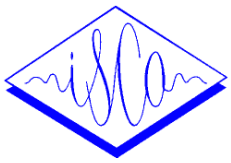

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

CHINESE FOUR TONE RECOGNITION BASED ON THE MODEL FOR PROCESS OF GENERATING F0 CONTOURS OF SENTENCES

Changfu Wang
Dept. of Electronic Engineering
Univ. of Science & Technology of China
Hefei, Anhui, P. R. China

Hiroya Fujisaki and Keikichi Hirose
Faculty of Engineering
University of Tokyo
7-3-1, Hongo, Bunkyo-ku, Tokyo, 113 Japan

ABSTRACT

This paper proposes a new method of recognizing the four tones, the method is based on the functional model for the process of generating F0 contours of sentences, so our method is different from proposed others. it is applied to the four tone recognition not only for monosyllables, but also for disyllables, phrases and even sentences. This method is adaptable to every speaker with a simple training procedure using only "MA" four tones. It can recognize the lexical tones using a few parameters. The recognition accuracy is over 99% for monosyllables, 93% for disyllables and idioms of four-syllable.

INTRODUCTION

Chinese is a syllabic and tonal language. Each chinese character is a monosyllable, while more than 50,000 characters are used in the written language, they are expressed by a total of approximately 1,300 syllables in the spoken language. If we disregard tones, these 1,300 syllables are reduced to about 400 segmentally different syllables. Thus the tones play an important role in distinguishing words, and the four tone recognition is indispensable for the recognition and understanding of the standard spoken language of Chinese (Putonghua).

Although speakers may vary in the average pitch, their F0 contours corresponding to a particular tone have a similar pattern which constitutes a distinctive feature in speech perception. The four tones are specified by the F0 contours: high level, rising, low dipping and high falling, known as the first, second, third and fourth tones respectively. The F0 contours of syllables belonging to the same tone class are roughly similar, but they are quite different in detail depending on individual syllable, speaker and environment in which the syllable is spoken. To cope with these differences, the four tone recognition should be carried out under guidance of the functional model for the process of generating F0 contours of sentences (or Fujisaki's model) [1]. According to the functional model, an F0 contour can be represented as the sum of respones of second order linear systems to two kinds of excitation commands as shown in Fig. 1 .

The impulse (phrase) commands generate the 'phrase components', which are assumed to represent intonation, and the step (accent) commands generate the 'accent components', which are assumed to represent local F0 contour variation in syllable domain. So the total response of the systems to all commands is

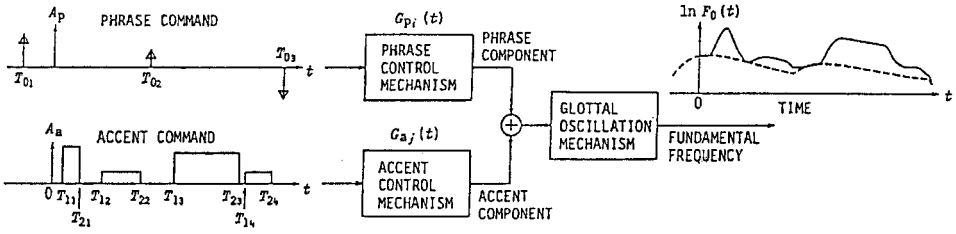

Fig .1. A functional model for process of generating sentence F0 contours.

$$\ln(F0) = \ln(Fmin) + \sum_i Api * Gpi(t - Ti) + \sum_j Aaj * [Gaj(t - T1j) - Gaj(t - T2j)]$$

$$= \ln(Fmin) + (\text{phrase components}) + (\text{accent components})$$

where

$$Gpi(t) = \alpha_i^2 t * \exp(-\alpha_i t) \quad \text{for} \quad t \geqslant 0$$
$$= 0 \quad \text{for} \quad t < 0$$

and

$$Gaj(t) = \min[1 - (1 + \beta_j t) * \exp(-\beta_j t), \theta_j] \quad \text{for} \quad t \geqslant 0$$
$$= 0 \quad \text{for} \quad t < 0$$

In order to simulate the F0 contours of Chinese four tones using Fujisaki's model, the accent commands are modified as shown in Fig. 2, known as tone commands, which generate the 'tone components' [2].

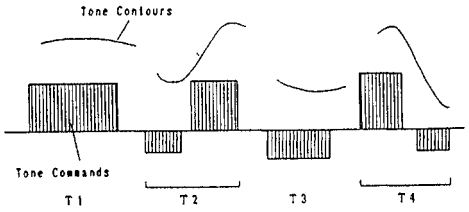

Fig. 2. Tone commands and tone contours.

