# Integration of Intonation in Trainable Speech Synthesis

*Lidong Luo and Xingchi Xian*

Speech and Hearing Research Center
State Key Laboratory on Machine Perception
Peking University, Beijing, China
{luold; xingcx}@cis.pku.edu.cn

## Abstract

Current developments in artificial speech synthesis place more emphasis on spectral continuities and diverse prosodic effects. The trainable HMM-based speech synthesis method has generated more continuous spectral structure than unit selection method in recent study, but the pitch contour generated by HMM-based method trends to be over-smoothed and lacks syllable variance in Chinese. In this paper, to synthesize speaker dependent speech with specific prosodic style, we model the global intonation in Chinese on the syllable scale with definition of pitch level and use pitch level prediction by statistical method to improve the prosodic effects of speech generated by the HMM-based synthesis method.

## 1. Introduction

Prosody is employed to express attitude, assumptions and attention in daily speech communication and has been studied by linguists, phoneticians, speech therapists for many decades [1] [2]. In recent artificial intelligence developments, people seek to communicate effectively with intelligent machines on a more personal and human level. To synthesize natural and human-sounding speech by computers, prosody plays an important role, which is related to pause, pitch, speech rate and loudness. Among the factors which weave the prosody, pitch or fundamental frequency (in this paper we consider pitch and fundamental frequency (F0) as the same) is the most characteristic.

As a tone-language, Chinese exhibits a more complex prosodic structure, especially the pitch contour which contains lexical tone and intonation information. The relationships between the lexical tone of local syllable and global intonation of a sentence are described as "the small ripples riding on top of large waves" [3].

Fujisaki model [4] has given an accurate approximation to observed F0 contours by revealing the physiological and physical evidence of the voicing process. In the model, 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 and 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 relationship between its parameters and the linguistic information is difficult to train and predict. Moreover, the fine pitch structures of Chinese lexical tones are difficult to be acquired by the combination of the positive and negative accent components in the model.

Yi Xu has focused on how lexical tones of Chinese were produced and perceived in continuous speech and has proposed the Target Approximation (TA) model [7] which considers the segmental phonemes, tones, and pitch accents as abstract units called pitch targets. In Chinese, pitch targets are separated into static targets-[high] and [low], and dynamic ones-[rise] and [fall], which are associated with the four lexical tones respectively. This model gives a more accurate description of lexical tone variations in the syllable than the Fujisaki model. However, the TA model needs labels on the onset and offset of the pitch target, and is difficult to implement on training speaker dependent prosodic styles. So the trainable HMM-based speech synthesis [8] stands out with its statistics method on large corpses 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-HMM method [9] where F0 can be considered as either voiced or unvoiced. HMM-based Chinese speech synthesis has been studied in [10]. Linguistic information such as part-of-speech, position of phonemes in syllables, or phrase and prosodic information such as ToBI labels are considered as different contextual features to model phonemes in the HMM scheme. The spectral parameter, log(F0) and their delta and delta-delta parameters are used to train the HMM model of each phoneme. When synthesizing pitch contour, this framework generates a sequence of F0 values for the maximum likelihood. In Chinese, this method figures out a good description on tone variations but cannot predict the whole intonation of a sentence or prosodic phrase because it just generates the F0 contour of phonemes for local optimization.

In this paper we propose an prosodic module to model the intonation for a specified speaker speaking Chinese-Mandarin. We assign a pitch level index to each syllable to model the pitch variations among syllables in one sentence. A statistical method is implemented to predict the pitch level indexes for a given sentence in text format. Then the pitch level indexes are converted to F0 values which can be considered as the intonation information and can be added to the F0 values generated by HMM-based method. In the second section, the pitch level is defined and the method for predicting the pitch level from linguistic information is elaborated. The HMM-based synthesis framework for Chinese is in Section 3 and the method for integrating intonation into HMM-based synthesis is presented in Section 4. In the discussion section, we conclude the proposed method and discuss the further work and other alternative approaches of the proposed method.

