A SENTENCE-PITCH-CONTOUR GENERATION METHOD USING VQ/HMM FOR MANDARIN TEXT-TO-SPEECH

Hung-Yan GU and Chung-Chieh YANG

Department of Electrical Engineering, National Taiwan University of Science and Technology, Taipei
root@guhy.ee.ntust.edu.tw, http://guhy.ee.ntust.edu.tw

ABSTRACT

In this paper, a method with sentence-wide optimization consideration is proposed to generate a Mandarin sentence's pitch-contour. The developed model is called the sentence pitch-contour HMM (SPC-HMM) due to its use of VQ (vector quantization) and HMM (hidden Markov model). To construct an SPC-HMM, the pitch-contours of the syllables from each training sentence are normalized on both time and pitch-height first. The method for pitch-height normalization is effective and newly developed here. After normalization, the pitch-contour of each training syllable is vector quantized. Then, the quantization code and lexical tones of adjacent syllables are combined to define the observation symbol sequences for HMM training. In the synthesis phase, when given a sentence and its relevant text-analysis information, the most probable observation sequence is generated by finding the sentence-wide largest probability path with a dynamic-programming based algorithm. We had conducted practical perception tests. It is found that the speech synthesized by using the sentence pitch-contour generated from out method is slightly better than uttered by an ordinary speaker. Besides, the comprehensibility of the synthesized speech is also promoted.

Keywords: Text-to-speech, pitch-contour, hidden Markov model, vector quantization

1. INTRODUCTION

In general, a TTS (text-to-speech) system is made of three main processing components, i.e., text analysis, prosodic parameter generation, and signal waveform synthesis [1]. When a Chinese sentence is to be synthesized, it is first analyzed by the text analysis component to segment it into a sequence of words, to set the boundaries of breath groups, and to determine the corresponding syllable and tone for each of its component characters. Then, the prosodic parameters, pitch-contour, duration, amplitude, and pause, for each syllable of the sentence are decided by the prosodic-parameter generation component. According to the given prosodic parameters, the signal waveform synthesis component then starts to synthesize clear and intelligible speech waveform.

We had studied the problem of signal waveform synthesis before [2,3]. A flexible synthesis method, called TIPW, is proposed which can eliminate the two important drawbacks, chorus and reverberation, found in PSOLA [4]. However, the pitch-contours of the syllables comprising a sentence play the dominant role for the naturalness level of the synthesized speech. Therefore, we decide to study the problem of sentence pitch-contour generation. In the past, many efforts had been made in studying the generation of pitch-contour. For example, The rule-based approach [5], the statistical approach [6], and the recurrent-neural-network approach [7]. Among the different methods, most of them select pitch-contour templates with just local optimization consideration. It is therefore questioned if a pitch-contour generation method can take sentence-wide optimization consideration and is explicitly controllable (not like a black box as an artificial neural network). With this purpose in mind, we found that a HMM (or finite state model) based method is just what we want.

From relevant literature, we know that a syllable at the beginning of a sentence is usually uttered with higher pitch than that at the end, i.e., the phenomenon of declining. To model this phenomenon, three prosodic states representing sentence-initial, sentence-middle, and sentence-final, are adopted. However, we do not know how to segment a sentence's syllables into these states explicitly. Therefore, we imagine these prosodic states are hidden and will represent them by the hidden states of a straight left-to-right HMM. Besides the influence of prosodic states, the lexical tones of a syllable and its adjacent syllables also have strong influences. Therefore, we will combine adjacent syllables' lexical tones and pitch-contour VQ code to form observations for such a HMM. We do not consider other minor factor (e.g., syllable-type and position in a word) because we have just a limited number of training sentences. This should not confine the extensibility of such model framework. Such a HMM based model is called sentence pitch-contour HMM (SPC-HMM) because the most probable observation sequence is generated, in the synthesis phase, by finding the sentence-wide largest probability path with a dynamic programming based algorithm. For a generated observation sequence, the corresponding sequence of syllable pitch-contour VQ code can be simply decoded as the inverse of observation symbol encoding.

