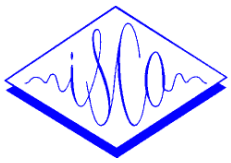

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

RECOGNITION OF CHINESE TONES IN MONOSYLLABIC AND DISYLLABIC SPEECH USING HMM

Xinhui Hu and *Keikichi Hirose*
xinhui@gavo.t.u-tokyo.ac.jp hirose@gavo.t.u-tokyo.ac.jp

Dept. of Electronic Engineering, University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113 Japan

ABSTRACT

HMM-based tone recognition methods were developed for monosyllabic and disyllabic speech of standard Chinese. Two dimensional feature vectors were used for these methods to represent well both macroscopic and microscopic features of fundamental frequency contours well. In order to realize a function of speaker normalization in the methods, an offset was introduced to the fundamental frequency. It was shown experimentally that the best function was obtained when mean fundamental frequency averaged over several word utterances of the speaker being used. The words should include those of every tone types equally. As for the disyllabic tone recognition, the developed method does not require segmentation process into syllables. Besides four lexical tone models, two models were added to represent the half 3rd tone and the first-syllabic 4th tone in 4th tone plus 4th tone sequence. The unvoiced region of a disyllable usually corresponds to the initial consonant of the second syllable. A model was also assigned to this region to reduce the coarticulation effect between two syllables. As for the neutralized tone, it was included in the 4th tone group tentatively, and, after the HMM-based recognition process, it was separated depending on the durational difference. With the developed methods above, correct recognition rate of 98.5% was achieved for monosyllables of multiple speakers, and, for disyllables of a speaker, 94.5% was obtained.

1. INTRODUCTION

In Standard Chinese, one Chinese character is pronounced in one of around 1300 distinguishable syllable sounds. As it is well known, Chinese is a typical tone language where a syllable possesses several tone types with fundamental frequency contours quite different from each other. In the case of Standard Chinese, 4 tone types are possible and, therefore, the number of syllables is reduced to 417 in phonemic level. Tone has a function of differentiating meaning. For example, 'mai3' has the meaning 'to buy,' while 'mai4' has the opposite meaning 'to sell.' (Numbers indicate tone types.) Therefore, importance of the tone recognition is very high in Chinese speech recognition and understanding as compared to the recognition of accent types in Japanese or the detection of stressed syllables in English. High performances have already been achieved for monosyllabic tone recognition in several methods, such as one approximating fundamental frequency contour with one or two straight lines [1], and another using logarithmic fundamental period (normalized by average period) and its first differential as the observation vector of discrete HMM [2]. In the case of polysyllables and continuous speech, however, several major problems will arise in tone recognition mainly due to the inter-syllabic effects on prosodic features of tones, and methods with high performance have not been reported. From this point of view, an HMM-based tone recognition method was first developed and investigated for monosyllabic words to find out the optimal code book size and the optimal method for averaging fundamental frequencies used as the offset value of speaker normalization. Investigation was also conducted on the feature vectors and the combination of macroscopic and microscopic features of fundamental frequency contours was shown to provide a good result. The method was then extended to disyllabic tone recognition. A method of concatenated learning was adopted to avoid the process of segmenting into syllables. Discussions were mainly conducted on the selection of tone models to cope with the problems on tone sandhi including tone neutralization. Discussions were also conducted on how to represent the unvoiced region of the second syllable.