## 2. Modeling of Intonation

HMM-based synthesis method trains HMM model for each phoneme. Limited by training data and model size, the pitch

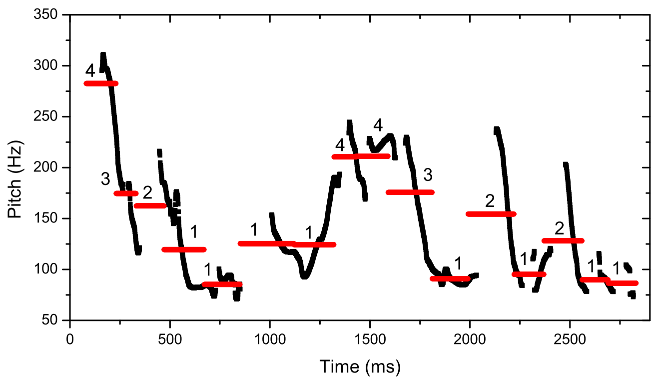


Figure 1: *Pitch level labels in sentence 'zhe4 ci4 chu1 ban3 de0 shi4 mao2 ze2 dong1 chuan2 de0 jian4 guo2 hou4 bu4 fen0'; the curve is the F0 contour; the vertical dash lines are mean F0 values of each syllable and the numbers are the pitch levels (4-level is used).*

contour generated by HMM method is plain in intonation as shown in the center plot of Figure 2. Inspired by Fujisaki model which divides pitch contour into phrase component and accent component. We model the intonation on syllable scale in Chinese. To reflect the pitch variations of different syllables in one sentence, the definition of pitch level is introduced.

### 2.1. Definition of pitch levels

The present study is based on the corpus of a male speaker, which contains 2000 sentences. The sentences are not segmented by punctuation in text but by salient silence in speech. The speech signals are sampled at 16 kHz, and encoded with 16 bits PCM. The word segmentation information and syllables with lexical tones are given and the syllables are automatically segmented by speech recognition method. To get accurate F0 values, the F0 extraction method in STRAIGHT [11] is adopted, in which F0 is 0 if the frame is unvoiced and the frame shift is 1ms. With the exacted F0 values and syllable boundaries, the mean F0 value of each syllable is calculated as follow:

$$F_{mean}(i) = \frac{1}{T(f(k) \neq 0)} \sum_{k=t_{onset}}^{t_{offset}} f(k) \quad (1)$$

where $F_{mean}(i)$ is the mean F0 value of syllable $i$ ; $f(k)$ is the F0 value of the $k_{th}$ frame; $t_{onset}$ means the beginning frame of the syllable and the $t_{offset}$ means the ending frame; $T$ is the number of frames in which F0 value is not 0 because the consonants are often unvoiced.

Figure 1 illustrates the calculation of mean F0 values of syllables. we consider the mean F0 values as the global intonation of the sentence and the pitch variation within one syllable as local information.

The average of maximum mean F0 values over 2000 sentences $(F_{mean})_{amax}$ = 247.6Hz, the average of minimum ones $(F_{mean})_{amin}$ = 86.9Hz; the standard deviations are 12.7Hz and 40.9Hz respectively. So we consider the pitch range of this speaker is between 247.6Hz and 86.9Hz, and the range is equally divided into several areas (Level Area 1-N with F0 increases). If the mean F0 value of a syllable falls in Level Area 1, the pitch level of this syllable is assigned to Index 1.

How many pitch levels should be used? At least 2 levels are needed for high and low intonation. And more pitch levels can depict the pitch variations among syllables more vividly. However, with the number of pitch levels increasing, the prediction precision of the pitch level index will decrease, which is listed in Table 2. To make balance, we use the four-level method to model the speech intonation.

### 2.2. Prediction of pitch levels

#### 2.2.1. *Training*

The prediction of pitch levels now turns to a sequential data labeling problem in machine learning. The methods usually used for sequential supervised learning are sliding-window methods, hidden Markov models, maximum entropy Markov models and conditional random fields (CRF) [12]. The CRF method outperforms others by solving the label bias problem in a principled way because it has a single exponential model for the joint probability of the entire sequence of labels for a given observation, so the weights of different features at different places can be traded off. For all the advantages of CRF, we used it for training our pitch levels on the current database. The features selected for training are listed in Table 1. As shown in Table 1, the position of the syllable in the sentence plays an important role in pitch generation. And pauses caused by punctuation such as commas or full stops will lead to different pitch levels. Moreover, the pitch is closely related to the vowel in the syllable, therefore we consider the Chinese character and lexical tone of the syllable. Also word structure of two or three syllables conveys a semantic meaning in Chinese, so different positions in one word affect the pitch of syllable. For example, some speaker may treat the last syllable of one word with more stress but some may treat them with less stress. When training with CRF model, the features of the adjacency are taken into consideration.

