音声言語情報処理 29－4
(1999. 12. 20)

逐次クラスタリングと線形判別分析による中国語声調核の検出

張勁松、広瀬啓吉

東京大学工学系研究科電子情報工学専門課程

東京都文京区本郷 7-3-1

zjs@gavo.t.u-tokyo.ac.jp, hirose@gavo.t.u-tokyo.ac.jp

あらまし　我々は以前、音節の基本周波数(F0)パターンのうち、声調核が声調の識別に有効な情報を有する部分であることを示した。ここでは、連続音声の音節F0 パターンに対して、声調核を検出する試みに付いて紹介する。開発した手法では、まず音節F0 パターンを逐次クラスタリングによって数個(3 個以下)のF0 線分に分割する。次に、線形判別に基づく予測器によって、どの線分が声調核に対応するかをけいてする。検出された声調核に着目して連続音声の声調認識を行ったところ、提案手法の有効性が確認された。

キーワード 声調核, 逐次クラスタリング、線形区別分析 ,仮説検定、声調認識.

CHINESE TONE NUCLEI DETECTION BASED ON SEQUENTIAL CLUSTERRING AND LINEAR DISCRIMINANT ANALYSIS

Jinsong Zhang, Keikichi Hirose

Department of Information and Communication Engineering

School of Engineering, University of Tokyo

Bunkyo-ku, Tokyo,113-8656, Japan

zjs@gavo.t.u-tokyo.ac.jp, hirose@gavo.t.u-tokyo.ac.jp

Abstract　We have suggested that tone-nucleus of a syllable fundamental frequency (F0) contour contains the information for tone discrimination. This paper introduces our approach toward the robust method to detect tone-nucleus for each syllable in continuous speech. The developed method first segments a syllable F0 contour into several (3 or less) linear F0 loci by the sequential clustering algorithm, and then decides the one corresponding to the tone-nucleus using a predictor based on linear discriminant analysis technique. Tone recognition of continuous speech using the detected tone-nuclei showed the effectiveness of the developed method.

key words　Tone-nucleus, Sequential clustering, Linear discriminant analysis, Hypothesis test, Tone recognition.

1 Introduction

Robust tone recognition is desirable in developing Chinese speech recognition and understanding systems. The first reason is that pitch tones play an important phonemic role in Chinese: each syllable corresponds to a morpheme (ideographic character) and is associated with a pitch tone(usually referred to as a lexical tone). Syllables with different lexical tones may have different meanings even if they own the same segmental structure. Chinese speech recognition systems must be capable of dealing with the phonemic role of the lexical tones. The second reason is that tone recognition is necessary for automatic detection of intonation structure such as prosodic phrase boundaries, foci, stress locations, etc. This kind of information is helpful to deal with problems such as ungrammatical utterances, disfluencies, syntax ambiguities, dialogue topic detection, to develop spontaneous spoken dialogue systems.

There are four basic lexical tones (referred to as Tones 1, 2, 3, 4, respectively) and a neutral tone (referred to as Tone 5) in Mandarin Chinese. The four basic lexical tones are characterized by their perceptually distinctive pitch patterns. According to the different tonality perception, linguists usually call them: the high-level, mid-rising, low-dipping and high-falling tones [1, pp.25-26.]. The neutral tone is said to have no specific pitch pattern, and is highly dependent on the preceding tone and usually perceived to be short and light [1, pp.35-36]. Fundamental frequency (hence F0) contours are the main acoustic correlates with pitch pattern, thus the most important features for tone recognition.

Previous studies have shown that it is not difficult to recognize isolated lexical tones, but it is very difficult to realize robust tone recognition for continuous speech [2, 3]. This was ascribed to the fact that lexical tones show quite consistent F0 patterns when uttered in isolation, but show complex variations in continuous speech[3]. The complex F0 variations originate from the mechanic-physiological realization of the compound intonation functions. On the one hand, F0 contours are the main acoustic manifestations of not only the lexical tones, but also other kinds of intonation functions including prosodic phrase grouping, foci, sentential pragmatics and etc. Such confounded nature of information in F0 contours may obscure the F0 variations cueing lexical tones. For example, since focus is usually related to F0-range-expansion of non-final focused words and F0-range-suppression of post-focus words, F0 contour of a Tone 1 which usually shows high and flat pattern may change into a rising or falling shape [4]. On the other hand, F0 contours, which reflect the periods of the successive human vocal cords' vibrations, are to vary due to articulatory constraints. For example, voicing or unvoicing in syllable initial segments may lead to quite different syllable F0 contours even when the lexical tones are the same [5]. The inertial characteristic of the bio-mechanical vibrations makes the neighboring tones interfere with each other extensively so that early portion of a syllable F0 contour tends to vary according to the carryover effect of the preceding tone[4].

