社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2006－SLP－64 (19)
2006／12／21

# 声調核モデルとニューラルネットワークを用いた標準中国語連続音声の声調認識

王 暁東[I] 広瀬 啓吉[I] 張 勁松[II] 峯松 信明[I] 江 振宇[III] 王 逸如[III] 廖 元甫[IV]

I 東京大学 〒113-8656 東京都 文京区 本郷 7-3-1
情報通信研究機構 〒619-0288 京都府 けいはんな学研都市 光台２－２－２
III台湾交通大学 新竹市大学路 1001 号
IV台北科技大学 台北市忠孝東路三段一号

Tel. 03-5841-6667

E-mail: （wxd, hirose, minematsu）@gavo.t.u-tokyo.ac.jp, jinsong.zhang@nict.go.jp, gene.am91g@nctu.edu.tw, yrwang@cc.nctu.edu.tw, yfliao@ntut.edu.tw

あらまし 連続音声における標準中国語の声調は、複雑な特性を示す。このため、声調認識は未解決の課題として、多く研究されている。ここでは、２つの手法を統合することによって、声調認識の高度化を図る。２つの手法のうち１つ目は、声調核モデルで、当該音節の F0 パターンを特徴付ける部分に着目することにより、前後の音節への過渡部分の影響（声調結合）を押さえるものである。この際、声調核を自動抽出することが求められるが、それを高精度で行う手法も開発した。２つ目は、多層パーセプトロン（MLP）を声調認識に用いることにより、HMM では困難であった（F0 と持続時間といった）声調に関係する異種の特徴を利用することを可能とすることである。実験の結果、（軽声を含めた）声調認識誤りが、声調核モデルから得られる特徴を用いた HMM では 14.5%、声調核モデルを用いない MLP では 14.1%であるのに対し、手法を統合した場合、12.8%に減少した（10%の誤り軽減）。この結果は、手法の統合が声調認識の性能向上に有効であることを示すものである。

キーワード 標準中国語、声調認識、声調核モデル、声調核自動抽出、MLP

# Tone Recognition of Continuous Mandarin Speech Based on Tone Nucleus Model and Neural Network

Xiao-Dong Wang[I], Keikichi Hirose[I], Jin-Song Zhang[II], Nobuaki Minematsu[I], Chen-Yu Chiang[III], Yih-Ru Wang[III], Yuan-Fu Liao[IV]

The University of Tokyo 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-865 Japan
NiCT 2-2-2 Hikaridai "Keihanna Science City", Kyoto 619-0288 Japan
IIINational Chiao Tung University 1001 Ta Hsueh Road, Hsinchu, Taiwan 300, ROC
IVNational Taipei University of Technology 1,Sec. 3, Chung-Hsiao E. Rd. Taipei 106, Taiwan, ROC

Tel. 03-5841-6667

E-mail: I （wxd, hirose, minematsu）@gavo.t.u-tokyo.ac.jp, II jinsong.zhang@nict.go.jp, III gene.am91g@nctu.edu.tw, yrwang@cc.nctu.edu.tw, IV yfliao@ntut.edu.tw

**Abstract** Automatic tone recognition of Mandarin continuous speech is still far from a resolved task due to complex feature variations for the five tones. In this paper we propose to integrate two independent techniques together to challenge the task: The first one is the tone nucleus modeling that specifies a critical segment in the $F_0$ contour of a syllable to avoid influences from transitional $F_0$ loci. Here, a method of automatically detecting tone nucleus is crucial. We have developed such a method. The second one is to adopt MLP (Multi-layer Perceptron) as tone recognizer to accept heterogeneous features correlated to tones but difficult to be used in a HMM (Hidden Markov Model) tone recognizer. Experiments showed: When the two individual techniques were adopted independently, HMM-based tone nucleus modeling got 14.5% tone recognition error, and MLP recognizer got 14.1%. On the other hand, the integrated approach achieved a relative error reduction by about 10%, and got 12.8% tone recognition error. The improved performances indicate the proposal is at a right direction.

