Tone Recognition of Continuous Speech of Standard Chinese Using Neural Network and Tone Nucleus Model

Keikichi Hirose[1], Hui Hu[2], Xiaodong Wang[2], & Nobuaki Minematsu[3]

[1]Dept. of Inf. and Commu. Engineering, School of Inf. Science and Tech.
[2]Dept. of Electronic Engineering, School of Engineering
[3]Dept. of Frontier Informatics, School of Frontier Sciences
University of Tokyo, Tokyo, Japan
{hirose, huhui, wxd, mine}@gavo.t.u-tokyo.ac.jp

Abstract

A method is developed for recognizing lexical tone types of Standard Chinese syllables in continuous speech. Neural network (four-layered perceptron) is adopted as classifier. The method includes two steps; first recognizing tone types using prosodic features of voiced part, and then re-recognizing by viewing only on tone nucleus, which is a portion of the syllable showing rather stable fundamental frequency (F_0) contour regardless of tone types of the preceding and following syllables. The voiced part (or tone nucleus) is divided into 20 segments, and F_0, delta-F_0, F_0 slope and short-term energy of each segment are served as inputs to the neural network. In order to cope with tone coarticulation, prosodic feature parameters for the last 5 segments of the preceding syllable and the initial 5 segments of the following syllable are included in the neural network inputs. Information on syllable length is also added to the inputs. Tone recognition experiment was conducted for a female speaker's utterances included in HKU96 corpus. The average recognition rate was 86.5 % including neutral tone syllables, when the tone nucleus model was not used. It increased to 86.9 %, when the model was used. The obtained rate is higher by more than 3 points as compared to that obtained by the hidden-Markov-model-based tone recognizer developed by the authors formerly.

Index Terms: tone recognition, tone nucleus model, neural network, Standard Chinese

1. Introduction

As a typical tone language, a Chinese syllable can possess several meanings according its tone types. In the case of Standard Chinese, around 1,300 distinguishable sounds are possible if tone types are taken into account, though the number of monosyllabic sounds is only around 420 in phonemic level. Therefore, tone recognition comes an important issue in Chinese speech recognition. However, in most current systems for continuous speech recognition, syllable tone types are not utilized; semantic ambiguities are dissolved using language models. This situation comes from the complexity of acoustic manifestations of syllable tone types when they are uttered in continuous speech; a reliable method for tone recognition has not yet been developed for continuous speech.

Tone types of syllables are well characterized by their fundamental frequency (F_0) contours, when they are uttered in isolation. However, in continuous speech, their F_0 contours are changed to a large extent due to effects from neighboring syllables. Representing F_0 sequences and other prosodic features in the framework of hidden Markov model (HMM) is a possible answer for coping with this tone coarticulation [1, 2]. By incorporating several sophisticated schemes, such as context dependent modeling, tone nucleus modeling, and so on, the average tone recognition rate exceeded 80 %, given the syllable boundary information [3]. The score is rather high, but still insufficient to be included in the speech recognition process. Another candidate of tone recognition is to use neural networks, which shows high discrimination ability when dealing with a limited number of categories. When syllable boundaries are known, a neural-network-based method can easily be constructed with high performance [4]. Influenced this situation, we recently developed a tone recognition method using a neural network in four layered configuration. By taking prosodic features of preceding and proceeding syllables into account and by incorporating the tone nucleus model, a tone recognition performance surpassing that for our HMM-based method was obtained.

The rest of the paper is organized as follows: A brief explanation of the tone nucleus model is given in Section 2 with a method of automatic extraction of tone nucleus from observed F_0 contour. The proposed method is outlined in Section 3. Section 3 also includes explanation on acoustic parameters for tone recognition. The configuration of neural network is shown in Section 4. In Section 5, tone recognition experiments are explained in detail with a brief explanation on the speech material. After a discussion on how we can improve the performance in Section 6, Section 7 concludes the paper.