Although the substantial F0 variations confuse the automatic tone recognizer very much, they were found to exert very few influences on tonality perception by human beings. Phonetic studies ascribed this to the complex masking effects of speech-specific mode of perception[6]: F0 variations can be classified into Intrinsic and Extrinsic F0 loci, where Intrinsic F0 loci are the F0 variations that are non-deliberately produced, but occurring as the transition loci to the deliberate F0 targets, and exert no influences on human pitch perception. Extrinsic F0 loci are the deliberately produced target F0 and serve as the acoustic cues for pitch perception. Other studies have shown that the possible intrinsic transition F0 loci may appear in the early portion [5, 6, 7, 8] and in the ending portion of a syllable F0 contour[7]. Furthermore, F0 observations have showed that the intrinsic transition may occupy a large portion of the syllable F0 contour: 30% in [5], 50-100 msec after the release of the initial consonant in [6, p. 76] (accounting for about 40% proportion in the given example[p.65]), and more than 50% when the rising tone starting from the middle of the rhyme [9]. Since conventional tone recognizers based on hidden Markov Modes (HMMs) or neural net usually take whole syllable F0 contours into accounts as important acoustic cues for the lexical tones[2, 3], they can not discriminate the possible intrinsic and extrinsic F0 loci. And this is the probable reason for the low tone recognition accuracy of tone recognizers as compared with human perception.

In view of the above considerations, we have recently proposed a new scheme to deal with the possible intrinsic F0 loci in a syllable F0 contour for tone recognition. We proposed an "F0 segmental structure model" of Chinese syllable F0 contours, in which each syllable F0 contour consists of three sequential segments: onset-course, tone-nucleus and offset-course[12]. Tone-nucleus corresponds to the deliberate tone targets and is the extrinsic F0 loci, and the other two are possible intrinsic F0 transition loci. Among the three segments, only tone-nucleus is obligatory and the other two are optional. Tone recognition based on acoustic features of tone nuclei reported about 5% tone accuracy improvement form the conventional one observing full syllable features through a speaker-independent disyllable words task [10] and a speaker-dependent continuous speech task [11]. Furthermore, since tone-nuclei correspond to deliberate F0 targets, they may serve as more robust cues than the full syllable F0 contour for the higher-level intonation functions like prosodic phrase boundaries and foci. Our further study showed that it was easier for human labelers to label possible tone coarticulation effects by focusing on the tone-nuclei[13].

One important prerequisite for these approaches is that tone-nuclei need to be located in the syllable F0 contours first. In the earlier period, we tried to use phoneme recognizer's output to locate tone-nuclei[10]. Although it worked for the disyllable words task, it showed some problems when applied to continuous speech task, such as a time-consuming work of hand labeling and segmentation errors. Instead, we developed a new method to automatically locate tone-nuclei in continuous speech [11], which only uses F0 and power features together with phoneme and syllable segmentation information which are assumed to be available from conventional acoustic recognizer's output. The new method consists of three steps: the first step is F0 contour segmentation based on sequential clustering algorithm, the second step is tone-nucleus prediction based on linear discriminant analysis of a number of acoustic features, and the third step is tone-nucleus smoothing.

This report gives a detailed description of the pro-

posed tone-nuclei detection method, and the rest is organized as: Section 2 for the speaker-dependent speech database used in this study, Section 3 for the proposed tone-nucleus F0 segmental structure model of Chinese syllable F0 contours, Section 4 for the proposed tone-nuclei detection method, Section 5 for the evaluation of the effectiveness of the proposed tone-nuclei detection method through tone recognition of continuous speech. Finally we conclude in the section 6.

