The Statistical Model of Chinese Word Contours Based on Fuzzy Clustering Method

Jianhua Tao Lianhong Cai Yuzuo Zhong
Information group, Department of Computer Science
Tsinghua University, Beijing, 100084
E-mail: tjh@tts.cs.tsinghua.edu.cn clh-dcs@mail.tsinghua.edu.cn

ABSTRACT

With the aim of constructing a set of prosodic rules enabling to generate high-quality synthetic speech of Chinese, tone concatenation features were investigated for Chinese words. A statistical model is developed for Chinese word pitch contours based on fuzzy clustering and analysis method. The clustering results shows that word contours are not only depending on the different combination of the tones of the adjacent syllables, but also related nearly to the phonetics of them. More research also shows that word contours are also influenced by different surroundings in the sentence, such as the position of the word, the stress degree of the word, the distance between the current word and the stressed word, the mood of the sentence, etc. The paper studies both the word contour models for isolated di-syllables and the distribution characters of them within different surroundings. It helps us greatly to develop the high-quality prosodic parameters in our TTS system.

KEYWORDS:Word contours, Fuzzy clustering, Speech synthesis

1. INTRODUCTION

As we know, Chinese is a tonal language, where four lexical tones exit for a syllable: namely, tone 1 characterized by a high-flat F0 contour, tone 2 characterized by a rising contour, tone 3 characterized by a low-dip contour, and tone 4 characterized by a failing contour form high F0. Prosody of spoken Chinese has two major factors, tone and intonation of sentence, which is produced to express emotion. The relation of tone and intonation in the Standard Chinese had been described:" the small ripples riding on large waves" (Y.R. Chao), the large waves represent stress changing, and the small ripples represent syllable or word contours. Thus, the sentence pitch contour in Standard Chinese can also be considerd being composed of a number of word tone-sandhi contours (Z.J.Wu)[3].

Mainly, word contours contain di-syllable contours, tri-syllable contours, and quad-syllable contours etc. There are more than 20,000 di-syllables of Chinese, and much more tri-syllables and quad-syllables. Even the number of di-syllables, which are used frequently, is more than 6,000. It's impossible to develop models for each of the words in Chinese.

Most important character in the word contours is the tone coarticulation between two adjacent syllables. Statistical results show that most of the figures of their contours in the same surrounding of the sentence are much similar. The paper uses fuzzy clustering method to classify a large number of the isolated di-syllable pitch contours automatically, which will be described in section [2]. It can be thought that the tri-syllable's and the quad-syllable's contour model can be extended from the di-syllable's.

Clustering results show that there are some rules in the transitional segment of the word contours, at which word contours can be classified not only by the different combination of the tones of the syllables but also by the initial of the next

syllables. The model of isolated word will be established in section [4].

Because the word contours is nearly related to the surroundings of the sentence, the paper studies the distribution characters of them in different surroundings[5].

The model of word contours is used successfully in the TTS system. Lots of practice shows that it can generate high-quality pitch contours in the sentence.

2. CLUSTERING METHOD OF WORD TONE

In this paper, 641 words of isolated di-syllable are selected carefully for clustering and statistic, in order to find the rules of the tone coarticulation and generate the word pitch contour model of isolated di-syllables.

The words are parsed by speech-analysis tool, which is developed by us for calculating the speech pitch automatically. As we know, all of the di-syllable tones can be divided into 20 types based on the different combination of the tones, such as 1-1, 2-2, 3-4, etc.

Ten points are selected on equal space, $t_0, t_1, \sim t_9$, from which the pitch value $f_j(t_i)$ can be thought to describe the main figures of the pitch. The pitch value of the previous syllable in the word is denoted by $fr_j(t_i)$, and next is $fn_j(t_i)$. Where, t_i stands for the i'th point in the syllable, and j denotes different word.