2. MONOSYLLABIC TONE RECOGNITION

2.1 Extraction of Tone Features

A Chinese syllable can be decomposed into two major parts, viz., an initial consonantal part and a final vocalic part. There are cases with no initial consonantal part. On the other hand, the final vocalic part always exists and consist of a single vowel or a diphthong followed or not followed by an ending nasal. Initial consonantal part may or may not be voiced, while the final vocalic part is always voiced. Since a Chinese tone is mainly characterized by the fundamental frequency contour of the voiced part, usually called as the 'tone contour,' monosyllabic tone recognition can be conducted only by inspecting this contour. **Fig. 1** shows an example of tone contour for each of four tone types. Henceforth, T_1, T_2, T_3 and T_4 respectively indicate tones 1, 2, 3 and 4. From observations on concrete tone contours, it is known that there are large variations in the beginning and the end parts of a contour. In order to obtain a good result without extra process for these parts, two short periods with 1/10 of the total contour length were excluded respectively from both sides of the contour. The fundamental frequencies were extracted frame by frame using autocorrelation function with frame length proportional to the time lag. Extraction errors, such as half- and double-pitch errors, were corrected by comparing the extracted contour with the smoothed one [3]. If the voiced part included frames without fundamental frequencies extracted, they were compensated by

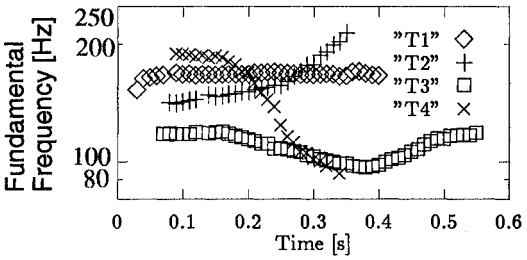

Fig. 1 An example of tone contour for each of Chinese monosyllabic tones.

the interpolation scheme using frequencies of adjacent 6 frames of preceding and succeeding parts to produce a continuous tone contour. The following tone feature vector was then calculated for the contour:

$$\begin{cases} U_i = a_i, \\ V_i = \log F_i - \log F_{offset}, \end{cases} \quad (1)$$

where F_i is the fundamental frequency at ith frame, a_i is degree of inclination of tone contour at ith frame defined as the slope coefficient of recursive line for 5 successive frames ($i-4 \sim i$), and F_{offset} denotes the offset value for the speaker normalization obtained as the averaged value of fundamental frequencies over several speech samples of the speaker. The elements U_i and V_i respectively represent macroscopic and microscopic features of a tone contour.

After the extraction of tone feature vectors, vector quantization (VQ) was utilized to convert them into a sequence of symbols which will be applied to the discrete Hidden Markov models for tone recognition. The algorithm used in the current method is the LBG algorithm. The topology shown in **Fig. 2** was adopted for the HMM.

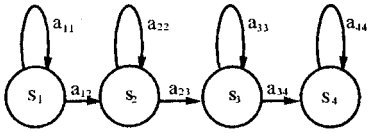

Fig. 2 Topology of monosyllabic HMM.

In order to find the following two values, experiments were conducted on the monosyllabic tone recognition.

- Optimal code book size.
- Optimal offset value for speaker normalization.

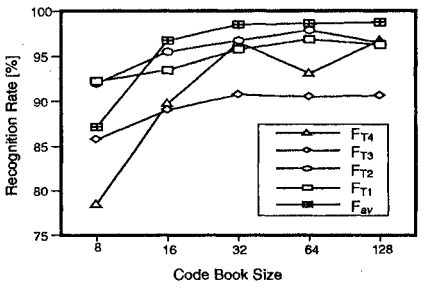

Fig. 3 Results of monosyllabic tone recognition.

2.2 Experiment

Utterances of 200 monosyllabic words from each of 7 male and 8 female speakers were recorded and used for the experiment. Those from 3 male and 3 female speakers were used as the training data of the models, while the others were used for the recognition. The recognition result is shown in **Fig. 3** as the average for the 10 speakers. Here, $F_{T1}, F_{T2}, F_{T3}, F_{T4}$ and F_{av} respectively denote the fundamental frequency averaged over several utterances of T_1, T_2, T_3, T_4 and that regardless of tone types, which are used as the F_{offset} of the speaker.

The results indicate that the recognition rates increase as the code book size increases, but are saturated around the code book size of 32. As for the F_{offset}, best result was obtained for F_{av}. From these results, the code book size was fixed to 32 and F_{av} was adopted as F_{offset} for the rest of the paper.