The F0 contours produced by Fujisaki's model are not only close to the observed ones, but it might also be reasonable approximatipon to physical process of pitch control in human.

PRINCIPLE OF THE FOUR TONE RECOGNITION

It is clear that searching for tone commands is a method of recognizing the four tones, but it might be quite difficult to do that at present. According to Fujisaki's model, the ln(F0) contour always consists of two components: phrase component which varies slowly and tone component which varies quickly. The first tone component is response of a second order linear system to a long positive step–wise command, the second tone component is response to a secondary negative command followed by a major positive one, the third tone component is response to a longer negative command, and the fourth tone component is response to a major positive command followed by a secondary negative one, so the ln(F0) contour is level and highest averagely for the first tone, rising and higher for the second tone, dipping and lowest for the third tone, falling and higher for the fourth tone as shown as in Fig. 2. These are very close to the observed ln(F0) contours(see Fig. 3). Therefore we can recognize the four tones using these features.

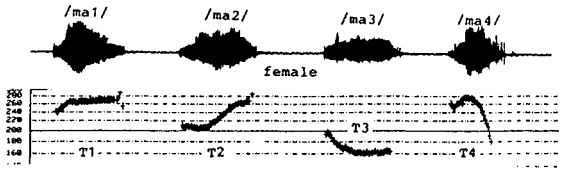

Fig. 3. "MA" four tone contours spoken by a female.

THE FOUR TONE RECOGNITION OF MONOSYLLABLE

In the case of monosyllables, because their duration is usually short, we can regard ln(Fmin)+(phrase component) as constant, the remainder is the tone component, we take the average of "MA" four tone ln(F0) contours, and regard the average as ln(Fmin)+(phrase component) (call AP), the arithmetic average of ln(F0) contour relative to AP is called APR, at the same time, we calculate the SLOPE of the straight line which approximates the syllable ln(F0) contour by the least mean square error criterion. So the APR is always larger than zero for the first tone, less than zero for the third tone, usually larger than zero for the second and fourth tones. The SLOPE is always close to zero for the first tone, much larger than zero for the second tone, much less than zero for the fourth tone, usually close to zero for the third tone. Therefore the feature parameters are expected to distribute in four separated regions as shown in Fig. 4, it is not difficult to make decision of lexical tone for monosyllables.

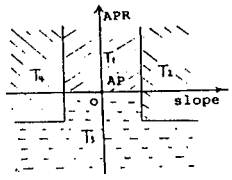

Fig. 4. Distribution of the four tone feature parameters in the SLOPE–APR plane.

THE FOUR TONE RECOGNITION OF DISYLLABLES AND FOUR–SYLLABLE IDIOMS

When isolated syllable tones are recognized, we think of ln(Fmim)+ (phrase component) as constant AP. Since the duration of disyllable and phrase is generally longer than that of monosyllable, we must take account of effect of the phrase component decay

$$\text{phrase component} = Ap\alpha^2 t * \exp(-\alpha t)$$

Let $Ap = 0.25$, $\alpha = 3.0$, and assume that $-T0$(phrase command location before the onset of an utterance) is 0.2 sec., we can calculate approximately the drop for the succeeding syllable in disyllable and phrase.

When syllables are uttered isolately, the relations between the lexical tone classes and their ln(F0) contours are quite regular and clear, although these ln(F0) contours belonging to the same tone class are fairly different in detail. However, when syllables are uttered continuously, because of the interaction between two adjacent syllables, the relation may be varied in some degree, and the variation cause some difficulties in the four tone recognition.

Through investigation of a great number of ln(F0) contours of disyllables and phrases of four–syllable, we found that the ln(F0) contour of a syllable always tries to keep continuity with its adjacent syllables, therefore deviates from its original pattern in monosyllable(see Fig. 5). In order to compensate the devtiation , the factors causing the deviation should be analyzed quantitatively using Fujisaki's model.