2. Tone nucleus model

In Standard Chinese, there are four lexical tones attached to each syllable. They are referred to as T1, T2, T3 and T4, which are characterized by high-level, high-rising, low dipping, and high-falling F_0 contours, respectively. Besides the lexical tones, there is also a so-called neutral tone (T0), which does not possess its inherent shape in the F_0 contour. Its F_0 contour varies largely with the preceding tone. The neutral tone occurs not only on certain particles; any lexical tones can be neutralized in an unstressed syllable, for example, in the second syllable of some bi-syllabic words.

For a syllable F_0 contour, only its later portion, approximately corresponding to the final vocalic part, is regarded to bear tonal information, whereas the early portion is

regarded as physiological transition period from the previous tone. It was also found that there are often cases where voicing period in the ending portion of a syllable also forms a transition period of vocal vibration and contributes nothing to the tonality. From this consideration, a tone nucleus model, which divides a syllable F_0 contour into three segments according to their roles in the tone generation process, was proposed [3]. The three segments are called onset course, tone nucleus, and offset course, respectively, which are defined as follows:

1. Onset course is an F_0 transition from the preceding syllable to the onset target of the tone nucleus. This segment covers the initial consonant and the transition period of the final vocalic part.
2. Tone nucleus is a portion where F_0 contour keeps the basic pattern of the tone unless it is affected by high-level prosodic factors such as neutralization, contextual effect, focus, phrasing, and etc. This segment covers the nucleus of the final vocalic part.
3. Offset course is an F_0 transition from the offset target of the tone nucleus to the following syllable. This segment holds the ending course of the final vocalic part.

Figure 1 illustrates syllable F_0 contours with possible articulatory transitions for the four lexical tones. It shows how the three segments are defined on the F_0 contours. Among the three segments, only tone nucleus is obligatory, whereas the other two segments are optional; their appearance depends on voicing characteristics of initial consonant, syllable duration, context, and etc.

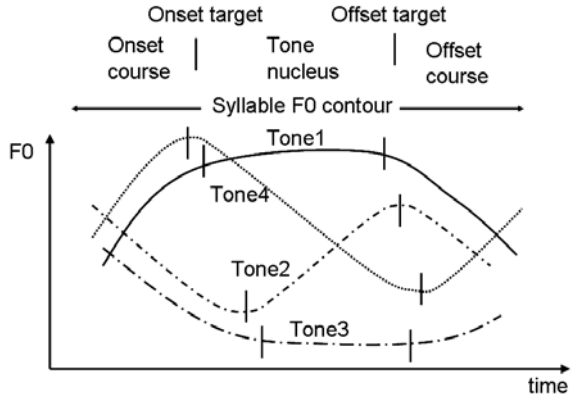

Figure 1*: Tone nuclei for the four lexical tones.*

Given a syllable F_0 contour, its tone nucleus is determined through the following process based on the K-means iterative method. Detail of segmentation procedure is found in [3].

1. Initial segmentation: Divide the contour into three segments with equal lengths. Assume the distribution of vector (F_0, delta-F_0) of each point of each portion as multivariate Gaussian, and calculate its mean vector and covariance matrix. Here, delta-F_0 of point i is defined as $F_0(i)$- $F_0(i$-1).
2. Re-segmentation: Each point of the contour is grouped to one of three segments, which has the highest likelihood. According to the result, the contour is re-segmented into three parts. Calculate mean vectors and covariance matrices for these parts.
3. Repeat the process 2 till the segmentation converges. The second part is assumed as the tone nucleus.
4. When the conversion process is failed or when the detected tone nucleus is shorter than 60 ms, the re-segmentation process is ignored and the second part of the initial segmentation is assumed as the tone nucleus.

