Journal Of Harmonized Research (JOHR)

Journal Of Harmonized Research in Engineering
2(4), 2014, 367-371

ISSN 2347 – 7393

Original Research Article

A SENTENCE-PITCH-CONTOUR MODEL FOR BODO LANGUAGE USING HIDDEN MARKOV MODEL (HMM) AND VECTOR QUANTIZATION (VQ)

Laba kr. Thakuria

Department of Instrumentation and USIC
Gauhati University, India

Abstract:- Through this paper weintroduce a concept to implement a sentence's pitch-contour model with sentence-wide optimization. This is also called the sentence pitch-contour using HMM(Hidden Markov Model) & VQ (vector quantization) . Here each training sentence are normalized for the pitch-contours of the syllables. This is basically effective for pitch height normalization and after normalization, the pitch-contour of each training syllable is then vector quantized(VQ). The quantization code and lexical tones of adjacent syllables arethen combined to define for HMM training. Using a dynamic-programming in the synthesis phase, the probable observation sequence is produced by finding the sentence wide largest probability path.The pitch-contours of the syllables comprising a sentence which play the main dominant role for the naturalness of the synthesized speech.

Keywords: Bodo, HMM, SPC, VQ, optimization, pitch, contour, dynamic programming.

I. Introduction

A TTS (Text-to-Speech) system is made of three main processing components, i.e. (i) text analysis,(ii) prosodic parameter generation, and (iii) signal waveform synthesis. When a Bodo sentence is synthesized, then it is first analyzed by the text analysis component to segment it into a sequence of words to determine the corresponding syllable and tone for each of its component characters. The prosodic parameters, pitch-contour, duration, amplitude, and pause, for each syllable of the sentence are decided by the prosodic-parameter generation component. Due to the given prosodic parameters, the signal wave form synthesis component then starts to synthesize .A synthesis method, called TIPW, is proposed through which we can eliminate the two important drawbacks. The pitch-contours of the syllables comprising a sentence, play the main role for the naturalness level of the synthesized speech. Earlier many experiments has been done in studying the generation of pitch-contour. For example, the rule-based approach, the statistical approach, and the recurrent-neural network approach. From the relevant literature, we come to know that a syllable at the beginning of a sentence is usually

For Correspondence:
thakurialabaATgmail.com
Received on: October 2014
Accepted after revision: November 2014
Downloaded from: www.johronline.com

uttered with higher pitch than that at the end. Tomodel this, three prosodic states representing sentence-initial, sentence-middle, and sentence-final, are adopted. Besides the effect of prosodic states, the lexicaltones of a syllable and its adjacent syllables also have strong effect. Therefore, we combine adjacent syllables, lexical tones and pitch-contour VQ code to form observations for such a HMM.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. In case of a generated observation sequence, the corresponding sequence of syllable pitch-contour VQ code can be decoded as the inverse of observation symbol encoding. 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 inSection 3. In Section 4, SPC-HMM is evaluated by perception tests.

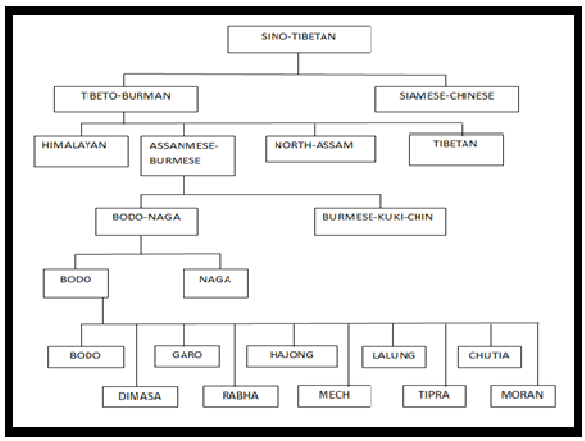
II. Bodo Lanaguage

The Bodo language is originated to a Sino-Tibetan language which is closely related to the Dimasa language. The Bodo speaking areas of Assam is from Dhubri to the west of Sadiya. The population of Boro speakers according is increasing now and census reports of Bodo tribe, however, comprises only the Bodos. The dialects spoken by Bodo in this area could be broadly sub-divided into three main groups:

1. The Western Boro , {(Sønabari) WBD}:
2. The Eastern Boro , {(Sanzari) EBD} and
3. The Southern Boro , {(Hazari) SBD}.

The Western Boro dialects are spoken in the districts of Kokrajhar and Bongaigaonand Eastern dialects are basically in the districts of Barpeta, Nalbari and Kamrup and some parts of Darrang, Udalguri. The Western Boro dialect has considered as the Standard Dialect and has developed a written form as well. The difference between the two dialect groups are mainly on the phonological and lexical part. The University Grants Commission (UGC) has also included Bodo as subject in NET examination. Bodo language is written using the Devanagari script. Some researchers have suggested that the language used to use a now-lost script called Deodhai. There is a difference in using the letters in Bodo than the Devanagari.