2 SPEECH DATABASE

Female speaker (0f) utterances in the speech corpus HKU96, published by Hong Kong University, were used in the study. There are several reasons for selecting only one speaker's data: the first one is that the problem dealt with is on the intra-speaker phonetic variations. Testing using a speaker-dependent database is necessary to prevent co-influences of inter-speaker phonetic variations. The second one is that the speaker 0f was a phonetician who designed and was fully involved in developing the data corpus HKU96. Her speech was much more natural than those of other speakers, which often sounded more like a "Read" style. For two reasons of easily manageable data size and enough data for statistical training, we took 500 utterances labeled from cs0f0001 to cs0f0500 as training set(6419 syllables), and 200 utterances labeled from cs0f0501 to cs0f700 as testing set(2567 syllables). Average utterance length is 12.8 syllables in both the training and testing sets. The utterances have average speech rates ranging from 3.8 to 4.9 per a second, with local rates sometimes exceeding 5 syllables per a second.

The corpus offers phoneme, syllable and lexical tone labels together with orthographic transcriptions. All the labels were manually checked to correct possible errors. Spectrograms, and sometimes together with F0 contours, were referred to align segmental units. Phonological rules were guided manual modifications of tone labels according to their real tonalities. For example, according to the Tone 3-Tone 3 tone sandhi rule, the first one often changes to Tone 2. According to some other morphophonemic tone sandhi rules[1], syllables like "yi" and "bu" have tonal alternations depending on its following tones. F0 was extracted using the integrated F0 tracking algorithm (IFTA) [14] with manual error correction. The window size for log-energy analysis was 20ms. Frame shift for both F0 and log-energy was 10ms.

3 TONE NUCLEUS MODEL

Tone Nucleus model [12], as illustrated in Fig. 1, suggests that a syllable F0 contour may consist of three segments: onset course, tone-nucleus and offset course. Among the three segments, only the tone-nucleus is obligatory, whereas the other two are intrinsic F0 transition loci and their appearances are optional.

- Tone-nucleus: the segment contains the most critical information for tonality perception. The beginning and ending points of a tone-nucleus correspond to the Tone-onset and Tone-offset.
- Onset Course: the asymptotic F0 transition locus to the Tone-onset target from a preceding vocal cord vibration state.
- Offset Course: the F0 transition locus from the Tone-offset target to a succeeding vocal cord vibration state.

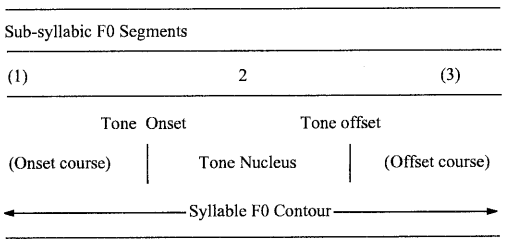

Figure 1: Illustration of the proposed F0 segmental structure model of Chinese syllable F0 contours. Optional F0 segments and their indices are indicated by parentheses; only the tone-nucleus is obligatory.

Pitch values like low(L) or high(H) at tone onset and offset are the distinctive targets characterizing the four basic lexical tones [Table 1]. When lexical tones are uttered in isolation, their F0 contours usually conform very well to the perceptual pitch patterns, i.e. L pitch is associated with low F0 values and H pitch is associated with high F0 values. Thus Tone 1 has a high and flat F0 contour in the speaker's F0 range, Tone 2 has a rising F0 contour, Tone 3 has a low dipping F0 contour, and Tone 4 has a high falling F0 contour.

targets	Tone 1	Tone 2	Tone 3	Tone 4
Onset	H	L	L	H
Offset	H	H	L	L

Table 1: Pitch targets of the four lexical tones. "H","L" depict high and low targets respectively.

Due to inertial characteristics of the vocal cord vibration system, each segment in the Tone-nucleus model usually shows a linear F0 locus in a syllable F0 contour. And one syllable F0 contour can be represented by no more than three linear F0 loci. Figure 2 illustrates some frequently observable syllable F0 contours in continuous speech, and the segmentation based on the Tone-nucleus model. We can see that only tone-nuclei delimited by the vertical sticks keep the typical patterns of their associated tones, whereas other segments deviate. Tone recognizers can be prevented from the influences of those intrinsic transition loci by focusing on the tone-nuclei, if they can be located beforehand.

