

MODELLING AND DECISION TREE BASED PREDICTION OF PITCH CONTOUR IN IBM MANDARIN SPEECH SYNTHESIS SYSTEM

Xiaochuan NIU, Liqin SHEN, Weibin ZHU and Qin SHI

IBM China Research Laboratory, Beijing
niuxc, shenlq, zhuweib, shiqin@cn.ibm.com

ABSTRACT

In this paper, a method of pitch contour modelling based on the hidden Markov model (HMM) states of an acoustic unit is presented. A pair of vectors is computed from the alignment of the speech data with the acoustic unit's HMM states. The pitch contour feature of the acoustic unit is represented by the vector pair so that the variants of the acoustic unit's pitch contour can be measured and compared. Using this model, pitch contour decision trees are constructed for phones in Mandarin from a single speaker's continuous reading speech database. The trees are used in the Mandarin speech synthesis system, which is trained over the same database, to predict the pitch contour of a certain phone according to its phone context. The naturalness of the synthesized Mandarin speech is highly improved.

1. INTRODUCTION

The prosodic modelling, especially the modelling of pitch contour, is relatively more important in a Mandarin speech synthesis system than in other systems of non-tonal languages, because the tones of Mandarin syllables not only compose the intonation pattern of sentences but also determine the meaning of words. Although there are only five lexical tone patterns (i.e. 1st, 2nd, 3rd, 4th and neutral tone) applied on a certain syllable in Mandarin, the pitch contours appearing in the continuous speech are very sensitive to the context of the speech. Many tone sandhi rules of Mandarin have been summarised by the phoneticians ([1]). These rules should be converted into quantitative presentation when they are used to guide the variations of the pitch contours associated with the tone patterns. A quantitative pitch contours generation model has been presented by Fujisaki ([2]). However, the complexity of the variations of the pitch contour patterns in Mandarin is still a key problem to be solved.

This paper introduces a method to model the pitch contour of a certain acoustic unit. The pitch contours of the phones in Mandarin are modelled in the proposed means and the decision tree approach ([3]) is used to

model the context-dependent phone pitch contour. When the IBM trainable speech synthesis system ([4]) for Mandarin is built, the prototypes of these context-dependent phone pitch contours are used to generate the target pitch contours of the phones to be synthesised. Through this pitch contour prediction approach, the synthesised Mandarin speech achieves high naturalness.

The structure of this paper is as follows. In section 2, the construction of the basic Mandarin speech synthesis system is briefed. Then the method of pitch contour modelling is described in section 3. In section 4 is the description of the modelling of decision tree based context-dependent phone pitch contour and its usage for pitch contour prediction in the synthesis system. The experimental results are presented and some discussion is made in section 5. Finally, the future work is briefly discussed in section 6.

2. BASIC SYNTHESIS SYSTEM

The IBM trainable speech synthesis system is potentially language independent. Since the details of constructing an English system have been described in [4, 5, 6], only the language specific parts of the Mandarin system and what are necessary for understanding the following sections are described here.

The training data used to build the system is approximately one hour's continuous reading speech data. It consists of the waveform signals in one channel and the corresponding laryngographic signals in the other channel. The laryngographic signals are processed to determine the pitch synchronous points of the waveform signals. The waveform signals are then divided into the unvoiced and voiced sections according to the pitch synchronous points. The unvoiced sections are segmented into frames in a uniformed frame shift rate (6ms frames at a 3ms frame rate), while the voiced sections are segmented into pitch synchronous frames.

The Mel frequency cepstral features of each of the frames are extracted. After that, a set of left-to-right HMMs of the Mandarin phones are created and trained. The phone set is the same as what is defined in the IBM speech recognition system for Mandarin ([7]), but the HMMs of phones are defined differently. The Initials in the phone set are represented by 3-state models, while the Finals are represented by 6-state models in order to depict more complicate variations of the pitch contour. The Mel frequency cepstral features vector of each frame is aligned with the appropriate phone state using the Viterbi algorithm so that each phone state has an inventory of waveform segments to be selected when synthesising.

The acoustic (HMM) decision trees of the phone states are built by asking questions about the immediate left and right phone contexts. Each leaf of the acoustic trees is an index to a group of waveform segments over which the dynamic programming (d. p.) search for optimal concatenate segments is to be performed. As presented in [4], the core mechanism to ensure the quality of the synthesised speech is the d. p. search algorithm and the associated cost functions including continuity cost, duration cost, pitch cost and energy cost. The required parameters of the acoustic unit, the pitch contour and the energy should be fed to restrain the d. p. processing to select appropriate segments. To predict the energy, phone energy decision trees are built according to the energy of each frame concerning two phones of the phone context in each direction.

As to the pitch contour, which is the combination of duration and pitch, the prediction is based on the duration decision trees and pitch decision trees, whose construction depends on the phone pitch contour model we present.