The family structure of Bodolanguage is as given below



Firgure -1 : Sino-Tibetan structure of Bodo

The Bodo phonemes consists of 6(six) vowels and 16(sixteen) consonants. Two semi vowels out of 16 consonants. They are as shown below-

a. Vowels : **अ**, **आ**, **इ**, **उ**, **ए**, **औ**

b. Consonants : **ख**, **ग**, **ङ**, **ज**, **थ**, **द**, **न**, **फ**, **ब**, **म**, **र**, **ल**, **स**, **ह**

c. Semi Vowels:**य**, **व**

III. Training Phase

A) Time and Pitch Normalization

We have decided to represent a syllable pitch-contour as a vector of 16 frequency heights (Hz) computed at 16 normalized time points over a syllable's voiced part. If a time point is located between the two adjacent pitch periods center points, its corresponding frequency height is then evaluated as the inverse of the weighting sum of the two pitch periods' lengths. The pitch height normalization is necessary because the training sentences are usually recorded in many days with different emotions, and have large variations among the sentences average pitch heights. Here, we an effective normalization method is proposed, with which only one utterance is required for each training sentence. The

procedure of this method is :(a) For the *i*'th training sentence, compute its *j*'th syllable's average pitch-height *Ej*in logarithmic scale. i.e,

15

$$Ej = 1/16 \sum_{k=0} P_{jk}, P_{jk} = \log(f_{jk}) , \quad (1)$$

where*fjk*is the frequency height at the normalized timepoint *k*. Then, compute this sentence's average pitchheightS*i* as

n

$$Si = 1/n \sum_{j=1} Ej \quad (2)$$

where*n* represents the number of syllables in training sentence.(b) Compute the grant average pitch-height, *Sa*, across alltraining sentences. i.e,

St

$$Sa = 1/St \sum_{i=1} Si \quad (3)$$

where *St* represents the number of training sentences.(c) Compute the pitch-height adjusting value, δi , for the*i*'th training sentence as

$$\delta i = Si - Sa \quad (4)$$

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

$$pjk = p\,jk - \delta i , k=0,1, ..., 15, j=1,2, ..., n \quad (5)$$

It is simple and it can indeed eliminate most abnormal pitch-contour transitions between syllables. The 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. Collecting the syllables, from all training sentences, which are divided to the first segment and Then, compute the mean pitch-height, *M0,k,* of these syllables that are pronounced in the *k*'th lexical tone. In the same way, the mean pitch-height, *M1,k* and *M2,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'spitch-height difference *dj and t*hen, compute the mean difference *d* for this sentence. i.e,

$$dj = Ej - M_{l,k}, \quad (6)$$

$$l=[(j-1)/n.3], j=1,2,\ldots n \quad -$$

$$d = (d1 + d2 + ... + dn) / n \quad (7)$$

where *Ej* 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)Due to the mean difference *d* , normalize the pitch-contour of the *j*'th syllable as

$$pjk = p\,jk - d , k = 0,1,...,15, j =1,2,.., n, \quad (8)$$

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

B) Vector Quantization

The training syllables' pitch-contour vectors are classified according to their lexical tones. For each lexical tone, we have used Generalized Lloyd Algorithm to perform VQ code book training. It is not always better because larger code book size will result in larger observation space for SPC-HMM and larger observation space means coarser HMM parameter estimation.

