A STUDY ON ROBUST SEGMENTATION AND LOCATION OF TONE NUCLEI IN CHINESE CONTINUOUS SPEECH

Jin-Song Zhang[†] *and Keikichi Hirose*[‡]

[†] ATR Spoken Language Translation Research Laboratories
2-2-2 Kansai Science City, Kyoto 619-0288 Japan
[‡] Dept. of Frontier Informatics, School of Frontier Sciences, University of Tokyo
Bunkyo-ku, Tokyo, 113-0033, Japan
{jinsong.zhang@atr.co.jp, hirose@gavo.t.u-tokyo.ac.jp

ABSTRACT

Tone nuclei in continuous speech are regarded as efficient targets for either tone recognition or intonation function decomposition. This paper presents our statistically robust method to segment and locate tone nuclei in continuous speech. The method includes: an iterative segmental K-means segmentation of the tonal F0 contours, which is further aided with T-Test based segment almalgamation. And a linear discriminant function based tone nucleus discriminator, whose features are selected by the sequential feature selection method. The developed system achieved 97.5% tone nuclei correct rate on a speaker dependent task. The tone recognizer based on the detected tone nuclei improved tone recognition rate by more than 6% than the baseline ones using the full tonal syllable features.

1. INTRODUCTION

With the increase of the naturalness requirement, the automatic speech recognition and understanding (ASRU) systems are now never confined to providing a simple text transcription of the input speech, but also desired to be able to provide prosody information on such as speaker's intentions, attitudes, pragmatics, discourse structure and even speaker emotional status [1]. These kinds of information are known mainly acoustically realized by the fundamental frequency (F0) [2]. However, studies have shown there are rather complex variations in F0 contours, making the decomposition of prosody information from them very difficult. On the one hand, various kinds of information sources are conveyed by the only one-dimensional F0 feature, hence *inherently confounding* [3]. On the other hand, F0 contours, which reflect the periodicals of the successive human vocal cords' vibrations, are to vary due to *articulatory constraints*.

In the case of Chinese, F0 contours become more complex due to the phonemic role of pitch tones. There are four basic tonal F0 patterns, with each the same syllable become different morphemes or different words. An efficient way to squeeze information from Chinese speech F0 contours should be able to deal with the variations arising from the *articulatory constraints*, the syllable-level tones and the high-level prosody information.

The *Tone Nucleus* model [4], a linguistic framework resulting from the phonetic findings in decades, possibly offers such an way to deal with the variations to guide information decomposition. It 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 optional, non-deliberately produced articulatory transition F0 loci. Fig. 1 illustrates some frequently observed tonal F0 variations in continuous speech and their tone nuclei notations.

Studies through automatic tone recognition showed that: tone recognition based on the features of only tone nuclei led to better performances than on those of whole syllable [6]; Tonal F0 normalization based on the onsets and the offsets of tone nuclei led to significant performance improvement [7]; The Chinese intonation model, Hypo-and-Hyper articulation model which was defined based on the Tone Nucleus model, is efficient to improve the tone recognition accuracy [7]. The tone nuclei are not only critical to tone recognition, but also critical to discern and, probably, automatically detect the high-level prosody information like prosodic phrasing and foci[4].

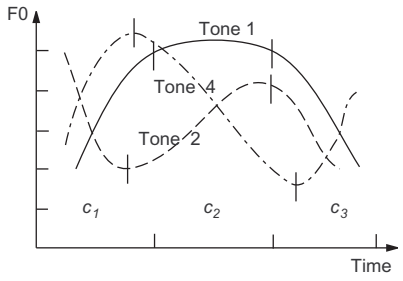

Fig. 1. Illustration of syllable F0 contours with possible articulatory transition loci for Tone 1, Tone 2 and Tone 4. The left and right vertical sticks in each contour correspond to the possible tone onset and offset F0 values, and the medium F0 locus delimited by the tone onset and offset in each contour represents the tone-nuclei. C1, C2 and C3 depict onset courses, tone-nuclei and offset courses.

One important prerequisite for the Tone Nucleus based approaches is that tone-nuclei need to be located in the syllable F0 contours first. However, there are several difficulties to robustly do this: First, it is not trivial to get appropriate segmentation of possible F0 loci. As the number of optional loci is unknown beforehand, it must be decided automatically during the segmentation. Also, it is not easy to find appropriate turning points in such F0 contours as the Tone 1 in Fig.1 using methods of peak and valley points [8]. The second is that we have little knowledge about what kinds of features are efficient to characterize the tone nuclei.

To deal with these problems, we have developed a robust tone-nuclei segmentation and location method based on statistical means in two steps: the first step is F0 contour segmentation based on the iterative segmental K-means segmentation procedure, with which a T-Test based decision

of segment amalgamation is combined. The merit of T-Test based amalgation is that no physical thresholds but a statistic evidence was used to decide when to amalgamate two segments. The second step is tone-nucleus decision based on linear discriminant analysis (LDA) function. A sub-optimal feature selection method, the sequential forward selection (SFS) [9], was realized to select a few number of prosodic features to form an efficient discriminant function. The use of LDA and SFS helped to learn the statistic distributional characteristics of tone nuclei. This paper gives a detailed description of the proposed tone nucleus segmentation and location method, together with the experimental results of tone nuclei detection and tone recognition.