**Keyword** Mandarin, Tone Recognition, Tone Nucleus model, Automatic Tone Nucleus Extraction, MLP.

## 1. Introduction

Each Mandarin syllable is pronounced with a pitch tone to manifest a specific morpheme. There are four lexical tones (Tone 1 to Tone 4) and a neutral tone (Tone 5). The number of tonal syllables is about 1,300, and it is reduced to about 410 when tone discriminations are discarded [1]. Fundamental frequency (hereafter F0) is the major acoustic feature to manifest the four basic tones [1,2], and a short duration seems more likely to characterize a neutral tone [1]. Although F0 contours show rather stable patterns for isolate tones, those of continuous speech may vary significantly from those isolate tones [2,9]. This is because F0s are also used to convey other information simultaneously including high-level intonation structure like prosody phrasing, sentence foci and etc. besides lexical tones [2,3,9]. The nature of F0s is inherently interplayed.

Automatic tone recognition or classification is desirable for building Mandarin speech recognition and understanding systems: First, tone information helps to reduce the problem of homophone words during acoustic matching. Second, as sentence F0s are rather varied by underling tones, the tone information is necessary for automatically detecting other information like high-level intonation structures from the F0 contours. Such kind of information is important for understanding "rich information" transmitted from speakers [4].

Efforts on tone recognition have been continuously made along two lines in the past decades [5-8]: The first line is to check different F0 normalization methods to suppress effects of substantial F0 variations. Among a number of methods, Tone Nucleus model [7,8] defines sentential F0s as the concatenation of target points (two per tone), tone nuclei (F0 loci between two in-tone targets) and transition loci, and suggests using features of tone nuclei for tone recognition. Compared with other methods, it not only provides a clear linguistic meaning for the normalization process, but also shows explicit potentials for detecting intonation structure from the target points. Papers [7,8] reported tone recognition studies by HMM based tone nucleus models. The second line of efforts is to look for appropriate tonal acoustic models and recognizer. Previous studies have tried linear regression model, Gaussian mixture model, HMM, Neural network (NN) and etc. Compared with other methods like HMM, NN has a capability to exploit heterogeneous features including both continuous and categorical variables, and showed good tone recognition performances in [5, 6, 10].

To achieve better tone recognition performance, we propose that it is a good try to integrate the two above-mentioned efficient approaches. The reasons lie in that: the tone nucleus approaches in [7,8] used HMM as the basic tone models, but HMMs are difficult in incorporating other features like duration and category ones than frame synchronized ones. However, we know that short duration is an important characteristic of the neutral tone, and syllable positions in a sentence have significant influences on their F0s [2,9]. The low performances of the neutral tone in [7,8] might be improved by exploiting these features. For this purpose, NN is a good choice. On the other hand, simple F0 normalizations in the NN studies of [5,6] failed to provide a clear linguistic meaning, thus limited further improvements and application of the approach. If tone nucleus modeling can be successfully integrated with NN approach, the integration will become a potentially good solution to the problem of high-level prosody information detection.

This paper integrates the two approaches of tone nucleus modeling and MLP neural network for tone recognition: Section 2 introduces the Tone Nucleus model and the procedure of automatic tone nucleus extraction. Section 3 illustrates our MLP based tone recognizer. Section 4 presents experimental database, comparative baseline systems and experimental results. Section 5 discusses the experimental results and finally we conclude the paper in Section 6.

## 2. Tone Nucleus Modeling

### 2.1. Tone Nucleus Model

The Tone Nucleus model is a linguistic framework proposed in [7], which offers a possible systematic framework to deal with F0 variations that result from both articulatory constraints and confounded intonation function, for tone recognition and intonation function detection.

As illustrated in Fig. 1, a syllable F0 contour could be divided into three segments: onset course, tone nucleus and offset course. Tone nucleus is a piece of F0 contour that represents pitch targets of the lexical tone, which contains the most critical information for tonality perception. Onset course is the asymptotic F0 transition locus to the tone-onset target from a preceding vocal cords' vibration state. Offset course is the F0 transition locus from the tone-offset target to a succeeding vocal cords' vibration state.