3. Method outline and acoustic parameters used for tone recognition

The proposed method consists in two steps: in the first step, the acoustic parameters for the voiced part of the syllable in question are used and in the second step, those only for the tone nucleus are used. Since reliable results are not obtained for tone nucleus detection in the case of Tone 3 and neutral tone, whole voiced parts are used again in the second step for syllables recognized as those tone types in the first step; tone nucleus model is adopted only for Tones 1, 2 and 4. In the both steps, the voiced part and the tone nucleus are divided into 20 segments with equal lengths, and, for each segment, averages of F_0, delta-F_0, and short-term energy of its constituting points are calculated to be input parameters of the neural network for tone recognition. In the current work, interval of two succeeding points is set to 2 msec. The frame length of the analysis is 30 msec. The number of division is decided to 20 through a preliminary experiment; the best result is obtained among 3-, 9-, 20- and 40-segments. F_0 slope of each segment defined as (F_0e - F_0b)/L is also added to input parameters. Here, F_0e, F_0b and L denote F_0 of the last point, that of the initial point, and length of the segment, respectively. To cope with tone coarticulation effect, F_0, delta-F_0, short-term energy and F_0 slope of the last 5 segments of the preceding syllable and the first 5 segments of the following syllable are added to input parameters of the tone type recognition of the current syllable. In order to facilitate the tone recognition process using the neural network, all the input parameters are normalized so that their values are mostly between 0 and 1. As for the slope parameter, it is grouped into three classes, which are represented using three input elements (S1, S2, S3) as follows.

1. Slope < -12 (Hz/msec): S1=1, S2=0, S3=0
2. -12 (Hz/msec) ≤ Slope ≤ 6 (Hz/msec): S1=0, S2=1, S3=0
3. Slope > 6 (Hz/msec): S1=0, S2=0, S3=1

A better result is obtainable with this special treatment of the slope parameter as compared to simply representing them using one input element. As for F_0, delta-F_0, short-term energy, one input element is used to represent each one. Totally, 6 elements are allotted to represent acoustic features of each segment.

Beside parameters for each segment, averages of F_0, delta-F_0, short-term energy and F_0 slope among 20 segments are added to the input parameters as features representing the whole syllable. The lengths of syllable and voiced part (or tone nucleus) are also added, resulting in six input elements to represent the syllable as a whole.

In our former tone recognition using HMM, F_0 values were viewed in logarithmic scale [3]. However, in the current method, they are treated in linear scale, because a better result is obtainable. This situation may be related to the normalization of input parameters for neural network. Also, the situation may change if speaker independent tone recognition is addressed.

(The speech material for the current experiment is utterances by a female speaker.)

4. Neural network

Configuration of the neural network used for the tone recognition is a four-layered perceptron as shown in Fig. 2. Each small square symbol indicates an element. Based on the preliminary experiment, the number of hidden layers is set to two. Elements of the first five lines and those of 26th to 30th lines of the input layer accept acoustic features of the preceding syllable and those of the following syllable, respectively. The 6th to 25th lines are used to input the acoustic features of the current syllable in question. The bottom line is for the acoustic features representing the syllable as a whole. The output layer consists of 5 elements, each corresponding to a tone type.

The neural network was constructed using Stuttgart Neural Network Simulator [http://www-ra.informatik.uni-tuebingen.de/SNNS/UserManual/UserManual.html] and trained by the back propagation scheme. The number of training cycle was set to 1,000.