For the same combination of the tones, the average pitch contour can be calculate by,

$$\bar{F} = \frac{1}{20}\left[\sum_{i=0}^{9}\left(fr_j(t_i) + fn_j(t_i)\right)\right] \quad (1)$$

Assign: $fr_{ji}' = fr_j(t_i) - \bar{F}$ (2)

We can get the relational coefficient between two words contour,

$$R_{jk} = \frac{\sum_{i=0}^{9}\left[fr_{ji}' \times fr_{ki}' + fn_{ji}' \times fn_{ki}'\right]}{\sum_{i=0}^{9}\left[fr_{ji}' \times fr_{ji}' + fn_{ji}' \times fn_{ji}'\right] + \sum_{i=0}^{9}\left[fr_{ki}' \times fr_{ki}' + fn_{ki}' \times fn_{ki}'\right]} \quad (3)$$

For all the words, a similar matrix ($M \times M$) can be get,

$$A = \begin{bmatrix} R_{00} & R_{01} & \ldots & R_{0M} \\ R_{10} & R_{11} & \ldots & R_{1M} \\ \ldots & \ldots & \ldots & \ldots \\ R_{M0} & R_{M1} & \ldots & R_{MM} \end{bmatrix} \quad (4)$$

Where, M is the number of words, which are being analyzed.

Thus, the word pitch contours can be classified accurately, using max-tree method from the similar matrix. Thus, the model of the isolated di-syllable pitch contour $S_{ij}(t)$ can be gotten by fuzzy clustering. Where, i stands for different tone combination, and j is the index of the different word tone model in the same tone combination.

3. CLUSTERING RESULT

The result of fuzzy clustering of the combination of the tone 2 and the tone 2 is shown in Tab.1. The threshold of the correlation coefficient is 0.92. It can be seen from the Tab 1, that the initials of the next syllables in the first type include p, c, b, t, x, t, z, f, j, h, sh, zh, etc, which are mainly surds. The typical characteristic of the pitch contours in this type is the large pitch fall in the transition between the adjacent syllables, which is shown as Pit 1(a).

It can also be seen that the initials of the next syllables in the second type include m, n, l, r, y, w, etc, which are mainly sonant. The pitch contours is transited fluently in the transitional segment of the adjacent syllables, which is shown as Pit 1(b).

There are also few of words in the other figures. Considering the accuracy of the calculation and the random of the speaker, those types can be neglected.

First type	Wang2pai2,tao2ci2,ping2bai2,cong2tou2,de2xing2,mao2tou2,hui2xuan2, jue2ze2,quan2cai2,fu2fu2,lian2jie2,men2fang2,tou2sheng2,rong2shi2,pian2pian2,zhi2xing2, pei2zhi2		
Second type	Feng2ying2,qin2na2, he2yi2,qiu2,ren2,chao2liu2,cheng2men2,he2mou2,fo2 men2,ming2yan2,mang2mang2, ti2e2, shei2ren2, zei2ren2,xi2wen2		
Third type	huang2li2	Sixth type	Ming2wen2
Forth type	xie2yang2	Seventh type	Cheng2lou2
Fifth type	shei2shei2,yang2pan2	Eighth type	Mao2lv2

Table 1: The result of fuzzy clustering of the 2-2 tone combination

From the above, we can see that the pitch contours in the segment of the transition of the word depend on not only the combination of the tones of isolated di-syllables but also the initial of the next syllables. The initial and the final of the previous syllables and the final of the next syllables do little affection on it.

Figure 1: Two types of the word tone model of the 2-2 tone

Type a: there exits a large pitch fall in transition.

Type b: the pitch contours is transited fluently in the transitional

4. WORD CONTOUR MODEL FOR ISOLATED DI-SYLLABLES

For each type, the model of isolated di-syllables can be gotten as following:

Calculate the square root error of the word contours, which are in the same clustering type,

$$\overline{fr_i} = \frac{1}{M}\sum_j fr_j(t_i) \qquad (5)$$

$$\overline{fn_i} = \frac{1}{M}\sum_j fn_j(t_i) \qquad (6)$$

$$\alpha = \frac{1}{N}\sum_{j=0}^{N-1}\sqrt{\frac{1}{20}\sum_{i=0}^{9}\left\{\left[fr_j(t_i)-\overline{fr_i}\right]^2+\left[fn_j(t_i)-\overline{fn_i}\right]^2\right\}} \qquad (7)$$

The model can be gotten by calculating the average pitch contours of the word. Because the points, which exceed the bias of the square root error, may lead large warp in the average pitch contours, they must be dismissed from the calculating firstly, which is shown as following.