Table 1: *Features selected for CRF training.*

| Index | Feature selected |
|---|---|
| 1 | Position of the syllable in the sentence |
| 2 | Pause index after the syllable |
| 3 | Chinese character of the syllable |
| 4 | Lexical tone of the syllable |
| 5 | Position of the syllable in the word |

#### 2.2.2. *Prediction*

With the CRF model trained in Section 2.2.1, the prediction precision is tested. Table 2 lists the prediction precisions of different numbers of pitch levels on the while corpus. The result shows with the increasing of pitch levels, the CRF model performs worse. But more pitch levels will have a better approximation of pitch contour, so four-level method is used instead of three-level. Table 3 lists the training performance of four-level method, when CRF model(in set) is trained and tested with 2000 sentences, and the CRF model(out set) uses 1900 sentences for training and 100 sentences for testing. For comparison, we train the ME model[18] (in set) and find that CRF model outperforms ME model.

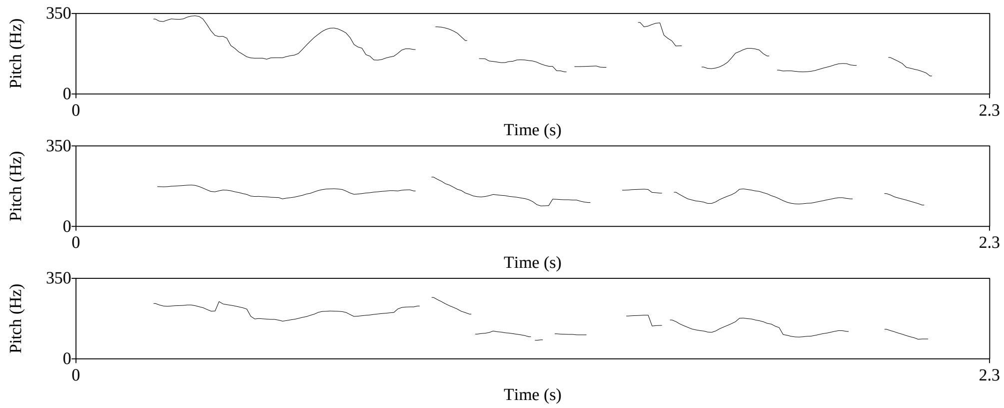


Figure 2: *The pitch contour of a natural speech sentence 'zhong1 guo2 neng2 yuan2 zhan4 lue4 de0 ji1 ben3 nei4 rong2 shi4' (top); the pitch contour generated by HMM (center); the modified pitch contour based on proposed method (bottom).*

Table 2: *Prediction results of different numbers of pitch levels.*

| Number of levels | 3 | 4 | 5 | 6 |
|---|---|---|---|---|
| Precision | 74.52% | 69.23% | 62.66% | 58.70% |

Table 3: *Prediction results of 4 pitch levels.*

| Method | CRF(in set) | CRF(out set) | ME(in set) |
|---|---|---|---|
| Precision | 69.23% | 57.10% | 53.64% |

# 3. HMM-based speech synthesis

[13] [14] has done attempts on speech synthesis by HMM modeling and Tokuda [15] [16] developed the HMM-based synthesis model to a trainable one regardless of language with a good synthesized quality based on parameter speech synthesis which is different from the waveform concatenation method. In the MSD-HMM, the F0 for each frame is considered as the training feature along with the spectral parameters. Moreover, state duration of each phoneme is modeled as a multi-dimensional Gaussian distribution. So spectrum, pitch, and duration are modeled simultaneously in the unified framework of HMM. The advantages of this model are language transparency, speaker dependency, and full-automation.