4 TONE-NUCLEUS DETECTION

For a given syllable F0 contour, in order to detect possible tone-nucleus, it should be first segmented into concatenation of linear F0 loci where numbers of F0 loci are unknown beforehand. Secondly we need to decide which segment is the possible tone-nucleus. There are some difficulties to do this: one is how to robustly segment the

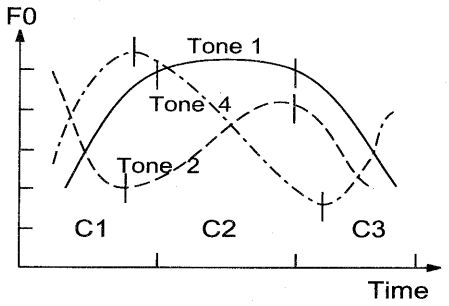

Figure 2: Illustration of syllable F0 contours with possible intrinsic transition loci for Tone 1, Tone 2 and Tone 4. The left and right vertical sticks in each contour correspond to the possible tone onset and offset F0 values, and the medium F0 segment delimited by the tone onset and offset in each contour represents tone-nucleus of the tone. C1, C2 and C3 depict onset courses, tone-nuclei and offset courses. Besides, Tone 3 has much more complicated variations than other three tones, and is not dealt here.

syllable F0 contours into sequential linear F0 loci. Segmentation method based on turning points like peak and valley points is not a good choice since there are possible extr-local peaks and valley points in F0 contours due to fluctuations. Also it is difficult to find a threshold to segment an F0 contour like that of Tone 1 in Figure 2, since the slope variation is gradual. Furthermore, number of linear F0 loci in a syllable F0 contour is uncertain; maybe one, two or three. The second difficulty is that we have little knowledge about acoustic distributional characteristics that may be used to detect tone-nuclei, albeit there were some related qualitative descriptions in some phonetic studies.

Hence, we proposed to segment syllable F0 contours by sequential clustering algorithm, and to amalgamate neighboring segments according whether they can stand Hypothesis test on equal means of F0 slope ratios. In order to understand acoustic dirstributional characteristics of tone-nuclei, we collected a number of acoustic features, such as time, energy and F0 related variables, to carry out variance analyses. A tone-nucleus predictor was finally developed via linear discriminant analysis (LDA) which could select the smallest subset of variables for the predictor with the minimum prediction error.

4.1 Syllable F0 Contour Segmentation Via Sequential Clustering

Sequential Clustering is a variant on the well-known K-means iterative procedure for clustering data. The variation lies in that the re-assignment of samples to the clusters is achieved by finding the optimum segmental state sequence, via the Viterbi algorithm, and then by backtracking along the optimal path. The procedure is illustrated in Figure 3.

Let $x_1, x_2, \cdots, x_j, \cdots, x_N$ denote the frame points of a syllable F0 contour. The jth point x_j has a two dimensional feature $(\log F0_j, \Delta \log F0_j)$. Each segment cluster contains $LEN_i, 1 \leq i \leq 3$, sequential frame points, and is represented by a centroid with normal distribution $N(M_i, \Sigma_i), 1 \leq i \leq 3$. Viterbi search is based on the minimum distance rule, where Mahalanobis distance is used to measure the distance between a frame point x_j and a segment centroid C_i.

$$dis(x_j, N(M_i, \Sigma_i)) = (x_j - M_i)^t \Sigma_i^{-1} (x_j - M_i)$$
$$1 \leq i \leq 3. \text{ j=1,2,}\cdots\text{,N.}$$

There are two principles to decide whether neighboring F0 segments should amalgamate or not: one is a phonetic principle and the other one is based on Hypothesis test on the equal means of the slope ratios of the two segments. Phonetic studies have revealed a duration threshold of 40-60 ms is necessary for a F0 contour to be perceivable[6, pp.70], thus it is reasonable to suggest that length of a tone-nucleus should be longer than 50 ms, corresponding 5 frames when shift of F0 being 10 ms. If after segmentation a syllable F0 contour consists of 3 segments, then the medium segment is corresponding to tone-nucleus according to the Tone Nucleus model. Its length is required to be longer than 5 frames.