3. DISYLLABIC TONE RECOGNITION

3.1 Characteristics of Disyllabic Tone Contours

Although tone contours of T_1 and T_2 are little affected by their location in a word, due to the mutual interaction of constituent syllables, tone contours of disyllabic words are usually different from those obtained as mere concatenation of monosyllabic contours, especially in the followings points:

1. In the case of T_3T_3, the tone contour of first T_3 has a similar shape to that of T_2. (First syllable in the T_3T_3 sequence should be treated as T_2.)
2. Except T_3T_3, the tone contour of T_3 in the first syllable position is heavily depressed to a nearly flat shape. This is the phenomenon known as 'half 3rd tone,' which is denoted here as T_{3h}.
3. The tone contour of first T_4 in T_4T_4 does not show a major downfall as in the case of isolated utterance. Henceforth, this type of T_4 shall be denoted by T_{4h}.
4. Occasionally the tone contour of a second syllable is neutralized (depressed) and lose its original shape. This phenomenon is known as 'light tone ,' denoted as T_0.

Adding to the points above, we should also note that a disyllabic tone contour is divided into two parts by an unvoiced region corresponding to the initial consonantal part of the second syllable, if it is voicedless. If the initial consonant is voiced or null, a continuous contour is observed. The above facts imply that the method developed for monosyllabic tone recognition cannot be used for disyllabic words only by minor modifications. Considerations are necessary especially on the selection of HMMs and on the processing of the unvoiced region in a contour.

3.2 Selection of Tone Models

The tone contour of T_{3h} lacks its rising part originally shown in T_3 in isolated utterances, and indicates a rather flat shape. As for the T_{4h}, although it has a similar declining contour as that of the T_4 in the second syllable, the contour is shorter in length and higher in fundamental frequency than the second T_4. Based on these observations, individual models were added for T_{3h} and T_{4h}.

3.3 Investigation of the Light Tone

Light tone syllable can only be seen in the rear or at the midst of polysyllables or sentences. It does not appear in isolated utterances of monosyllables. This is a phenomenon of tone neutralization where it has no definite shape of tone contour. In our speech data of disyllables, 36 light tone syllables have falling contours among the total of 39 light tone syllables. When we listened to these light tone syllables, they more or less sounded like T_4 syllables. Therefore, T_0 was temporarily treated as T_4 in the HMM recognition. To distinguish light tone from T_4, a post processing was conducted using durational information for the recognition result obtained from the HMM recognition process.

3.4 Representation of Unvoiced Region

Tone recognition process for monosyllables was limited to the voiced part with a continuous tone contour. In the case of disyllables, however, two different forms are possible for their tone contours. As mentioned already, one is a continuous curve when the initial consonant of the second syllable is voiced (or null), and another is a contour separated into two parts when the initial consonant of the second syllable is unvoiced. If disyllables are successfully segmented into syllables, the monosyllabic tone recognition can be adopted where the unvoiced region will be disregarded. The accurate segmentation may be possible for disyllables, but is rather hard for longer units. Segmentation-based methods may have only limited performances for continuous speech. Therefore, we have constructed a method for disyllabic tone recognition without pre-segmentation, taking the adaptability of the method to continuous speech into consideration. One of the major problems in adopting this method is how to represent the unvoiced region. In our method, (0,0) was assigned for each frame of the region. Concretely, when there was no unvoiced points during $i-4 \sim i$, the feature vectors were computed by Equs. **1**, otherwise, they were computed by the following Equs. **2**:

$$\begin{cases} U_i = 0, \\ V_i = 0. \end{cases} \tag{2}$$

3.5 Search of Unvoiced Region

Fundamental frequencies were extracted in the same way as in the monosyllablic tone recognition. Then, the unvoiced region in a disyllabic tone contour was searched. This information on the existence of the unvoiced region is necessary for the trainning of the unvoiced model. As the first step, voiceless parts with length longer than a threshold (for the current experiment, 5 sampling points) were searched for the whole period of the tone contour. If no such part was found, the initial consonant part of the second syllable was considered to be voiced (or null) and the contour was treated as continuous. In this case, the contour may sometimes contain frames without fundamental frequencies. Fundamental frequencies of these frames were compensated by the interpolation scheme using those for neighboring voiced frames. If one voiceless part was found longer than the threshold, it was regarded as the initial voiceless consonantal part of the second syllable. If more than two voiceless parts were found, longer one was selected as the voiceless consonantal part and, the other(s) was treated as the result of pitch extraction errors. The interpolation scheme was also adopted for the shorter part(s).