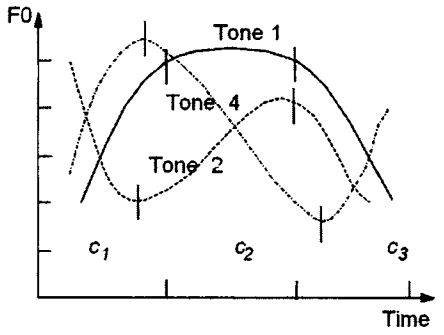


**Fig.1.** Illustration of tonal F0 contours with possible articulatory transitions for Tones 1, 2 and 4. The left and right vertical sticks in each contour correspond to the possible tone onset and offset locations, and the three F0 segments delimited by tone onset and offset are onset courses, tone nuclei and offset courses, depicted as $c_1$, $c_2$ and $c_3$.

## 2.2. Automatic Tone Nucleus Extraction

Tone nucleus modeling suggests to use features from tone nuclei for tone recognition [7,8]. For the purpose, tone nuclei must be located in sentential F0 contours before recognition. Fig.2 illustrates our algorithm for tone nucleus extraction, which is modified from the original one in [7].

Major steps of the algorithm in [7] include: First, segmentation initialization. The observation sequence, $\boldsymbol{O} = (o_1 o_2 \ldots o_N)$ calculated for the given syllable's voiced part, is divided into 3 segments with lengths of $n_1$, $n_2$ and $n_3$ frames. Then, the sub-sequence of segment $i$ is assumed to have the probability density function of the multivariate Gaussian $\Phi_i$, whose mean-vector $\mu_i$ and covariance matrix $\Sigma_i$ are given by:

$$\mu_i = \frac{1}{n_i}\sum_{k=1}^{n_i} o_k$$

$$\Sigma_i = \frac{1}{n_i}\sum_{k=1}^{n_i} (o_k - \mu_i)(o_k - \mu_i)'$$

$$(1 \le i \le 3)$$

Second, re-segmentation using the segmental K-means quantization algorithm [7] and re-estimation of $\Phi_i$ until convergence.

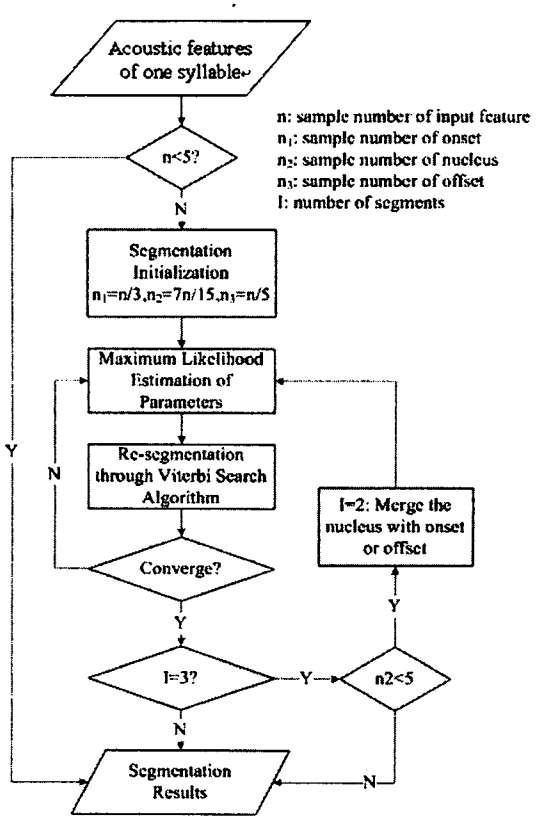


**Fig.2.** Illustration of tone nucleus extraction algorithm.

In this step, the likelihood

$$p(o_j|\Phi_i) = \frac{1}{2\pi|\Sigma_i|^{1/2}} \exp[-\frac{1}{2}(o_j - \mu_i)' \times \Sigma_i^{-1}(o_j - \mu_i)]$$

$$(1 \le j \le N)$$

is calculated for all observations in the Viterbi search to decide which segment the observation $o_j$ belongs to.

Third, check the resulting segmentation if it satisfies stop criterion. If the segment number is 3 and if the middle segment is shorter than 50ms, segmentation number is reduced to 2 and the second step is repeated.

Major modifications of current algorithm as compared to [7] are: First, autocorrelation coefficient computed during pitch tracking is incorporated together with delta logF0, power as the acoustic features. This is because tone nucleus is considered to be stable enough to reflect the tonality perception and autocorrelation coefficient here is one of the good indices to show the stability. Second, as reported in [9], the carry-over effect is much

larger than the anticipate effect. Therefore, the initial segmentation is changed from uniform ratio to current ratio, 5:7:3, to correspond to the unbalanced articulatory effects from previous and succeeding syllables to some extent. Third, we replaced the T-test based segments merge in [7] by a simple method: just let the Viterbi decoding allow state skip to realize segment merge. Fourth, when the resulting segment number is 2, we simply select the one as the nucleus that includes most points of the original middle candidate segment.