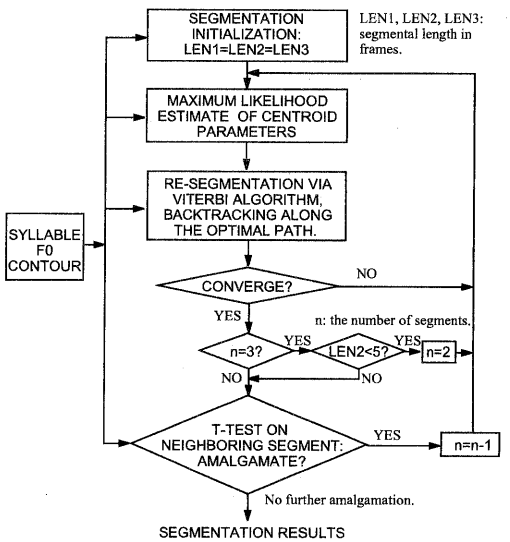

Figure 3: The sequential clustering procedure used to segment a syllable F0 contour into a possible number of linear F0 loci.

Estimator of slope ratio K_i of a F0 segment C_i is equal to the maximum likelihood estimate of the $\Delta \log F0$ of the segment centroid:

$$\hat{K}_i = \frac{1}{LEN_i} \sum_{j=1}^{LEN_i} \Delta \log F0_j = \frac{1}{LEN_i} (\log F0_{LEN_i} - \log F0_0)$$
$$x_j \in C_i \text{ and } x_0 \in C_{i-1}$$

After a syllable F0 contour is segmented into $n, 1 \leq n \leq 3$, segments, we will try to test if any two neighboring segments have the equal means of slope ratio, i.e.

$$H_0 : K_i = K_{i+1}$$
$$H_1 : K_i \neq K_{i+1}$$

by using a test statistic $T_{i,i+1}$,

$$T_{i,i+1} = \frac{\hat{K}_i - \hat{K}_{i+1}}{\sqrt{MSE(\frac{1}{LEN_i} + \frac{1}{LEN_{i+1}})}}$$

where

$$MSE = \max \left\{ \begin{array}{l} \frac{\sum_{i=1}^{n} \sum_{j=1}^{LEN_i} (\Delta \log F0_{ij} - \hat{K}_i)^2}{N-n} \\ MSE_0 \end{array} \right.$$

MSE_0 is the prescribed minimum allowance for micro-fluctuations.

$T_{i,i+1}$ has a Student T distribution with N-n degree of freedom. The critical point for the two-tailed test at α level of significance is

$$cp = t_{N-n,1-\alpha/2} \sqrt{MSE(\frac{1}{LEN_i} + \frac{1}{LEN_{i+1}})}$$

We reject H_0 whenever $|\hat{K}_i - \hat{K}_{i+1}|$ exceeds the computed critical point. If H_0 cannot be rejected at α level, this means the two neighboring F0 loci have possible equal means of slope ratio, then the two segments will amalgamate. A syllable F0 contour may be segmented into one, two or three segments, among them the tone-nucleus is decided. For one-segment result, the whole syllable F0 contour is the tone-nucleus. For three-segment result, the medium segment is the tone-nucleus. For two-segment result, the tone-nucleus is predicted via linear discriminant analysis.

4.2 Tone-nuclei Prediction Via Linear Discriminant Analysis

We used group index "-1" or "1" to indicate that either the former or the latter segment of a two-segment F0 contour is the possible tone-nucleus. In discriminant analysis, a linear combination of a number of independent variables is formed and serves as the basis for assigning cases to groups. The linear discriminant equation is like,

$$D = B_0 + B_1 X_1 + B_2 X_2 + \cdots + B_p X_p$$

where the X's are the values of the independent variables and the B's are coefficients estimated from the training data. D is called the discriminant score. For a two-segment F0 contour, it is assigned to group "-1" or "1" depending on the discriminant score being closer to "-1" or "1". The B's are chosen during training period so that the discriminant scores differ as much as possible between the two groups, i.e., the ratio

$$\frac{\text{between-group sum of squares}}{\text{within-group sum of squares}}$$

is a maximum.

We collected a number of acoustic features from the speech wave including time, energy and F0 related features for the discriminant analysis. Among a large set of features, the following ones (refer to Figure 4) show to be important for characterizing tone-nuclei from other intrinsic F0 loci:

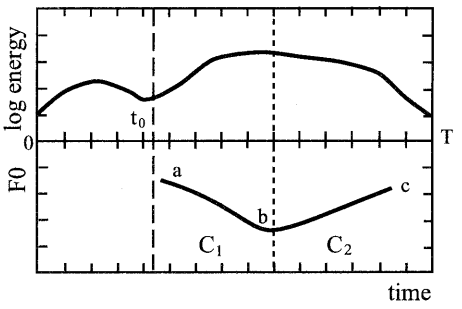

Figure 4: Illustration of a syllable F0 and energy contours. The broken line indicates the C-V boundary at t_0, the dotted line indicates the critical point b dividing the F0 contour into two segments C_1 and C_2, a and c indicate the starting and ending voicing points delimiting the syllable F0 contour.

