are based on the Mandarin database released for Blizzard Challenge 2009 [17], of which 1000 sentences are used—900 sentences for training and 100 sentences for testing. All the sentences are uttered by one professional female speaker in standard Mandarin Chinese. The speech signals are sampled at 16 kHz, encoded with 16 bits PCM, and windowed by a 25-ms window shifted every 5-ms. To get accurate F0 values, the F0 extraction method in STRAIGHT [18] is adopted, in which F0 is 0 if the frame is unvoiced. The baseline Mandarin Text-to-Speech system is developed by SHRC, Peking University, based on MSD-HMMs [19]. The contextual features used for training the CART based syllable register model are showed in table 1 (there are 15 features selected in total).

Features Selected
The syllable info. (the forward, current and backward)
The tone info. (the forward, current and backward syllable)
The prominence info. (the forward, current, and backward syllable)
Position of the current syllable in the current prosodic word (from forward and backward)
Position of the current syllable in the current prosodic phrase (from forward and backward)
The boundary type before current syllable (4 types in total)
The boundary type after current syllable (4 types in total)

Table 1. Contextual features used in CART-based syllable register model training

Two objective and one subjective measures are used to evaluate the performance of the approach proposed in this paper. The two objective measures are F0 distortion in the root mean squared error (RMSE) and the correlation coefficient between the natural (the speech of the original speaker) and synthesized F0 trajectory over aligned frames. In the subjective evaluation, a preference listening test on the synthesized speech was conducted.

In order to compare the performance of MSD-HMMs (the baseline) and the improved approach clearly, we trained nine models using 100, 200, ..., 900 sentences respectively for each system.

Figure 2 shows the RMSE comparison of the two systems between the natural F0 trajectories as the training sentences increase from 100 to 900. It can be clearly seen from the figure that the improved approach gains a notable RMSE decline; especially when the training data is not sufficient, the improved approach has a significant better performance—when using 100 sentences as training data, the RMSE of the baseline is 33.39 Hz while the improved approach is 31.98 Hz (a decline of 4.2%); when training sentences is 900, the RMSE is 29.22Hz and 28.76Hz

respectively (a decline of 1.6%). Figure 3 shows the number of better sentences in RMSE among all the 100 test sentences; for example, of the 100 testing sentences, the improved approach has 87 sentences with lower RMSE than the baseline while the baseline system synthesizes 13 sentences with lower RMSE.

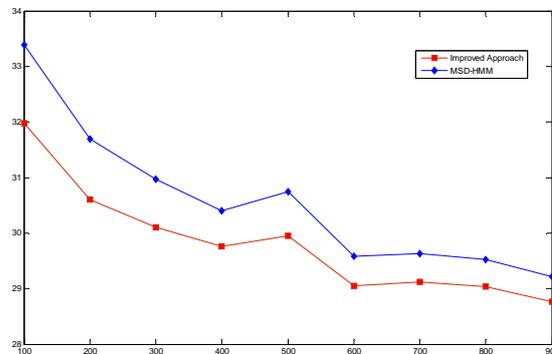


Figure 2: The RMSE comparison of the two systems as the training sentences increase from 100 to 900

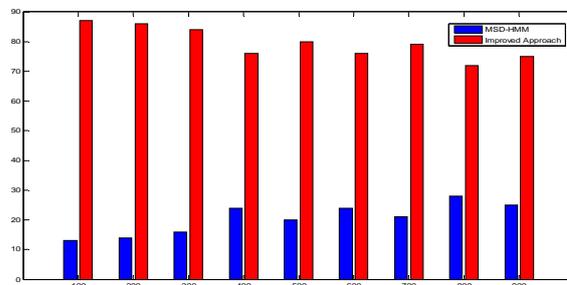


Figure 3: The number of better sentences in RMSE among all the 100 test sentences as the training sentences increase from 100 to 900

Figure 4 shows the correlation coefficient comparison of the two systems between the natural F0 trajectories, as the training sentences increase from 100 to 900. The bar chart in Figure 5 shows the number of better sentences in correlation coefficient among all the 100 test sentences.

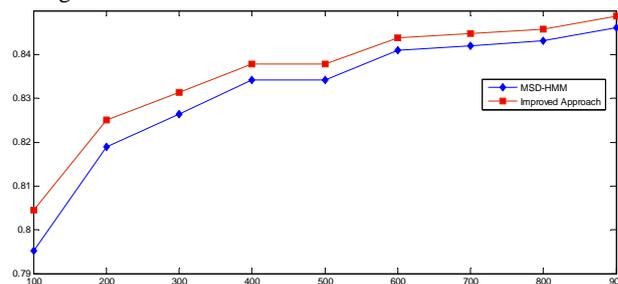


Figure 4: The correlation coefficient comparison of the two systems as the training sentences increase from 100 to 900

It can be discerned from the two figures that the correlation coefficient gains a steady increase with training data, and that the improved approach has a better correlation coefficient score in all of the nine models. However, this improvement in correlation coefficient gets slighter as the training sentences increase (1.2% and 0.3% when the training sentences are 100 and 900 respectively); one explanation is that the correlation coefficient of the baseline is high enough and it prevents significant incensement. But the number of sentences that have better correlation coefficient score gains an encouraging

improvement (as is shown in Figure 5), and this improvement remains stable as the training sentences increase; in average, the improved approach synthesizes 69.6 sentences that has better correlation coefficient score than the baseline.