Although HMM has also been adopted by other researchers to generate pitch-contours [8,9], however, there are several obvious differences. In their studies, the observations of HMM represent the consecutive pitch-periods' lengths (so small time-scale), a different HMM is used to model the pitch-contours of a different syllable, and more importantly sentence-wide phenomenon such as declination is not considered. In contrast to their studies, only one HMM is constructed here for all sentences, and the observations of the HMM are the combination of pitch-contour VQ code and lexical tones, i.e., the time-scale is as large as a syllable. More importantly, SPC-HMM has sentence-wide phenomenon covered.

In the training phase of SPC-HMM, the main processing flow is as shown in Fig. 1 whereas in the synthesis phase, the main processing flow is as shown in Fig. 2. In Section 2, the functions of the blocks in Fig. 1 will be described. Then, the functions of the blocks in Fig. 2 will be explained in Section 3. In Section 4, SPC-HMM is evaluated by

perception tests.

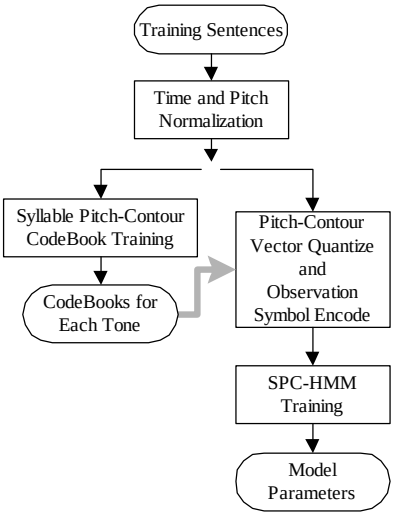

Fig. 1 Main flowchart for the training phase

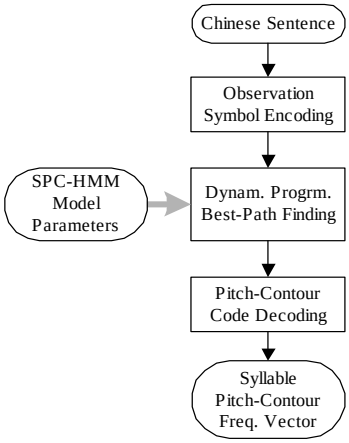

Fig. 2 Main flowchart for the synthesis phase

2. TRAINING PHASE

2.1 Time and Pitch Normalization

We decide to represent a syllable pitch-contour as a vector of 16 frequency heights (in Hz) computed at 16 normalized (i.e., equal spacing) time points over a syllable's voiced part. If a time point is located between two adjacent pitch periods' center points, its corresponding frequency height is then computed as the inverse of the weighting sum of the two pitch periods' lengths.

Pitch height normalization is needed because the training sentences are usually recorded in many days with different emotions, and have large variations among the sentences' average pitch heights. If normalization is not done, abnormal pitch-contour transition between two adjacent syllables will be often heard in the synthesized speech. Here, an effective normalization method is proposed, with which only one utterance is required for each training sentence. The procedure for this method is:

(a) For the *i*'th training sentence, compute its *j*'th syllable's average pitch-height E_j in logarithmic scale. That is,

$$E_j = \frac{1}{16}\sum_{k=0}^{15} p_{jk}\,, \quad p_{jk} = \mathbf{log}(f_{jk})\,, \tag{1}$$

where f_{jk} is the frequency height at the normalized time point *k*. Then, compute this sentence's average pitch-height S_i as

$$S_i = \frac{1}{n}\sum_{j=1}^{n} E_j\,, \tag{2}$$

where *n* represents the number of syllables in this training sentence.

(b) Compute the grant average pitch-height, ***Sa*****,** across all training sentences. That is,

$$Sa = \frac{1}{St}\sum_{i=1}^{St} S_i \tag{3}$$

where *St* represents the number of training sentences.

(c) Compute the pitch-height adjusting value, δ_i, for the *i*'th training sentence as

$$\delta_i = S_i - Sa \tag{4}$$

(d) According to δ_i, normalize the pitch contour of the *j*'th syllable of the *i*'th training sentence as

$$\bar{p}_{jk} = p_{jk} - \delta_i\,, \quad k=0,1,\ldots,15, \quad j=1,2,\ldots,n \tag{5}$$

Although the method explained above seems simple, it can indeed eliminate most abnormal pitch-contour transitions between syllables. To further eliminate unacceptable pitch-contour transitions, we have studied another sentence pitch-height normalization method. This method is applied to the resultant pitch-contours obtained from the prior normalization method. The procedure for this method is:

(a) Uniformly divide each training sentence into three segments. Collect the syllables, from all training sentences, which are divided to the first segment. Then, compute the mean pitch-height, $M_{0,k}$, of these syllables that are pronounced in the *k*'th lexical tone. Similarly, the mean pitch-height, $M_{1,k}$ and $M_{2,k}$, for those syllables divided to the second and third segments can be computed also.