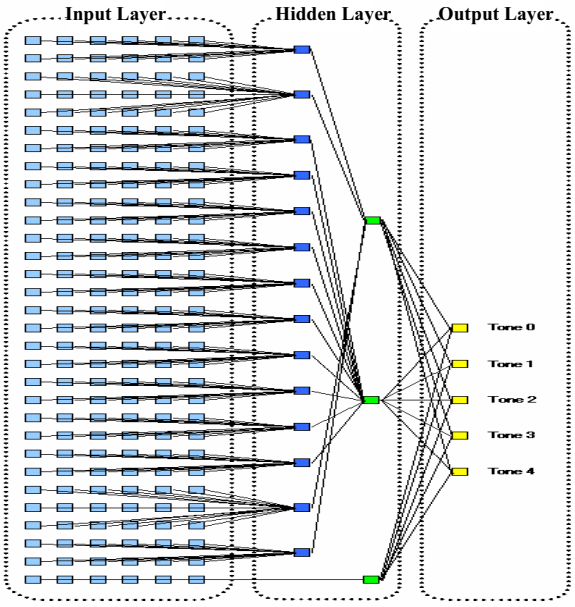

Figure 2*: Configuration of the neural network.*

5. Experiment

5.1. Speech Material

Experiment on the tone recognition was conducted for speech data uttered by a female speaker (0f), included in HKU96 (Hong Kong University 1996) Mandarin Speech Corpus. We used her data so that a direct comparison is possible between the neural-network-based method proposed in this paper and the HMM-based method developed by the authors formerly [3]. Also, she was involved in developing the HKU96 as a phonetician, and her speech sounded quite natural as compared to that by other speakers. Although the current experiment is speaker dependent, it will be enough to show the advantage of the proposed method as compared to the former one. The issue of speaker independent tone recognition is left for the future work. We took 150 sentences, which included a total of 1,955 syllables (54 T0 syllables, 396 T1 syllables, 400 T2 syllables, 247 T3 syllables and 858 T4 syllables). The cross validation was conducted for the tone recognition experiment; each time, 100 syllables are left for the testing and the rest 1,855 syllables are used for training the neural network. The total of 20 runs are conducted by changing syllables for testing. The test syllables are 55 for the last run.

The proposed method requires syllable boundaries to calculate parameters used for tone recognition. For the current experiment, those labeled in the corpus are used.

5.2. Experiment without tone nucleus model

Experiment for tone recognition was first conducted without tone nucleus model (without the second step). The average tone recognition rate 86.5 % was obtained. As shown in Table 1, recognition rate for neutral tone (Tone 0) is quite low. The recognition rate is the highest for Tone 4, being followed by Tone 2, Tone 1, and Tone 3.

Table 1. *Confusion matrix of tone recognition when tone nucleus model is not used. Each number is in %. The numbers in parentheses are total numbers of syllables in the data used for the experiment.*

Result Input	Tone 0	Tone 1	Tone 2	Tone 3	Tone 4
Tone 0 (54)	*55.56*	1.85	16.67	9.26	16.67
Tone 1 (396)	2.02	*82.32*	6.82	1.26	7.58
Tone 2 (400)	1.75	6.50	*85.00*	3.50	3.25
Tone 3 (247)	1.62	0.81	4.86	*78.95*	13.77
Tone 4 (858)	1.63	2.10	0.70	2.33	*93.24*

5.3. Experiment with tone nucleus model

Table 2: *Comparison of recognition rates in % without and with tone nucleus model.*

Tone Type	Without	With
Tone 0 (54)	55.56	50.00
Tone 1 (396)	82.32	82.83
Tone 2 (400)	85.00	86.50
Tone 3 (247)	78.95	80.16
Tone 4 (858)	93.24	93.12
Average	86.49	86.85

Then tone recognition was conducted with tone nucleus model (including the second step). As shown in Table 2, an improvement is clear for Tone 3, while some degradation

occurred for Tone 0. Totally a slight improvement was realized. In our former tone recognition experiment based on HMM modeling, the tone recognition rates for Tone 0, Tone 1, Tone 2, Tone 3, Tone 4 were respectively 31.7 %, 83.6 %, 84.5 %, 68.0 %, and 90.7 %. A remarkable improvement in Tone 3 recognition was realized by the neural-network-based method.