C) Observation Symbol Encoding

After that 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 i.e. an observation at time *t* is defined as

$$Ot = X_{t-1} \times X_t \times X_{t+1} \times V_t$$

$$= 200X_{t-1} + 40X_t + 8X_{t-1} + Vt \quad (9)$$

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. We 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 computed this difficulty by building two simplified SPCHMM, in which observations are defined as fewer factors' combinations. i,e.

$$Ot \equiv XtxXt+1x\ Vt \quad (10)$$

$$Ot \equiv Xt-1\ xXtx.Vt \quad (11)$$

In this case for the first level and the second level downgrades and we have values in the two ranges, 1,500 to 1,690 and 1,700 to 1,890 respectively.

D) SPC-HMM Training

The parameters, *aij* and *bj*(*k*), of the original and the two downgraded SPC-HMM can be trained

independently. The segmental K-means algorithm is used. The insufficiency of training sentences (400 sentences of 3,945syllables), we have adopted a sharing method. i.e, 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. we add a new parameter, $cj(k)$, to record the average pitch-height difference between the former two syllables' pitch-contours, whose lexical tone are combined to obtain observation k instate j. The t'th syllable of a sentence, the pitch-height difference, Wt, is defined as $Wt = WFt - WBt\text{-}1$ (12) , where WFt represent the front-pitch-height for the t'th syllable and $WBt\text{-}1$ represent the back-pitch-height for the$(t\text{-}1)$'th syllable.

IV. Pitch-Contour Generation

We have extended the commonly used 2D dynamic programming algorithm or called Viterbi algorithm to solve this 3D dynamic programming problem. So, the observation sequence can be obtained for a given lexical tone sequence. The pitch-contour VQ code sequence is decoded from the best observation sequence, and each VQ code is used to retrieve its correspondent time-and-pitchnormalized frequency vector. Under this condition the sentence pitch contour is called Mode-A generation. Therefore, we have studied another SPC-HMM based pitch contour generation method, which is called Mode-B generation method. Through this method, the breath-break and word-boundary information from text analysis component is used to set the state transition sequence in SPC-HMM.

In the first breath group, the syllables are uniformly divided into three states while the syllables in the second and latter groups are uniformly divided into states 1 and 2.

V. Perception Test

20 speakersare invited to evaluate the SPC-HMM based sentence pitch-contour generation methods. In this evaluation of comprehensibility, 12 different sentences are divided into 3 sets with equal difficulty. Every set is assigned to one of the 3 test conditions, i.e., sentence pitch contour generation with simple rules with SPC-HMM based Mode-A method, and with SPCHMM based Mode-B method. Then, for each speaker, the three test conditions are randomly permutated and the sentences assigned to each condition are synthesized. After listening to each synthesized sentence, the invited speaker is requested to write down the Bodo sentence that he or she heard. The comprehensibility is defined as the average ratio of correctly written characters over the total characters. In the evaluation of prosody-preference score, the speechuttered by the second author is defined as having 7 points while the perfect prosody has 10 points. For each person, the speech uttered by the second authoris played first, then the speech synthesized under the 3 test conditions are played respectively. The invited speaker 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 85.3% for the previous version to more than 94% when using SPC-HMM based generation methods. It is surprising that the speech synthesized by using the SPC-HMM Mode-B generation method is evaluated to have preference score of 7.8 points, which is slightly higher than the speech uttered by the second author. Also, this score, 7.8 is apparently higher than the scores for the speech synthesized by using simple rules and the SPC-HMM based Mode-A method.

VI. Conclusion

This paper, we have studied and proposed a sentence pitch-contour (SPC) generation model using HMM tomodel implicit prosodic states and VQ to classify each lexical tone's syllable pitch-contours into 8 classes. This proposed 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 have proposed an effective pitch-height normalization method. With 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.

VII. Acknowledgment

This work is done under the department of Instrumentation & USIC, Gauhati University. I am very much thankful to Prof. P. H. Talukdar, supervisor speech research centre, Gauhati University for his constant cooperation and guidance for the completion of the paper.

VIII. 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 SynthesisMethod: TIPW”, ISCSLP (Singapore), SS-B3, 1998.

[4] Modulines, E. and F. Charpentier, “Pitch-SynchronousWaveform Processing Techniques for Text-to-SpeechSynthesis Using Diphones”, Speech Communication, pp. 453-467, 1990.

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

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

[7] Chen, S. H., S. H. Hwang and Y. R. Wang, "An RNNbasedProsodic Information Synthesizer for MandarinText-to-Speech", IEEE trans. Speech and AudioProcessing, Vol. 6, No.3, pp. 226-239, 1998.

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

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

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