(b) For the *i*'th training sentence, compute its *j*'th syllable's pitch-height difference d_j. Then, compute the mean difference $\bar{d}$ for this sentence. That is,

$$d_j = E_j - M_{l,k}\,, \quad l = \left\lfloor \frac{(j-1)}{n}\cdot 3 \right\rfloor, \quad j=1,2,\ldots,n \tag{6}$$

$$\bar{d} = (d_1 + d_2 + \ldots + d_n)/n \tag{7}$$

where E_j is the renewed pitch-height from the prior normalization method, *l* is the segment number that the *j*'th syllable is divided to, *k* is the tone number that the *j*'th syllable is pronounced, and *n* is the number of syllables in the *i*'th training sentence.

(c) According to the mean difference $\bar{d}$, normalize the pitch-contour of the *j*'th syllable as

$$\bar{p}_{jk} = p_{jk} - \bar{d}\,, \quad k=0,1,\ldots,15, \quad j=1,2,\ldots,n, \tag{8}$$

where p_{jk}, *k*=0,1, …, 15, is the *j*'th syllable's pitch-contour obtained from the prior normalization method.

In pitch-contour codebook training, three conditions, no pitch height normalization, normalization using the first method, and normalization using the two methods above, are tested. The average VQ errors obtained are, 0.0398, 0.0330, and 0.0308 (about 3.7Hz at 120Hz) respectively,

i.e., VQ error will become smaller as more normalization methods are used.

2.2 Vector Quantization

After time and pitch height normalization, the training syllables' pitch-contour vectors are classified according to their lexical tones. Then, for each lexical tone, we use GLA (Generalized Lloyd Algorithm) to perform VQ codebook training [10]. Apparently, the average quantization error will become smaller when the codebook size become larger. However, this is not always good because larger codebook size will result in larger observation space for SPC-HMM, and larger observation space means coarser HMM parameter estimation. That is, a tradeoff should be made. Here, we set each lexical tone's codebook size to be 8 according to preliminary experiment results.

2.3 Observation Symbol Encoding

The lexical tones of three adjacent syllables are combined with the pitch-contour VQ code of the middle syllable to define its corresponding discrete observation. That is, an observation at time t is defined as

$$O_t \equiv X_{t-1} \times X_t \times X_{t+1} \times V_t, \quad 0 \le X_t \le 4, \ 0 \le V_t \le 7 \\ = 200X_{t-1} + 40X_t + 8X_{t+1} + V_t \tag{9}$$

where Xt represents the lexical tone number of the t'th syllable in a training sentence and Vt represents the VQ code of the t'th syllable's pitch-contour. When t=1, $Xt\text{-}1$ is undefined and is therefore removed, and Ot is defined in the range, $1{,}000 \le Xt \times Xt{+}1 \times Vt < 1{,}200$. Similarly, when t represents the last syllable, Ot is defined in the range, $1{,}200 \le Xt\text{-}1 \times Xt \times Vt < 1{,}400$.

Besides, consider the condition that some three-lexical-tone combinations seen in the synthesis phase may not be seen in the training phase due to insufficient training sentences. We resolve this difficulty by building two simplified SPC-HMM, in which observations are defined as fewer factors' combinations. That is,

$$O_t \equiv X_t \times X_{t+1} \times V_t \tag{10}$$

$$O_t \equiv X_{t-1} \times X_t \times V_t \tag{11}$$

for the first level and the second level downgrades and have values in the two ranges, 1,400 to 1,599 and 1,600 to 1,799 respectively. Then, when an observation is not seen in the training phase, its occurrence probability is looked up from the corresponding observation in the downgraded SPC-HMM but divided by a constant (e.g., 100,000) to avoid model biasing.

2.4 SPC-HMM Training

The parameters, a_{ij} and $b_j(k)$, of the original and the two downgraded SPC-HMM can be trained independently since no downgrading occurs in the training phase. The segmental K-means algorithm is used here [10]. About the insufficiency of training sentences (375 sentences of 2,925 syllables), we have adopted a sharing method. That is, when an observation is seen, 0.0001 of its occurrence is shared to the nearest observation that has same lexical tone combination but differs in pitch-contour VQ code.