In our system, the context dependent and tonal phoneme HMMs are trained with 5 states which are left-to-right with no skipping. The contextual information used is as follows:

- Preceding, current, succeeding phoneme
- Tone of preceding, current, succeeding phoneme and position of current phoneme in current word
- Part-of-speech of preceding, current, succeeding word
- Number of syllables in preceding, current, succeeding word and position of current word in the sentence
- Number of syllables and words in the sentence

The speech signal is windowed by a 25ms Blackman window with 5ms frame shift; the spectrum feature vector is consisted of 19 mel-cepstral coefficients [17] including the 0th coefficient, their delta and delta-delta coefficients and the F0 parameter contains log(F0), its delta and delta-delta values.

# 4. Integration of intonation information

## 4.1. Generation method

With a given text, the word segmentation and part-of-speech of words are analyzed. With the feature listed in Table 1, the pitch levels of the syllables are predicted with CRF method. From the statistics in Section 2.1, we know the pitch range of the speaker is between $(F_{mean})_{amax}$ and $(F_{mean})_{amin}$ so the range is equally divided into 4 areas and the mean pitch value of each area is assigned to the corresponding pitch level. In our corpus, the pitch Level 1-4 corresponds to 107.0Hz,147.2Hz,187.3Hz and 227.5Hz respectively.

Regarding pitch level as a bias parameter, the modification of pitch goes as steps:

- With syllable boundaries generated by HMM models, pitch value of the $i_{th}$ frame $F_{level}(i)$ is converted from pitch level of the syllable.
- The pitch values converted from pitch levels are not continuous, so spline smooth method is used.
- With the pitch values generated by HMM models, $F_{HMM}(i)$ for the $i_{th}$ frame, the mean F0 value $F_{mean}$ of each syllable is calculated.
- For each frame, $F_{mean}$ is subtracted from $F_{HMM}(i)$.
- The modified F0 value is $F_{HMM}(i)$-$F_{mean}$+$F_{level}(i)$
- For the unvoiced frames, the F0 value remains 0 as generated by HMM models.

## 4.2. Experiment and Result

1900 sentences of the corpus are used for training CRF-based pitch level model as in Section 2.2 and HMM models as in Section 3. The remaining 100 sentences are used for out-set test.

The bottom plot in Figure 2 shows the pitch contour after pitch modification with intonation information, which leads more pitch variations in the whole sentence compared with the pitch contour generated by HMM method (the center plot in Figure 2). Speech generated by the proposed method is more expressive than the speech produced by HMM model because the intonation information plays an important role in the hearing perception of speech. However the quality of the speech is not improved significantly (with even degradations sometimes.)

Since the speech generated by HMM method is not the same length as the natural speech, we use Dynamic Time Warping (DTW) method to align the two speech and calculate the minimum distance as the RMSE value. Table 4 lists the RMSE of the speech generated by HMM method and of the proposed intonation method compared with the natural speech. The second row of Table 4 is CRF prediction rate for pitch level; the 100% rate means the pitch level of each syllable is predicted correctly and under this condition the proposed model has a better result than HMM method. With the decrease of prediction rate, the proposed method has worse RMSE but it can still make the generated speech more expressive with pitch variations.

Table 4: *RMSE of the pitch value of HMM model and of the proposed integration method with the natural speech*

| | In-set test (1900 sentences) | | Out-set test (100 sentences) |
|---|---|---|---|
| Pitch level prediction rate | 100 % | 69.23 % | 57.10 % |
| HMM method [Hz] | 16.86 | | 20.18 |
| Proposed method [Hz] | 15.28 | 17.63 | 21.10 |

## 5. Discussion

In this paper we treat the intonation and lexical tone in Chinese as different parts and model them separately with the pitch level prediction for intonation and HMM-based method for tones. For each syllable mean F0 value is calculated. With the pitch range of the speaker defined the pitch levels (Level 1-4) are assigned. However, there is no clear evidence that the levels should be 4 but not 3 or 5. Though the modeling is not precise enough, the prediction by the conditional random fields (CRF) method with linguistic contextual information can surmise the tendency of the given text (which is not limited to declaration intonation.)