- Normalized temporal location of point a in Figure 4 with respect to syllable onset: $t_{a0} = t_a/T$.
- Normalized temporal location of point a to Consonant-Vowel (CV) boundary: $t_{a1} = \frac{t_a - t_0}{T - t_0}$.
- Normalized temporal location of point b to syllable onset: $t_{b0} = t_b/T$.
- Normalized temporal location of point b to CV boundary: $t_{b1} = \frac{t_b - t_0}{T - t_0}$.
- Normalized temporal location of point c to CV boundary: $t_{c1} = \frac{t_c - t_0}{T - t_0}$.
- Duration of C_1 in frames: LEN_1.
- Duration of C_2 in frames: LEN_2.
- Normalized duration of C_1: $dur_1 = t_{b1} - t_{a1}$.
- Normalized duration of C_2: $dur_2 = t_{c1} - t_{b1}$.
- Energy slope ratio of C_1: $\Delta \log P_{C1} = \frac{\log P_b - \log P_a}{LEN_1 - 1}$.
- Energy slope ratio of C_2: $\Delta \log P_{C2} = \frac{\log P_c - \log P_b}{LEN_2 - 1}$.
- Segmental energy-sum ratio: $\xi = \frac{\text{energy sum of } C_2}{\text{energy sum of } C_1} = \frac{\sum_{j=b}^{c} \log P_j}{\sum_{k=a}^{b} \log P_k}$

where t_a, t_b, and t_c depict time locations of points a, b and c in Figure 4, P_a, P_b and P_c depict frame energy at points a, b and c.

We took a half of 500 training utterances, i.e. utterances labeled from cs0f0001 to cs0f0250, as the data for linear discriminant analysis. After F0 contour segmentation, we got 289 cases for group "-1" and 375 cases for group "1". Table 2 give the statistics such as Group-means (GM), Group-standard-deviations(GSD), univariate F-ratio(F) and Significance(S) for the various features.

From Table 2, we can see nearly all the variables except t_{a0} and t_{a1} have different distributions for the two groups "-1" and "1" above 0.999 probability. We

variables	GM		GSD		F	S
	-1	1	-1	1		
t_{a0}	.350	.315	.121	.132	12.0	0.001
t_{a1}	-.011	-.038	.133	.187	4.3	.039
t_{b0}	.756	.487	.092	.150	723.3	.000
t_{b1}	.617	.230	.141	.185	873.5	.000
t_{c1}	.834	.765	.120	.207	25.2	.000
$LEN1$	10.2	5.0	3.354	1.813	647.0	.000
$LEN2$	3.0	8.5	1.670	3.168	718.9	.000
dur_1	.705	.336	.161	.166	825.3	.000
dur_2	.291	.602	.118	.200	549.9	.000
$\Delta \log P_{C1}$	-.046	.069	.077	.134	171.0	.000
$\Delta \log P_{C2}$	-.304	-.085	.185	.101	373.5	.000
ξ	.272	2.365	.241	1.547	519.4	.000

Table 2: A number of statistics for group "-1" and "1" computed from 664 cases in LDA training data.

also can see the temporal location of point b is of the highest importance for discriminating the two groups. And t_{b1}, the normalized temporal location to CV boundary, is a better discriminating feature than t_{b0} which is normalized to the syllable onset. It is also interesting to note that if the former F0 segment (C_1) is tone-nucleus, it tends to have a descending energy slope ($\Delta \log P_{C1} = -.046$), whereas when the latter F0 segment (C_2) is tone-nucleus, the former segment (C_1) tends to have a rising energy slope ($\Delta \log P_{C1} = .069$). Since the variables may have correlations which will affect their effectiveness for discrimination, the well-known variable selection procedure, Stepwise Selection, is used to find the best subset of variables for group separation. In our experiments, a final subset consisting of 5 variables was achieved: dur_1, dur_2, $\Delta \log P_{C1}$, $\Delta \log P_{C2}$, ξ. The trained tone-nucleus predictor had a prediction accuracy of 95.18% for the 664 cases in the LDA training data.