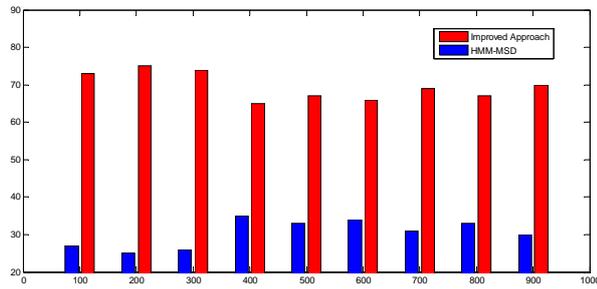


Figure 5: The number of better sentences in correlation coefficient among all the 100 test sentences as the training sentences increase from 100 to 900

The subjective evaluation compared the baseline and the improved approach systems over 60 sentences randomly selected from the 100 test sentences mentioned above (using 900 sentences as training data). The subjects were six native Mandarin speakers, among them four are master students majoring in spoken language processing and two are senior students. Forty sentences were randomly chosen from the evaluation sentences for each subject. Samples were presented in a random order for each test sentence and the subjects were asked to choose their preferred one. Note that the subjects could select "No Preference" if they had no preference. Table 2 shows the preference test results. It can be seen from the table that the improved approach has a significant higher preference score than the baseline, we can conclude that the improved approach has the ability to generate more natural speech than the baseline because it has contained the intonation information in F0 generation.

Baseline	Improved Approach	No Preference
32.38	50.95	16.67

Table 2. The preference score (%) between the MSD-HMMs (baseline) and the improved approach

5. Conclusions

This paper proposed an approach to improve the effect of F0 trajectory prediction in mandarin speech synthesis in the framework of MSD-HMMs. In the proposed approach, the intonation, which is predicted by context-dependent decision trees, is integrated to the F0 trajectory generated by the MSD-HMMs as a weighted bias term. The F0 trajectory generated in this way encompasses both the phone-level information and global intonation of the sentence. Thus, the new F0 trajectory will be much closer to the original speech than that generated by the conventional MSD-HMMs, and the naturalness of the speech will also be notably improved. The experiments indicate that it has an encouraging improvement in the prosodic effectiveness of Mandarin speech synthesis in both objective and subjective evaluations.

As we know, there is far more prosodic information in spoken language than intonation; then future work will focus on exploring more elaborate model that can provide more

accurate prosodic prediction in the generation of F0 trajectory in speech synthesis.

6. Acknowledgements

The authors would like to thank USTC iFLYTEK CO., LTD. for the Mandarin corpus they released in Blizzard Challenge 2009. Special thanks to Xian Xingchi and all the other members of the speech synthesis group in SHRC, Peking University for their great help and support.

7. References

- [1] D. R. Ladd, *Intonational Phonology*, Cambridge University Press, 1996.
- [2] C. Lee, H. Li, L. Lee, R. Wang, Q. Huo, *Advances in Chinese spoken language processing*, World Scientific Publishing Co. Pte. Ltd., 2007.
- [3] A. Cruttenden, *Intonation*, 2nd Edition, Cambridge University Press, 1997.
- [4] H. Fujisaki, K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentence of Japanese", *J. Acoustic. Soc. Jpa* (E), 5(4), pp. 233-242, 1984.
- [5] C. Wang, H. Fujisaki, S. Ohno, T. Kodama, "Analysis and Synthesis of the Four Tones in Connected Speech of the standard Chinese Based on a Command-Response Model", Proc. of *Eurospeech*, 1999.
- [6] Y. Xu, Q. E. Wang, "Pitch Targets and Their Realization: Evidence from Mandarin Chinese, Speech Communication", 33(2001), pp. 319-337, 2001.
- [7] S. Prom-on, Y. Xu, B. Thipakorn, "Modeling Tone and Intonation in Mandarin and English as a Process of Target Approximation", *The Journal of the Acoustical Society of America*, 125(1), pp. 405-424, 2009.
- [8] K. Tokuda, T. Kobayashi, S. Imai, "Speech parameter generation from HMM using dynamic features", Proc. of *ICASSP*, pp. 660-663, 1995.
- [9] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Multi-space Probability Distribution HMM", *IEICE Trans. Inf. & Syst.*, E85-D (3), pp. 455-464, 2002.
- [10] L. Luo, X. Xian, "Integration of intonation in trainable speech synthesis", In *SP-2008*, pp. 75-78.
- [11] J. Latorre, M. Akamine. "Multilevel Parametric-base F0 Model for Speech Synthesis", in Proc. *Interspeech*, pp. 2274-2277, 2008.
- [12] H. Zen, N. Braunschweiler, "Context-dependent additive log F0 model for HMM-based speech synthesis," in Proc. *Interspeech*, pp.2091-2094, 2009.
- [13] J. Cao, "Intonation Structure of Spoken Chinese: Universality and Characteristics", in *From Traditional Phonology To Modern Speech Processing*, 2004.
- [14] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*, Chapman & Hall, New York, 1993.
- [15] T. Toda and K. Tokuda, "Speech Parameter Generation Algorithm Considering Global Variance for HMM-Based Speech Synthesis", in Proc. of *Interspeech*, 2005.
- [16] 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.
- [17] http://www.synsig.org/index.php/Blizzard_Challenge_2009.
- [18] H. Kawahara, I. Masuda-Katsuse, A. Cheveign, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency based F0 extraction: Possible role of a repetitive structure in sounds", *Speech Communication*, pp. 187-207, 1999.
- [19] H. Zen, T. Toda, M. Nakamura and K. Tokuda, "Details of Nitech HMM-based Speech Synthesis System for Blizzard Challenge 2005", *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 1, pp. 325-333, 2007.