## 3. MLP Tone Recognition

We use an MLP neural network as our tone recognizer, as illustrated in Fig. 3. The MLP has 3 layers: the input layer has a varying number of nodes depending on the specific features used. The hidden layer was optimized to have 250 nodes according to the amount of training database of this study. The output layer consists of five nodes, each corresponding to one of the five tones. Both the hidden and output layers use a standard sigmoid function. The training algorithm for the MLP is the back propagation algorithm. During recognition, the node (of the output layer) with the highest score is accepted as the final answer.

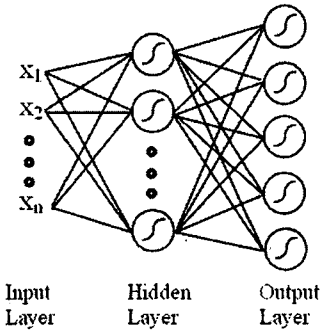


**Fig.3.** MLP structure for tone recognizer.

The acoustic features fed into the MLP recognizer are calculated from tone nuclei extracted beforehand. To make the features robust against F0 fluctuations, each tone nucleus is further divided into 3 uniform sub-segments to compute features of the following 7 groups:

1) Mean energy, mean logF0 and F0 slope of each sub-segment of the focused tone nucleus (a total of 9 parameters);
2) LogF0's of onset point, midpoint and offset point of the focused tone nucleus (3 parameters);
3) Duration of the syllabic F0 contour;
4) Mean energy, mean logF0 and F0 slope of the last sub-segment of the preceding tone, and those of the first sub-segment of the succeeding tone (a total of 6 parameters);
5) Four anchor features for the onset and the offset of the focused tone nucleus, as in [8].
6) Two transition durations: the duration from previous syllable's offset to current syllable's onset, and the duration from current syllable's offset to next syllable's onset.
7) Two binary flags indicating if the tone is located in the beginning or ending position in an utterance.

Totally, the feature vector includes 27 values.

## 4. Experiments and Results

### 4.1. Speech Database

Tone recognition experiments have been carried out on the data of one female speaker in corpus HKU96, published by Hong Kong University [7]. F0 was extracted by ESPS package without manual error correction. 500 utterances from cs0f0001 to cs0f0500 were used as training set, while 200 utterances from cs0f0501 to cs0f0700 were used as testing set. The number of syllables is 6,419 for training set and 2,567 for testing set. Phonetic segmentation of the corpus was checked and used as the prior information in the tone recognition experiments.

### 4.2. Two Baseline Comparative Systems

The basic scheme of NN approach in [6] was repeated to serve as the first baseline tone recognition system. The input feature vector has 20 values those in Groups 1, 3, 4, 6 and 7 described in Section 3. The calculation method is also different here by one point: they are computed from whole syllable features instead of tone nuclei. The second baseline system adopted one of our previous realization reported in [8]: anchor-based normalized F0 features appended to the standard feature vector (logF0, frame energy and their first, second order time derivatives). The tone models used continuous density HMMs. As the training and testing data were totally the same as our current approach, we directly used performances reported in [8] without repeating the tone recognition experiments.

### 4.3. Tone Nucleus Detection Results

As tone nuclei detection is not the major objective of

this study, the 1st author paid a loose visual inspection on the detection result. If the segmentation of a tone nucleus does not significantly depart from the supposed place, it is generally accepted as correct. The inspection was carried out on utterances labeled from cs0f0251-cs0f0300. Table 1 shows the inspection results totally and separately for the three cases: cases where 1 segment, 2 segments and 3 segments are kept after the tone nucleus extraction is section 2.2.