The average pitch can be gotten,

$$FR_i = \frac{1}{M}\sum_j fr_j(t_i) \qquad (8)$$

$$FN_i = \frac{1}{M}\sum_j fn_j(t_i) \qquad (9)$$

Where $fr_j(t_i)$, $fn_j(t_i)$ stands for pitch point of the same contour type, and $fr_j(t_i)-\overline{fr_i} \le \alpha$, $fn_j(t_i)-\overline{fn_i} \le \alpha$.

To generate high-quality word pitch contours, the result points FR_i and FN_i must be smoothed by b-spline curve according to the different duration in such surroundings of the sentence.

5. DISTRIBUTION OF THE WORD CONTOURS

Large experiments show that the word contours

will be changed a lot from the isolated word's if they are putted into the sentences or the phrases. Research show that the modification of the word contours is related nearly to the syntactic structure of the sentence or the phrase and speaking surroundings, such as word position in sentence, the stress degree, stress position, sentence mood, syntactic structure of the sentence, etc.

5.1 Influenced by sentence mode and word position

As we know, there are many sentence moods in Chinese, such as declarative, interrogative, imperative, exclamative, etc. Thus, the trends in modification of word contours in different sentence modes are various, and the same word will have different word contours, being in different position of the sentence or the phrase.

There shows a declarative sentence in Pit2, from which you can see that the pitch range of the word of the combination of tone 4 and tone 4 has been changed, being the different position of the sentence.

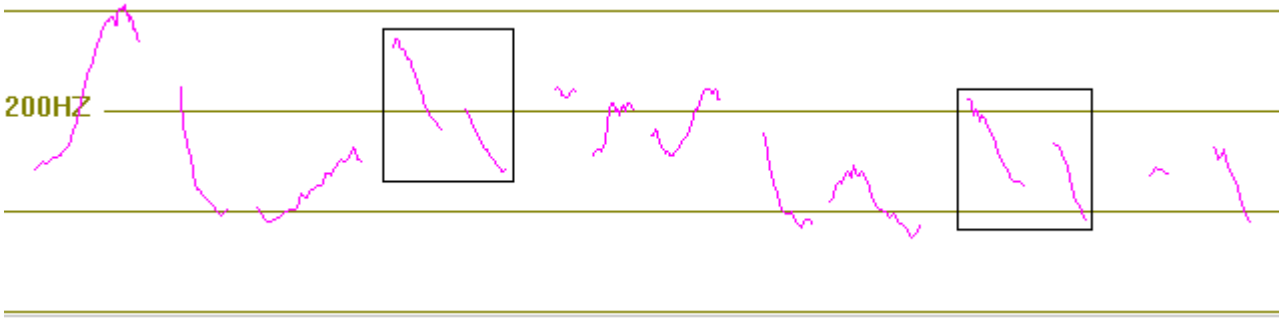

Figure 2: Pitch contour of the sentence,
"nuan3shui3liu2zheng4zai4tai4ping2yang2shui3yu4zai4du4chu1xian4" (暖水流正在太平洋水域再度出现)

5.2 Influenced by stress

If the word is stressed, the pitch contours of it will be changed either.

Pit3 show that the pitch range of a tri-syllable has been changed, being different degree of the stress.

When it is stressed, the pitch range of it will be increased [3].

Further research shows that the word contours will also be influenced by the stress degree of the adjacent words or syllables.