3.6 Training of HMM

A syllabic HMM used for disyllabic tone recognition has 4 states and 3 loops as shown in the upper side of **Fig. 4**.

One HMM was prepared for each of 4 tone types and T_{3h} and T_{4h}. The training of HMM was conducted using the speech material without syllable segmentation. Only the information on tone sequence for each disyllabic speech was referred to during the training process. Based on the tone sequence information, disyllabic HMM was produced by concatenating corresponding syllabic HMMs. The concatenation method is shown in **Fig. 4**. After the training of disyllabic HMMs, the decomposition was conducted in a converse manner as shown in **Fig. 4** to obtain a syllabic HMM for each tone.

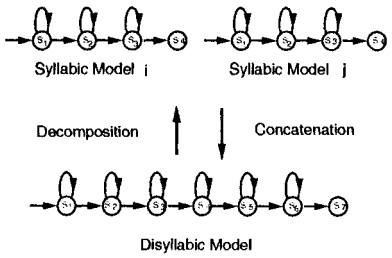

Fig. 4 Concatenation and decomposition of HMMs.

3.7 Recognition and Post Processing

Full Viterbi search was adopted for the matching process. As for the inter-syllabic transition, the transition probability to the last state of the preceding syllabic HMM was used as the transition probability to the first state of the succeeding syllabic HMM. The result of HMM-based search was given in the form of tone label sequences.

When the second syllable has no initial consonant, the whole tone contour includes no unvoiced region and, therefore, the contour is represented by one continuous curve. In this case, the contour of the transition part sometimes produces insertion errors. For example, disyllabic word 'qi4 wen1,' shown in **Fig. 5**, has one continuous contour of fundamental frequency. The transition part with rising contour will result in T_2. Therefore, the final result of $T_4T_2T_1$ sequence will be obtained . A post processing is necessary to reduce this type of insertion errors. The total procedure of the post processing is as follows:

- When the second region with a tone label is shorter than or equal to a threshold (for the current experiment, 5 frames), this region is merely regarded as the insertion error and deleted simply from the result sequence.

- When the second region is longer than the threshold, it is merged to the first or the third region by checking its location in the disyllable. If the center of the second region is located before the center of

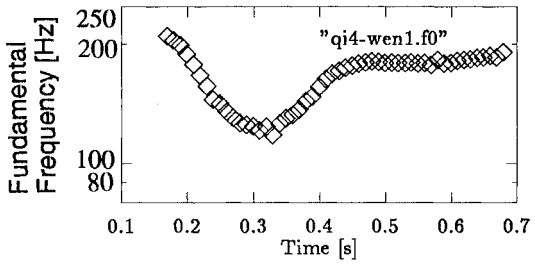

Fig. 5 Tone contour of the disyllabic word 'qi4-wen1.'

Table. 1 Mean syllable length of each tone in the 1st and in the 2nd syllable positions.

	T_1	T_2	T_3	T_4	T_0
1st Syllable	0.259	0.246	0.260	0.222	-
2nd Syllable	0.263	0.232	0.342	0.202	0.170

the disyllable, it is merged to the first region, and *vice versa.*

- For the merged part, the ratio is calculated for the length of two regions. If the ratio exceeds a threshold (5, for the current experiment), the tone label for the longer region is regarded as that of the merged part.