**Table 1.** Tone Nucleus detection performance for 50 utterances.

| | 1-seg. | 2-seg. | 3-seg. | Total |
|---|---|---|---|---|
| Number of tones | 14 | 119 | 533 | 666 |
| Nuclei accepted | 14 | 118 | 530 | 662 |
| Correct rate (%) | 100 | 99.2 | 99.4 | 99.4 |

### 4.4. Tone Recognition Results

Tone recognition performances are illustrated in Fig. 4 for total tone recognition errors for the two baseline systems and the proposed integrated approach. Tables 2, 3 and 4 give confusion matrices for the three systems for further detailed inspection.

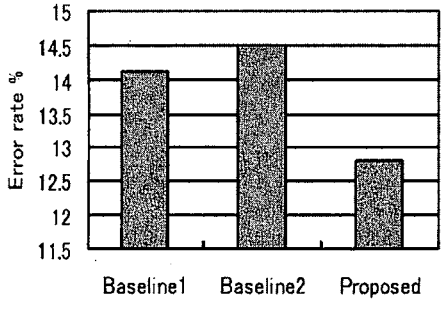


**Fig.4.** Tone recognition performances in error rates.

**Table 2.** Confusion matrix for 1st baseline NN system (in %).

| Tone | T1 | T2 | T3 | T4 | T5 |
|---|---|---|---|---|---|
| T1 | **74.2** | 8.0 | 0.2 | 16.7 | 0.9 |
| T2 | 6.1 | **87.2** | 1.7 | 4.4 | 0.6 |
| T3 | 0.9 | 5.0 | **81.3** | 11.1 | 1.8 |
| T4 | 2.9 | 1.3 | 1.4 | **93.5** | 0.9 |
| T5 | 4.4 | 3.8 | 5.1 | 13.9 | **72.8** |

**Table 3.** Confusion matrix for 2nd baseline HMM system (in %).

| Tone | T1 | T2 | T3 | T4 | T5 |
|---|---|---|---|---|---|
| T1 | **87.9** | 6.4 | 0 | 5.2 | 0.5 |
| T2 | 5.9 | **87.9** | 2.6 | 2.2 | 1.4 |
| T3 | 0.3 | 3.9 | **84.6** | 4.7 | 6.5 |
| T4 | 4.4 | 0.7 | 2.3 | **91.1** | 1.5 |
| T5 | 7.3 | 4.9 | 25.6 | 27.4 | **34.8** |

**Table 4.** Confusion matrix for the proposed system (in %).

| Tone | T1 | T2 | T3 | T4 | T5 |
|---|---|---|---|---|---|
| T1 | **75.7** | 6.34 | 0.6 | 15.6 | 1.7 |
| T2 | 4.4 | **88.9** | 2.1 | 3.6 | 1.1 |
| T3 | 0.6 | 3.5 | **84.8** | 9.3 | 1.2 |
| T4 | 1.4 | 0.9 | 2.1 | **94.8** | 0.8 |
| T5 | 3.2 | 2.5 | 10.8 | 15.8 | **67.7** |

## 5. Discussions

Comparing total tone recognition errors of the two baseline systems with that of the proposed integrated approach in Fig. 4, we can see:

1) The proposed approach achieved an absolute error reduction of 1.3% as compared to the 1st baseline of NN approach, equal to a relative error reduction of 9.2%. The difference of the two systems lies in whether to use full syllable or tone nucleus to calculate the features. The better performance indicates that tone nuclei do keep important and robust discriminating features for the tones.

2) The proposed approach got an absolute error reduction of 1.7% as compared to the 2nd baseline of HMM tone nucleus method, and the corresponding relative improvement was 11.7%. The improvement can be attributed to the use of NN and the two additional features: segmental durations and syllable positions in the sentence, which are difficult to be exploited in an HMM based approach.

Inspections into the confusion matrices in Tables 2, 3 and 4 may reveal more detailed differences among the three systems.

3) Comparison between Tables 2 and 4 indicates that the four basic lexical tones, Tone 1 to Tone 4, are all improved by the tone nucleus modeling. On the contrary, the neutral tone (Tone 5) reduced its accuracy significantly from 72.8% to 67.7% by the tone nucleus modeling. Reason for this might be ascribed to the fact that tone nuclei showed significantly shorter durations than the full syllables. This will increase the confusions between the four basic tones and the neutral tone, as the short duration is regarded as the key discriminating features for the neutral tone.