6. Discussion

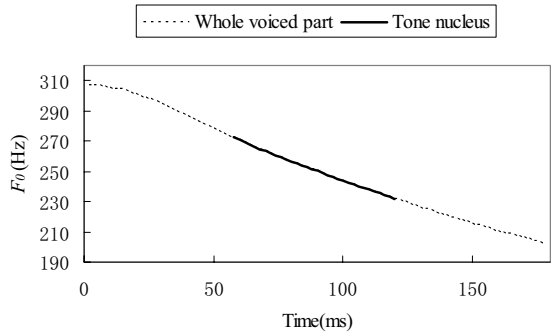

Figure 3*: An example of detected tone nucleus for Tone 4.*

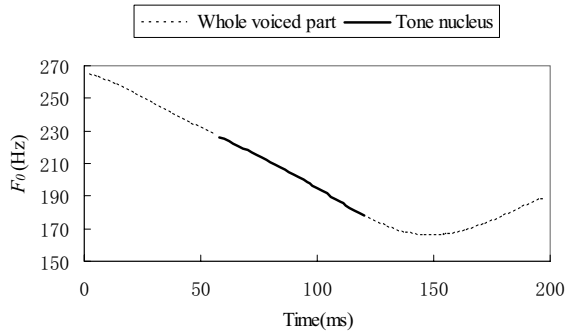

Figure 4*: An example of detected tone nucleus for Tone 3.*

Since tone coarticulation affects a lot on F_0 contours especially of Tone 3 syllables, accurate recognition of Tone 3 is a tough issue in tone recognition of continuous speech. Although the recognition rate for Tone 3 reached 80 % by the proposed method, the rate is still lower than other tone types. One possible answer to solve this situation is to develop a reliable method for tone nucleus detection for Tone 3. Figures 3 and 4 show examples of tone nuclei detected by the procedure explained in section 2 for Tone 4 and Tone 3, respectively. The detected nucleus for Tone 3 shows falling F_0 contour, which is quite close to that for Tone 4. This is the reason why we cannot use tone nucleus model for Tone 3. It is clear, the dipping pattern, which is typical for Tone 3, is not well detected by the current procedure. We are planning to add delta^2-F_0 to features of tone nucleus detection: useful for detecting parabolic shapes.

Another problem is the low recognition rate for Tone 0. Since Tone 0 does not show stable prosodic features, it is rather difficult to improve the performance only from information related to F_0, power and duration. Articulation of Tone 0 syllables may be neutralized, and some acoustic parameters representing this phenomenon need to be incorporated. Also, Tone 0 syllables often occur at sentence/phrase final. This information can be used to modify the "Tone 0 likelihood."

Furthermore, the proposed method assumes syllable boundaries, which are not labeled to the speech to be recognized. We are planning to implement the developed tone recognition method into a speech recognizer with two-pass search scheme: quick course search for the first pass and detailed search for the second pass. Tone recognition process will be included in the second pass. Since tone nucleus is considered to be robust to errors in syllable boundary locations, advantage of viewing only for tone nucleus over viewing whole syllable may come clearer.

7. Conclusion

A tone recognition method for continuous speech of Standard Chinese was developed using neural network as the classifier. To cope with tone coarticulation, the method views prosodic features not only for the syllable in question but also for the preceding and following syllables. Tone nucleus model is also adopted to cope with the tone coarticulation issue. As a whole, correct recognition rate exceeding 86 % was obtained: rate with 3 points better than that of our former work using HMM [3].

As for the future work, we will develop a method of stable detection of tone nucleus for Tone 3. Also, we are planning to apply the proposed method for other speakers (speaker independent tone recognition). As mentioned already, implementation of the method to a continuous speech recognition process is also in the scope of the near future work.

8. Acknowledgement

The authors' sincere thanks are due to Dr. Jin-Song Zhang, ATR for his useful discussion on tone nucleus.

9. References

[1] Yang, W.-J., Lee, J.-C., Chang Y.-C., and Wang, H.-C., "Hidden Markov model for Mandarin lexical tone recognition," *IEEE Trans. Acoustics, Speech and Signal Processing*, 36 (7), 988-922, 1988.

[2] Hu, X.-H and Hirose, K., "Tone recognition of Chinese dissyllables using hidden Markov models," *IEICE Trans. Information and Systems*, E78-D (6), 685-691, 1995.

[3] Zhang, J.-S and Hirose, K., "Tone nucleus modeling for Chinese lexical tone recognition," *Speech Communication*, 42 (3-4), 447-466, 2004.

[4] Chen, S.-H. and Wang, Y.-R., "Tone recognition of continuous Mandarin speech based on neural networks," *IEEE Trans. SAP*, 3 (2), 146-150, 1995.