According to the model, the tone component of ln(F0) contour is response of a second order linear system to tone commands. Under the condition of isolated utterance of syllables, the system is always in zero–state before the tone commands are input into it, while, under the condition of continuous utterance, the system may be in non–zero–state,

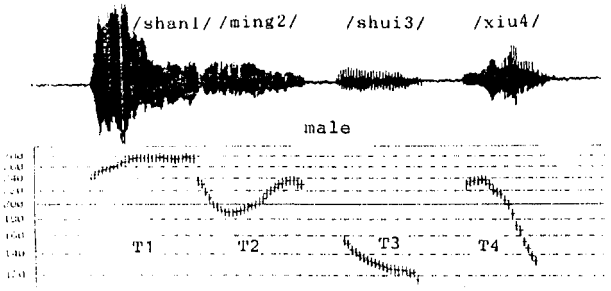

Fig. 5. The ln(F0) contours of 'SHAN(1)–MING(2)–SHUI(3)–XIU(4)'.

this is because, in later condition, when the tone commands of the succeeding syllable are input into it, the response to the tone commands of the preceding syllable may not decay completely. In order to obtain higher recognition accuracy for continuous speech using the method of recognizing the four tones of isolated syllables, we have to remove the zero–input response from succeeding tone component, if it exists. It is quite easy to do that, we cut one fourth length of the preceding syllable ln(F0) contour from its end, and take average of the length relative to AP, and assume that ln[F0(t)] of the last two frames of preceding syllable is equal to the average+AP, according to the average, the relative position of two adjacent syllables and the system's parameters $\beta = 20$ / sec.. we can calculate approximately the zero–input–response in succeeding syllable tone component, subtract the zero–input–response from the tone component, and obtain the zero–state–response of the succeeding syllable tone component, thus we can recognize the four tones of disyllables and idioms of four–syllable using the original method of recognizing the four tones of isolated syllables.

RECOGNITION RESULTS

Speech material were collected from five males and five females who speak modern standard Chinese. We selected 200 monosyllable words which distribute equally in four tone classes and contain almost all consonants and vowels of Chinese speech, 78 disyllable words which contain various combination of the four tones, and 35 idioms of four syllables which are spoken very often. Everybody spoke these materials once and recorded their voice in tapes, sampling rate is 10kHZ, the recognition experiment was carried out on VAX–Ⅱ, pitch was extracted by cepstrum analysis. Based on consideration of simplicity, we cut off small part of the ln(F0) contour near its end points in order to avoid the unfavourable influence, then we take the average APR of the ln(F0) contour relative to AP, calculate the SLOPE of straight line obtained by the least mean square error criterion, and make decision of lexical tone according to these feature parameters.

In the case of multi–syllable words, the transition part of two adjacent syllables is cut off, if their ln(F0) contours connect together. We first decide whether the system which

generates tone component is saturated or not according to the ln(F0) contour shapes of end–part in the preceding syllable and start–part in the succeeding syllable, if it is not, remove the zero–input–response from tone component for continuous speech, and take an account of the drop of AP, then calculate their APRs and SLOPEs and make decision of lexical tones.

Table 1, 2 and 3 show the recognition results, the recognition rate is over 99% for monosyllables, and 93% for disyllables and idioms of four –syllable.

Table 1. Monosyllable recognition results.

		recognition tone class			
		Tone 1	Tone 2	Tone 3	Tone 4
spoken tone class	T1	499	0	1	0
	T2	1	499	1	0
	T3	0	2	498	0
	T4	1	0	0	499

Table 2. Disyllable recognition results.

		recognition tone class			
		Tone 1	Tone 2	Tone 3	Tone 4
spoken tone class	T1	336	2	9	3
	T2	19	351	4	6
	T3	12	25	317	6
	T4	9	3	5	443

Table 3. Recognition results of four–syllable idioms.

		recognition tone class			
		Tone 1	Tone 2	Tone 3	Tone 4
spoken tone class	T1	356	1	2	1
	T2	28	427	3	2
	T3	11	31	173	15
	T4	1	0	5	344

CONCLUSIONS

According to our experiments, the four tone recognition based on Fujisaki's model is successful, the method is adaptive one for evNery speaker with simple training using "MA" four tones spoken by the speaker, the training for recognizing the four tones of monosyllables is effective for recognizing four tones in continuous speech. The selection of the feature parameters is appropriate, and decision of lexical tones is successful.

The factors which cause wrong decision are lightly reading, wrong segmentation of adjacent syllables, wrong extraction of F0 contour, tone of speakers and difficulty of selecting saturated parameters. Improvements are necessary for our present recognition method to overcome them.

ACKNOWLEDGMENT

The authors gratefully acknowledge their colleagues and friends at university of Tokyo and university of science and technology of China for their helps and supports.

REFERENCES

1. H. Fujisaki: 'Dynamic characteristics of voice fundamental frequency in speech and singing. Acoustical analysis and physiological interpretation.' Proceeding of the Fourth F.A.S.E Symposium(1981)
2. H. Fujisaki etc.: 'application of F0 contour command–response model to Chinese tones' J. Acoust. Soc. March,1988