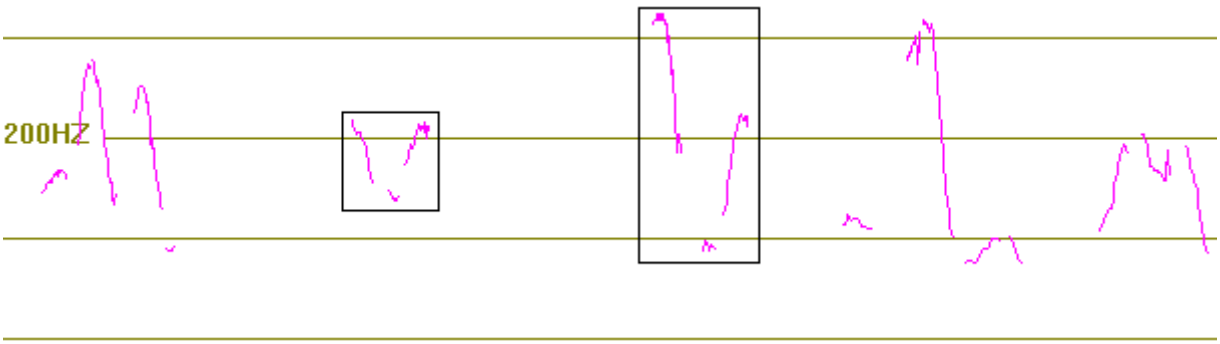

Figure 3: Pitch contour of the sentence,
"ni3gei3wo3guo4lai2gan4shen2me5gan4shen2me5ni3xin1li3ming2bai5bie2zhuang1suan4"
(你给我过来!干什么?干什么?!你心里明白,别装蒜。)

5.3 Influenced by other factors

Besides the factors, which are pointed out above, there are also some others, which can effect on the word contours, such as syntactic structure, silence between the words, etc. Their contributions to the pitch contours are in a large range.

5.4 Solution

It's a complicated problem to deal with lots of the parameters of the surroundings in sentence, which can affect the word contours. A neural network is developed, which can be thought to solve the problem very well. The notion of using a neural network, or other machine learning system, to implement components in a text-to-speech system is an attractive one. A system trained on actual speech may learn subtler nuances of variation in speech than can presently be incorporated fully into rule-based or concatenation text-to-speech system. System can also suit different styles of users. Though there has been some neural network models used in the TTS system now, most of them have been used to select the current synthetic units or generate a series of coded parameter vectors. They often require a large number of training data. And the training processing is usually slow.

We have developed a neural network with time delay in input data, which is detailed in reference [1]. The system establishes a relationship between the modification of the word contours and the complicated linguistic situation. The train databases of it are the linguistic labeling, prosodic labeling [1], and the isolated word contour, which is gotten from the output of the isolated word contour models. The output of the NN will be the real word contours in different sentences and different surroundings. The system doesn't need very large training database, but the grammatical structure of them must have been designed carefully. The data storage requirements in this system are also smaller than the others. It should also be easier to be trained on a new language than to determine a rule set for that language.

6. CONCLUSION

The models, considering the distributing of their characters in different surroundings, have been used in the TTS system successfully. It is the basis of the generation of the high-quality intonation of the sentence. The TTS system has been applied widely, such as in dialogic system, electronic lexicon, HTML reading, etc.

REFERENCE

[1] Tao Jianhua, Cai Lianhong, "The Context-based Method of Creating Chinese Prosodic Model",ISSPR98,VOL 2, P271-276

[2] Tao Jianhua, Hua Yiman, "Rule-Synthesis System of Chinese Based on Technology of PSOLA", Transaction of the Nanjing University, 1998，1 ,VOL34, NO.1, P85-89

[3] Wu,Z.J., "Tone-sandhi in sentences in Standard Chinese", Chinese of China, No.6,pp.439-450

[4] Yang Shunan, "Synthesis technology of the Mandarin Speech", Publishing of the Social Science, 1994, 4

[5] Cai Lianhong, Zhou Qiaofeng, "The prosodic modification method of the Chinese TTS system based on PSOLA technology", Transaction of the software, 1996

[6] Fujisaki,H.et al.,Analysis and modeling of tonal features in polysyllabic words and sentences of the Standard Chinese, ICSLP'90 Vol.2, pp841-844