4) Comparison between Tables 3 and 4 indicates: First, the neutral tone recognition is improved significantly from 34.8% to 67.7% (or, 72.8% in Table 2), indicating the effectiveness of duration feature for discriminating the neutral tone. Second, Tones 2, 3 and 4 got similar performances or improvements in the NN approach as compared to the HMM approach.

5) However, in Tables 3 and 4, Tone 1 got significant 12.2% degradation in the NN approach as compared to the HMM approach (75.7% vs. 87.9%). Although errors of recognizing Tone 1 as Tones 2, 3 and 5 keep the similar levels (differences are almost less than 1%). The error of Tone 1 as Tone 4 increased significantly from 5.2% of the HMM approach to 15.6% of the NN approach. Reasons for this degradation can be regarded as: First, Tone 1 and Tone 4 both have high pitch targets, and are tended to be confused [7,8,9]. Second, the database has a biased high frequency of Tone 4 (about 45%) due to the pronunciations of punctuation marks at the end of utterances, which are almost all Tone 4s. So it can be expected that there exists a high correlation between Tone 4, and pre-syllable pauses and syllable positions. Incorporation of such additional information might make the NN approach prefer more Tone 4 decisions. As a result, Tone 1 samples are more likely to be misrecognized as Tone 4s. Simultaneously, we can see that the improvement of Tone 4 in Table 4 mostly come from reduced errors of Tone 4 as Tone 1 in Table 3 (4.4% to 1.4%).

6) The direct solution to the problem in 5) is to use more unbiased training data to build the recognizer.

7) We regard the successful integrated approach as our start point for future efforts on Mandarin tone recognition and prosody information detection. The next step is to adopt the advanced multi-level prosody modeling based on tone nucleus model, as proposed in the study [8].

## 6. Conclusion

This paper presents our new approach on the tone recognition of Mandarin continuous speech by integrating two effective methods: tone nucleus modeling for F0 feature normalization and neutral network tone recognizer which exploited heterogeneous acoustic features including continuous and categorical variables. The improved performances showed the effectiveness of the proposal. In the future, we will continue our efforts to extend this study in the aspects of: 1) Porting the system to larger speaker independent databases. 2) Investigating the detection of high-level intonation structure based on the tone nucleus modeling.

## REFERENCE


[1] Y.-R. Chao, *A grammar of spoken Chinese*, University of California Press, Berkeley, 1968.

[2] Ch.-L. Shih,, "Tone and intonation in Mandarin", Working Papers, Cornell Phonetics Laboratory, pp.83-109, 1988.

[3] H. Fujisaki, "Prosody, Models, and Spontaneous Speech", in Y. Sagisaka et al, eds., Computing Prosody: computational models for processing spontaneous speech, Springer-Verlag, New York, pp.27-42, 1997.

[4] H, Niemann et al, "Using Prosodic Cues in Spoken Dialog Systems", Y.Kosarev (ed.), SPECOM'98 Workshop, ST-Petersburg, pp.17-28, 1998.

[5] T. Lee, P.C. Ching, L.W. Chan, Y.H. Cheng, and B. Mak, "Tone recognition of isolated Cantonese syllables", IEEE Trans. SAP, vol. 3, no. 3, pp. 204-209, 1995.

[6] S.-H. Chen and Y.-R. Wang, "Tone Recognition of Continuous Mandarin Speech Based on Neural Networks", IEEE Trans. SAP, vol. 3, no. 2, pp. 146-150, 1995.

[7] J.-S. Zhang and K. Hirose, "Tone Nucleus Modeling for Chinese Lexical Tone Recognition", Speech Communication, vol. 42, no. 4, pp. 447-466, 2004.

[8] J.-S. Zhang, S. Nakamura, and K. Hirose, "Tone Nucleus-based Multi-level Robust Acoustic Tonal Modeling of Sentential F0 Variations for Chinese Continuous Speech Tone Recognition", Speech Communication, vol. 46, no. 4, pp. 440-454, 2005.

[9] Y. Xu, "Contextual Tonal Variations in Mandarin", J. Phonetics, vol. 25, no. 1, pp. 61-83, 1997.

[10] K. Hirose, H. Hu, X. Wang, and N. Minematsu, "Tone recognition of continuous speech of Standard Chinese using neural network and tone nucleus model," Proc. Interspeech 2006 - ICSLP, Pittsburg, Thu1FoP.10, pp.2394-2397, 2006.