3. PITCH CONTOUR MODELLING

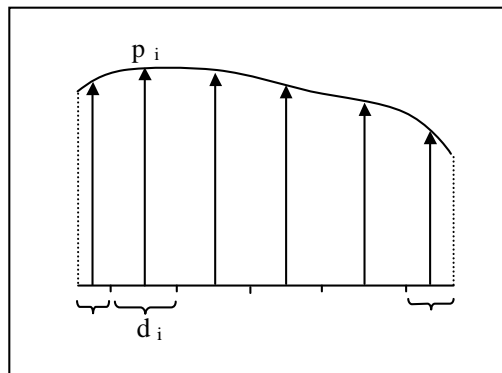
Given the left-to-right N-state HMM of an acoustic unit, when a piece of speech data is aligned with it, two N-dimension vectors can be computed. They are

$$V_D = [d_1, d_2, \dots, d_N]^T \text{ and}$$

$$V_P = [p_1, p_2, \dots, p_N]^T .$$

d_i denotes the duration staying in the i-th state of the HMM, while P_i is a value derived from the pitch points that are located in the corresponding d_i duration of time. Now the vector pair $\{V_D, V_P\}$ is a representation of the given acoustic unit's pitch contour and it can be regarded as a statistical variable dependent on the certain acoustic unit. Simply, it is assumed to have a Gaussian distribution.

As an example, the acoustic unit could be a certain phone of Mandarin. To compute the vector pair of a phone's pitch contour, the pitch mark points of the training speech data have to be firstly extracted from the associated laryngographic signals, so the time-point and the pitch value of each pitch sample are also obtained. Combining the pitch samples with the Viterbi alignment of each sentence, the pitch contour vector pair of each phone in the sentence can be computed as follows. The alignment determines the start time and the end time of each state of a phone, so the d_i of this phone is determined. As to P_i , an interpolated value in the middle of the start time and the end time of the i-th state is used when there are pitch samples between them, otherwise it is set to a huge negative to represent unvoiced feneme (Feneme is a term used to describe an individual HMM model position). Here the vector V_P can be regarded as the result of non-uniformed step sampling of a phone's pitch contour. It can also independently be considered the representation of the phone's pitch pattern after duration normalisation.



F 1. Computation of Pitch Contour Vector Pair

There could be other methods to compute the vector pair of pitch contour defined above, but the significations are the same. The intention of the vector pair modelling for pitch contour is to provide a measurement to compare the different pitch contours of a certain acoustic unit. Using this model, an approach of pitch contour prediction is developed.

4. PITCH CONTOUR PREDICTION

With the aid of the phone pitch contour model, the similarity between the variants of a certain phone's pitch contour can be measured. It is assumed that the variations are dependent on the contexts in which the phone occurs. The decision tree algorithm described in [3] is an appropriate approach to cluster the context dependent variations of a feature vector; the splitting of a tree's nodes implies the context rules of the variations.

We attempt to use this approach to model the context dependent phone pitch contour quantitatively and automatically. Since this approach is data-driven, it will be more efficient and flexible than to summarize any sandhi rules and convert them into quantitative pitch contour presentations.

4.1. Phone Pitch Contour Decision Trees

To build the pitch contour decision tree of a certain phone, all of the pitch contour vector pairs of this phone with its context information should be computed and gathered from the training data.

If the vector pair is treated as one huge vector, a binary tree can be constructed by asking the phone context questions to split nodes from the root to the leaves. The vectors belonging to each node are used to estimate a Gaussian distribution. The optimal phone context question for each splitting is selected according to the log-likelihood gain of the Gaussian. The splitting will keep on until the log-likelihood gain is less than some threshold or the number of the vectors associated to the node is less than a preset value.

If we assume that V_D and V_P in the vector pair are independent, two separate trees of duration and pitch can be constructed for each phone respectively. The method of building the trees is the same as above. This means of tree construction is implemented in our current system. Two phones of the phone context in each direction are taken into the consideration when the trees are built.

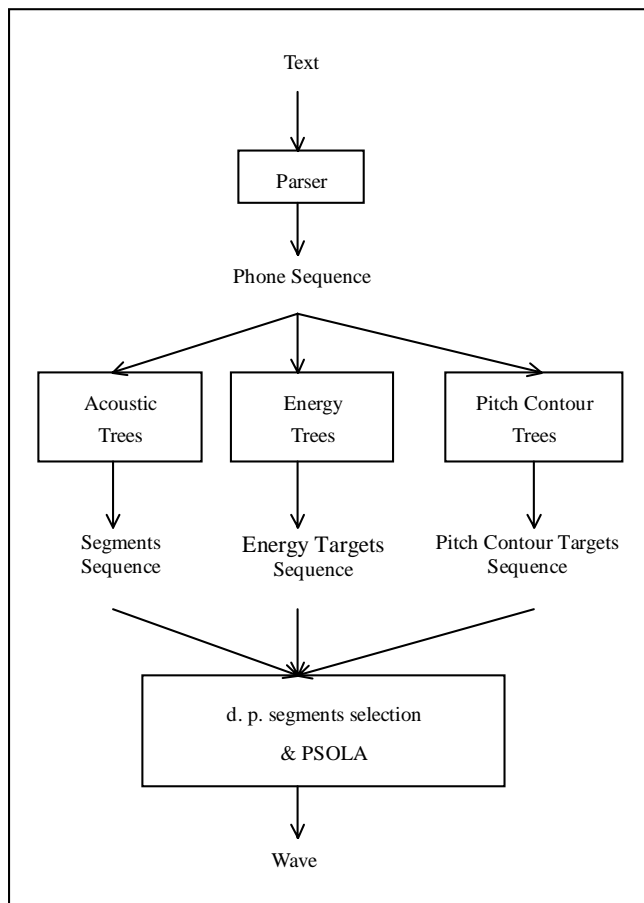
After the trees being built, all the vectors clustered to a certain leaf are used to estimate a Gaussian. The mean vector of the Gaussian is regarded as the prototype vector of the leaf. These trees and the prototype vectors attached to the leaves would be used to predict the pitch contour of the phones to be synthesised.