In original HMM, observations generated from a same state are assumed to be mutually independent. However, in a Mandarin sentence, adjacent syllables' pitch-contours have at least some degree of dependence. Therefore, we add a new parameter, $c_j(k)$, to record the average pitch-height difference between the former two syllables' pitch-contours, whose lexical tone are combined to obtain observation k in state j. For the t'th syllable of a sentence, the pitch-height difference, Wt, is defined as

$$W_t = WF_t - WB_{t-1} \tag{12}$$

where WF_t represent the front-pitch-height for the t'th syllable and WB_{t-1} represent the back-pitch-height for the $(t\text{-}1)$'th syllable. That is,

$$WF_t = \frac{1}{8}\sum_{j=0}^{7} p_{t,j} \quad , \quad WB_t = \frac{1}{8}\sum_{j=8}^{15} p_{t,j} \tag{13}$$

where $p_{t,j}$ represents the logarithmic frequency at the j'th normalized time point of the t'th syllable. When this $c_j(k)$ parameter is included, with equal weight to $b_j(k)$, into SPC-HMM, the average root-mean-square prediction error for a syllable's pitch-contour can be improved about 5%.

3. PITCH-CONTOUR GENERATION

In the synthesis phase, a given Chinese sentence is analyzed by text-analysis component first to derive its pronunciation syllable sequence. Then, every three adjacent syllables' lexical tones are combined to select the eight (eight codewords in each tone's pitch-contour codebook) possible observations for every time position. So, in addition to the time and state axes, the third axis, i.e., index to the eight possible observations, should also be considered. Here, we have extended the commonly used two-dimensional dynamic programming (DP) algorithm (or called Viterbi algorithm) to solve this three-dimensional DP problem. Therefore, a probabilistically best observation sequence can be found for a given lexical tone sequence. Then, the pitch-contour VQ code sequence is decoded from the best observation sequence, and each VQ code can be used to retrieve its correspondent time-and-pitch normalized frequency vector.

Note that structural prosodic information such as breath-breaks and word-boundaries are not used in the training and synthesis phases. The sentence pitch contour generated under this condition is called Mode-A generation, and may not be satisfactory for a long sentence. Therefore, we had studied another SPC-HMM based pitch contour generation method, called Mode-B generation method. In this method, the breath-break and word-boundary information from text-analysis component is used to set the state transition sequence in SPC-HMM. Although the state transition sequence is now fixed, the determination of pitch-contour VQ code for each syllable is still a two dimensional DP problem. For example, suppose there is a breath-break between the third and forth syllables of a sentence consisting of seven syllables. Then, the state transition sequence is set to 0,1,2, 1, 1, 2, 2 (the transition from state 2 to 1 is allowed here in the synthesis phase). In the first breath group, the syllables are uniformly divided to the three states while the syllables in the second and latter groups are uniformly divided to the states 1 and 2. As to the word boundary information, it is used to modify the state setting such that the two syllables of a two-character word are placed at a same state. Also, it must be satisfied that the last syllable of the first group must be placed at state 1 or 2 while the last syllables of the other groups must be placed at state 2 to form a wave-like state transition. With this integration of structural prosodic information, the

naturalness level of the generated sentence pitch contour has been improved a lot.

4. PERCEPTION TEST

Eighteen persons were invited to evaluate the SPC-HMM based sentence pitch-contour generation methods. In the evaluation of comprehensibility, 15 different sentences are divided into 3 sets with roughly equal difficulty. Each set is assigned to one of the 3 test conditions, i.e., sentence pitch-contour generation with simple rules (i.e., previous version), with SPC-HMM based Mode-A method, and with SPC-HMM based Mode-B method. Then, for each person, the three test conditions are randomly permutated and the sentences assigned to each condition are synthesized. After listening to each synthesized sentence, the invited person is requested to write down the Chinese sentence he heard. The comprehensibility is defined here as the average ratio of correctly written characters over total characters.