As for the pitch of lexical tone, we use the MSD-HMM speech synthesis framework, which can depict the fine structure of pitch variation on phoneme scale. To model one phoneme with different tone styles, we also use the contextual features as the labels of the phonemes. Moreover, the duration of phonemes is also modeled in the HMM scheme as a general Gaussian distribution. Consequently the duration information can be used to control the combination of intonation and tone parts.

The speech generated by the proposed method with intonation has better prosodic effects but would not provide precise pitch counter when the pitch level prediction is poor. Compared with Fujisaki model or TA model, this method need less manual labels on corpus and can be applied to different languages just with a database of the specified language. And this method is also flexible to different speakers or speaking styles because with different pitch levels, it cannot be constrained to a certain intonation.

This work is still preliminary. The definition of pitch level is not accurate enough and the pitch generated by HMM scheme cannot show large variation in syllable tones due to the generating regulation. The target model of Yi Xu should be applied to modify the generation regulation in the HMM-based synthesis.

## 6. References


[1] Huang, Xuedong; Acero, Alejandro; Hon, Hsiao-Wuen, 2001. Spoken Language Processing. Prentice Hall: 727-771.

[2] Cruttenden, Alan, 1997. Intonation, 2nd Edition. Cambridge: Cambridge University Press.

[3] Chao, Y.R., 1933. A preliminary study of English intonation and its Chinese equivalents. BIHP Supplement No.1.

[4] Fujisaki, H.; Hirose, K., 1984. Analysis of voice fundamental frequency contours for declarative sentences of Japanese. J.Acoust. Soc.Jpa (E), 5(4), 233-242.

[5] Wang, C.; Fujisaki, H.; Ohno, S.; Kodama, T., 1999. Analysis and Synthesis of the Four Tones in Connected Speech of the Standard Chinese Based on a Command-Response Model. In Proc.of EUROSPEECH99, Budapest; Hungary.

[6] Mixdorff, H., 2000. A Novel Approach to the Fully Automatic Extraction of Fujisaki Model Parameters. In Proc. of ICASSP 2000, vol.3, 1281-1284.

[7] Xu, Yi; Wang, Q. E., 2001. Pitch Targets and Their Realization: Evidence from Mandarin Chinese. Speech Communication. 33 (2001), 319-337.

[8] Tokuda, K.; Kobayashi, T.; Imai, S., 1995. Speech parameter generation from HMM using dynamic features, In Proc. of ICASSP 95, 660-663.

[9] Tokuda, K.; Masuko, T.; Miyazaki, N., 2002. Multi-Space Probability Distribution HMM. IEICE Trans. Inf. &Syst., vol.E85-D (3), 455-464.

[11] Kawahara, H.; Masuda-Katsuse, I.; Cheveign, A., 1999. 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 27 (1999), 187-207.

[12] Lafferty, J., McCallum, A., Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, 2001. In Proc. Int' Conf. Machine Learning 2001, 282-289.

[13] Ljolje, A.; Fallside, F., 1986. Synthesis of Natural Sounding Pitch Contours in Isolated Utterances Using Hidden Markov Models. IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-34, 1074-1080.

[14] Giustiniani, M.; Pierucci, P., 1991. Phonetic ergodic HMM for Speech Synthesis. In Proc. of EUROSPEECH 91, 349-352.

[15] Tokuda, K.; Masuko, T.; Miyazaki, N.; Kobayashi, T., 1999. Hidden Markov Models Based on Multi-Space Pobability Distribution for Pitch Pattern Modeling. In Proc. Of ICASSP 99, 229-232.

[16] http://hts.sp.nitech.ac.jp/

[10] Ling, Z.; Wu, Y.; Wang, Y,; Qin, L,; Wang, R, 2006 USTC System for Blizzard Challenge 2006: An Improved HMM-Based Speech Synthesis Method. In ICSLP Satellite Workshop, Blizzard Challenge, 2006.

[17] Imai, S., 1983. Cepstral Analysis Synthesis on the Mel-Frequency Scale. In Proc. of ICASSP 83, 93-96.

[18] Berger, A., Pietra, S., Pietra, V., 1996. A Maximum Entropy Approach to Natural Language Processing. Association for Computational Linguistics, Vol.22, No.1, 39-71.