- If three regions with tone labels still remain after the process above, one of them is deleted using the following rules:

  T_4 X Y $\Longrightarrow$ T_4 Y (X=T_2, T_3) (Y is arbitrary)

  Y T_4 X $\Longrightarrow$ Y T_4 (X=T_2, T_3) (Y is arbitrary)

  T_3T_2 X $\Longrightarrow$ T_3 X (X is arbitrary)

  X T_3T_2 $\Longrightarrow$ X T_2 X(is arbitrary)

  T_1T_4 X $\Longrightarrow$ T_1 X (X is arbitrary)

  T_1T_3 X $\Longrightarrow$ T_1 X (X is arbitrary)

  X T_2T_1 $\Longrightarrow$ X T_1 (X is T_3 or T_4)

For the cases not included in the above process, selection was conducted only by the longer one basis.

For the result of T_4, another post processing was conducted to further distinguish T_4 and light tone which was assumed within a T_4 model at the training stage. We used the mean lengths of light tone and T_4 to judge which the result belongs to. **Table. 1** shows the mean duration of each tone in two different syllable positions.

3.8 Experiments and Results

With the methods above, the disyllabic tone recognition experiments were conducted for one male speaker. The speech material was 357 of disyllabic utterances which include 39 neutralized tones (light tones). Of which, 200 utterances were used for the training data, the left 157 were used for the recognition. The experiments included following cases:

- Case A: Using only four lexical tone models.
- Case B: Adding T_{3h} model to the case A.
- Case C: Adding T_{4h} model to the case B.
- Case D: Using lexical tone models only, but different models for the first and the second syllables. The total number of models is 8.

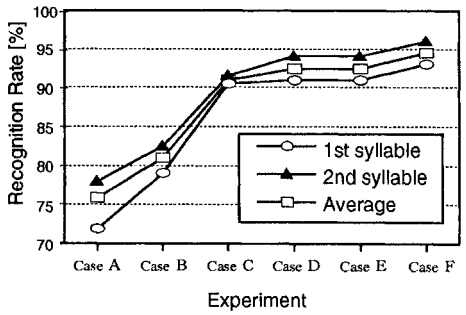

Fig. 6 Recognition results of disyllabic tones.

The above experiments were conducted with interpolating the unvoiced region linearly.

- Case E: Basically the same as the case C except the unvoiced region being represented by vector (0,0).
- Case F: Adding an individual model for the unvoiced region to the case E.

The recognition results are shown in **Fig. 6**. From the figure, it is known that, by adding T_{3h} and T_{4h} models, the recognition rates increased by 6% and 8.5% respectively. With the different models for each tone in different positions as in the case D, only little improvement was achieved. The recognition rate was almost the same with that obtained for the case E with 7 models. The validity of unvoiced model is clear if we compare the results of cases E and F. The use of unvoiced model has another advantage that it makes the post processing easier. The unvoiced model can be utilized to segment the disyllables with candidate tone labels into two portions. With the method of case F, the average recognition rate was reached to 94.5%.

4. SUMMARY

In this paper, HMM-based tone recognition methods were developed for standard Chinese monosyllables and disyllables. A combination of macroscopic and microscopic features was proposed. A proper offset for the feature vectors was found to be useful for speaker normalization. Addition of T_{3h}, T_{4h} and unvoiced models to the ordinary 4 lexical tone models were proved to be valid for the disyllabic tone recognition.

REFERENCES

[1] Wang Hangliang, Takayoshi Nakai, Hisayoshi Suzuki, "*Recognition of Chinese Tone Using Regression Lines*," Transactions on Institute of Electronics, Information and Communication Engineers, D, Vol. J71-D, No.2 pp.257-264, (1988-2).

[2] Wu-Ji Yang, Jyh-Chyang Lee and Yueh-Chin Chang, "*Hidden Markov Model for Mandarin Lexical Tone Recognition*," IEEE Transaction on ASSP, Vol.36, No.7, pp.988-992, (1988-7).

[3] Keikichi Hirose, Hiroya Fujisaki, Shigenbu Seto, "*A Sheme for Pitch Extraction of Speech Using Autocorrelation Function with Frame Length Proportional to the Time Lag*," Proc. ICASSP-92, San Francisco, Vol.I, pp.149-152 (1992-3).