In the evaluation of prosody-preference score, the speech uttered by the second author is defined as having 8 points while the perfect prosody has 10 points. For each person, the speech (reading an article) uttered by the second author is played first, then the speech synthesized under the 3 test conditions are played respectively. The invited person is requested to write down his prosody-preference score for each condition. The evaluation results are as shown in Table 1. From this table, it can be seen that the comprehensibility has been promoted from 81.2% for the previous version to more than 95% when using SPC-HMM based generation methods. Besides, it is surprising that the speech synthesized by using the SPC-HMM Mode-B generation method is evaluated to have preference score of 8.2 points, which is slightly higher than the speech uttered by the second author. Also, this score, 8.2, is apparently higher than the scores for the speech synthesized by using simple rules and the SPC-HMM based Mode-A method.

Table 1 Perception evaluation results.

	Simple rules	Mode-A	Mode-B
Comprehensibility	81.2%	95.1%	96.5%
Preference-Score	5.1	7.0	8.2

For those interested in evaluating the SPC-HMM based sentence pitch-contour generation method, we had set up a WWW home page at http://guhy.ee.ntust.edu.tw/gutts, on which an on-line inputted Big-5 Chinese sentence is synthesized immediately and its speech signal is then sent back for evaluation.

5. CONCLUSION

In this paper, we had studied and proposed a sentence pitch-contour (SPC) generation model using HMM to model implicit prosodic states and VQ to classify each lexical tone's syllable pitch-contours into 8 classes. This model is called SPC-HMM because in the generation of sentence pitch-contour, sentence-wide optimization consideration is taken into account, i.e., find the most probably syllable pitch-contour sequence by dynamic programming. In addition, we had proposed an effective pitch-height normalization method. By this normalization method, abnormal pitch-contour transitions between syllables can be nearly removed from the synthesized speech.

Although the structural prosodic information, breath breaks and word boundaries, are not used in training SPC-HMM, these information can still be utilized in the synthesis phase to set the state transition sequence, i.e. SPC-HMM based Mode-B generation method. The perception evaluations show that Mode-B generation method can indeed obtain prosodic-preference score slightly better than uttered by an ordinary person. It is a good idea to integrate the structural prosodic information directly into the model, SPC-HMM, but how to implement this idea needs to be studied.

REFERENCE

[1] Shih, C. and R. Sproat, "Issues in Text-to-Speech Conversion for Mandarin", Computational Linguistics & Chinese Language Processing, Vol. 1, No. 1, pp. 37-86, 1996.

[2] Gu, H. Y. and W. L. Shiu, "A Mandarin-syllable Signal Synthesis Method with Increased Flexibility in Duration, Tone and Timbre Control", Proc. Natl. Sci. Counc. ROC(A), Vol. 22, No.3, pp. 385-395, 1998.

[3] Gu, H. Y., "Notes for the Syllable-Signal Synthesis Method: TIPW", ISCSLP (Singapore), SS-B3, 1998.

[4] Modulines, E. and F. Charpentier, "Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones", Speech Communication, pp. 453-467, 1990.

[5] Lee, L. S., C. Y. Tseng and C. J. Hsieh, "Improved Tone Concatenation Rules in a Formant-based Chinese Text-to-Speech System", IEEE trans. Speech and Audio Processing, Vol. 1, pp. 287-294, 1993.

[6] Chen, S. H. and S. M. Lee, "A Statistical Model based Fundamental Frequency Synthesizer for Mandarin Speech", J. Acoust. Soc. Am., Vol. 92, No. 1, pp. 114-120, 1992.

[7] Chen, S. H., S. H. Hwang and Y. R. Wang, "An RNN-based Prosodic Information Synthesizer for Mandarin Text-to-Speech", IEEE trans. Speech and Audio Processing, Vol. 6, No.3, pp. 226-239, 1998.

[8] Ljolej, A. and F. Fallside, "Synthesis of Natural Sounding Pitch Contours in Isolated Utterances using Hidden Markov Models", IEEE trans. Acoust., Speech and Signal Processing, Vol. 34, No.5, pp. 1074-1079, Oct. 1986.

[9] Fukada, T., Y. Komori, T. Aso, and Y. Ohora, "A Study on Pitch Pattern Generation using HMM-based Statistical Information", Int. Conf. on Spoken Language Processing (Japan), pp. 723-726, 1994.

[10] Rabiner, L. and B. H. Juang, Fundamentals of Speech Recognition, Prentice-Hall, 1993

ACKNOWLEDGEMENT

This work was sponsored by national science council under the contract number NSC-89-2213-E-011-058.