4.2. Phone Pitch Contour Prediction

During the runtime of our synthesiser, a parser first converts the texts to be synthesised into the phone sequences. The acoustic (HMM) decision trees of the phone states are then traversed according to the phone context, so that each state of the phone has a set of waveform segments to be selected afterwards. The phone energy decision trees are also traversed to predict the energy target of each phone.

To predict the pitch contour, the duration vector trees and the pitch vector trees should be traversed to find a duration leaf sequence and a pitch leaf sequence according to the given phone contexts. For each leaf pair in the two sequences, a pair of vectors of a duration

prototype and a pitch prototype can be used to compute the pitch contour of the phone. This computation depends on how the pitch contour vector pair is compute in the training step. In our current system, we simply assign the value of each dimension of the pitch prototype to the middle of the corresponding feneme segment according to its duration prototype, if the value is not negative. After concatenating the phone pitch contours together, the target pitch contour of the phone sequence is obtained. It is then fed to the synthesiser for d. p. search of the waveform segments and synthesising.



F 2. The Diagram Of Synthesis Procedure

5. RESULTS AND DISCUSSION

The speech produced by our previous synthesis system without pitch prediction, which only has duration and energy prediction, has obviously good formant continuity. It mainly results from the mechanism of the d. p. search. This indicates that the syllable of Mandarin, which is intuitively a natural choice and is widely used, is not necessarily to be the basic acoustic unit of the

synthesis inventory from the viewpoint of naturalness. However, there are often some parts of the speech synthesised by the system without pitch prediction sounds unnatural in tone. This problem is related to the pitch contours of the selected segments. Although the Finals in our phone set have been defined with explicit tone patterns and so do the fenemes, the waveforms selected from the inventory may have inappropriate pitch contour variations.

By adding our pitch contour prediction process, those parts of the speech with unnatural tones are improved. This is because the predicted target pitch guides the d. p. search algorithm to select more appropriate segments from the aspect of pitch contour.

In conclusion, the method of modelling the pitch contour according to the HMM states of an acoustic unit is helpful for us to compare the pitch contours variants of the acoustic unit. In our current system, it is only used in the simplest form of it. The element of V_P can be redefined in a more complicated way so as to result in a finer representation of the unit's pitch contour. In predicting the pitch contour, we take great advantages of the decision tree approach. The context-dependent variants of the pitch contour are automatically trained from the database. Note that pitch contour is affected by not only the phone context but also the stress, the syntax and the sentence intonation pattern, the decision tree based pitch contour prediction approach focuses only locally comparing to the whole sentence level. It could be a primary step of more accurate pitch contour generation for a whole sentence.

6. FUTURE WORK

To improve our current synthesis system, the pitch contour model need to be refined and the variations of the pitch contour feature caused by the factors except the phone context must be analysed and modelled. A full intonation generation method at sentence level is the focus of our future work.

REFERENCES

- [1] Z. Wu, "Tone-sandhi in Standard Chinese Sentence", *Zhongguo Yuwen*, pp439-499, 1982. (In Chinese)
- [2] H. Fujisaki, "Modelling the Process of Fundamental Frequency Control of Speech for Synthesis of Tonal Features of Various languages", *Invited Plenary Lecture, 1997 China-Japan Symposium on Advanced Information Technology*.
- [3] L. R. Bahl, P. V. deSouza, P. S. Gopalakrishnan, and M. A. Picheny, "Context Dependent Vector Quantization for Continuous Speech Recognitions", *Proc. ICASSP'93 Vol. 2*, pp632-635.
- [4] R. E. Donovan, and E. M. Eide, "The IBM Trainable Speech Synthesis System", *Proc. ICSLP'98*

- [5] R. E. Donovan, "Trainable Speech Synthesis", *PhD. Thesis, Cambridge University Engineering Department. (1996)*
- [6] R. E. Donovan, and P. C. Woodland, "Improvements in an HMM-Based Speech Synthesiser", *Proc. Eurospeech95*.
- [7] C. J. Chen, R. A. Gopinath, M. D. Monkowski, M. A. Picheny, and K. Shen, "New Methods in Continuous Mandarin Speech Recognition", *Proc. Eurospeech97*, pp1543-1546.