After a syllable F0 contour is finally segmented into possible tone-nucleus and transition loci, it will be smoothed to keep only the tone-nucleus for tone recognition. A fraction of transition locus near the beginning or ending of the tone-nucleus is also kept for computation of F0 dynamic features.

5 TONE RECOGNITION

The effect of tone-nuclei detection may be estimated through application of detected tone-nuclei to lexical tone recognition. For tone recognition, we use CHMMs as tonal acoustic models. The number of tone HMMs is 5: among them 4 HMMs are for the 4 basic tones, and 1 for the neutral tone. Number of states is 5 for the four basic tones and 3 for the neutral tone. Mixture number is 6 for middle states and 2 for edging states in the case of 4 basic tones, and less number of mixtures for the neutral tone. Frame acoustic vector consists of 6 elements:

$$\log F0_i, \Delta \log F0_i, \Delta\Delta \log F0_i, \log p_i, \Delta \log p_i, \Delta\Delta \log p_i$$

The conventional method observing full syllable features acted as the baseline system for estimating the performance of our method based on ton-nuclei. Table 3 give the experimental results of the two methods.

tries	Tone1	Tone2	Tone3	Tone4	Tone5	Total
Base	69.2%	76.4%	70.0%	85.3%	30.5%	75.3%
New	83.6%	84.5%	68.0%	90.7%	31.7%	81.5%

Table 3: Tone recognition experimental results in percentages for the test dataset. "Base" and "New" represent the baseline system and the tone-nucleus approach.

6 CONCLUSION

From the tone recognition results, we can see that the proposed method increased by about 6% tone accuracy compared with the conventional one. Thus we may conclude that tone recognition based on tone-nuclei is a better method than the conventional one observing full syllable features, and that the proposed tone-nuclei detection method worked appropriately. We hope to extend this study to a larger database and speaker-independent cases.

References

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

[2] H.M. Wang et al, "Complete recognition of continuous Mandarin speech for Chinese language with very large vocabulary using limited training data", IEEE Trans. on SAP, 5, No.2, 1997, pp. 195-200.

[3] S. Chen, Y. Wang, "Tone recognition of continuous Mandarin speech based on Neural Networks", IEEE Trans. on SAP, 3, No. 2, 1995, pp.146-150.

[4] Y. Xu, "Effects of tone and focus on the formation and alignment of F0 contours", Journal of Phonetics, Vol. 27, 1999, pp.55-105.

[5] J. M. Howie, "On the domain of tone in Mandarin", Phonetica 30, 1974, pp.129-148.

[6] P. J. Rose, "On the non-equivalence of Fundamental frequency and pitch in tonal description", D. Bradley et al eds., Prosodic analysis and Asian linguistics, C-104, 1988, pp. 55-82.

[7] M.-C. Lin, "A perceptual study on the domain of tones in Beijing Mandarin", China ACTA Acustica, Vol.20, No.6, Nov. 1995, pp.437-445.

[8] Y. Xu, Q. Wang, "Pitch targets and their realization: Evidence from Chinese", forthcoming.

[9] Ch.-L. Shih, 88 "Tone and intonation in Mandarin", Working Papers, Cornell Phonetics Laboratory, Vol. 3, pp.83-109.

[10] J.-S. Zhang, K. Hirose, "A robust tone recognition method of Chinese based on subsyllabic F0 contours", Proc. of ICSLP98, 1998. pp. 703-706.

[11] K. Hirose and J.-S. Zhang, "Tone recognition of Chinese continuous speech using tone critical segments", Proc. of Eurospeech'99, 1999,pp.879-882.

[12] J.-S. Zhang, G. Kawai and K. Hirose, "Subsyllabic tone units for reducing physiological effects in automatic tone recognition for connected Mandarin Chinese", Proc. of ICPhS99, 1999, pp. 2367-2370.

[13] J.-S. Zhang, H. Kawanami, "Modeling carryover and anticipation effects for Chinese tone recognition", Proc. of Eurospeech'99,1999, pp.747-750.

[14] B.G. Secrest, and G.R. Doddington, "An integrated pitch tracking algorithm for speech systems", ICASSP, 1983, pp.1352-1355.