2. TONE NUCLEUS SEGMENTATION AND LOCATION

2.1. Iterative Segmental K-means Segmentation

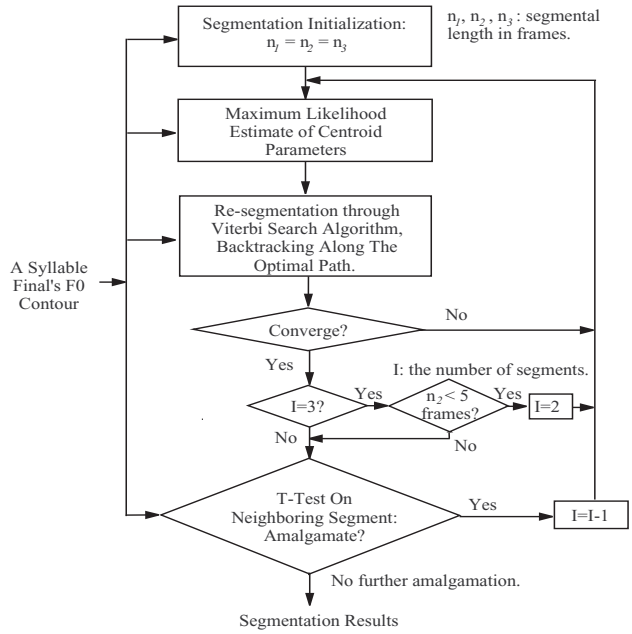

Fig. 3. The procedure to segment a syllable F0 contour.

Fig 3 illustrates the iterative Segmental K-means segmentation algorithm used to segment a syllable F0 contour into a few loci. The observation vector o_j is a two component vector $(\log F0_j, \Delta \log F0_j)$. The ith, $1 \le i \le I$, locus centroid is assumed to have the p.d.f. of the multivariate Gaussian $p(o|\Phi_i)$, where the parameter vector Φ_i include the mean-vector μ_i and covariance matrix Σ_i, which are obtained from the n_i observation points of the ith locus by the maximum likelihood estimates.

$$\hat{\mu}_i = \frac{1}{n_i}\sum_{k=1}^{n_i} o_k \quad (1)$$

$$\hat{\Sigma}_i = \frac{1}{n_i}\sum_{k=1}^{n_i}(o_k - \hat{\mu}_i)(o_k - \hat{\mu}_i)^t \quad (2)$$

In the re-segmentation step, the likelihood

$$p(o_j|\Phi_i) = \frac{1}{2\pi|\hat{\Sigma}_i|^{1/2}} \exp\left[-\frac{1}{2}(o_j - \hat{\mu}_i)^t \hat{\Sigma}_i^{-1}(o_j - \hat{\mu}_i)\right] \quad (3)$$

can be used in the Viterbi search to decide which locus the point o_j belongs to. When a segmentation becomes available, a check will be made whether two successive loci can be almalgamated or not based on the following two principles:

- Phonetic rule: a tone-nucleus should be longer than 50 ms, which is reported lowest limit for human pitch perception.
- Statistical rule: if there is no significant statistical evidence for distributional differences between two neighboring F0 loci, they are merged.

When a Final's F0 contour is divided into 3 loci, the medium locus should correspond to tone-nucleus according to the Tone Nucleus model, and its length is required to be longer than 50ms, i.e., $n_2 \ge 5$ according to the phonetic rule. Otherwise the number of loci, I, will be reduced to 2 and the segmentation will be repeated once more. Next, two successive loci of F0 contour will be checked whether to amalgamate into one or not based on T-Test on the slope ratio K of the linear regression function of $\log F0$ on time t. For the observation point $(t_j, \log F0_j)$ in the segment c_i:

$$\log F0_j = K_i t_j + C_i \text{ where } C_i \text{ is a consonant.}$$

K_i may be estimated by taking the average of $\Delta \log F0_j$ for the n_i points in c_i:

$$\hat{K}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} \Delta \log F0_j$$

We test the two hypotheses on if any two neighboring loci have the same slope ratio or not,

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

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

$$T_{i,i+1} = \frac{\hat{K}_i - \hat{K}_{i+1}}{\sqrt{S(\frac{1}{n_i} + \frac{1}{n_{i+1}})}} \quad (4)$$

$$S = \max\left\{ \frac{\sum_{i=1}^{I}\sum_{j=1}^{n_i}(\Delta \log F0_{ij} - \hat{K}_i)^2}{N-I},\ S_0 \right\}$$

where

S_0 : the micro-fluctuations allowance, calculated by the average variance of $\Delta \log F0$ of whole utterance.

N : number of voicing points in the Final's F0 contour

I : number of F0 loci

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

$$cp = t_{N-I,1-\alpha/2}\sqrt{S(\frac{1}{n_i} + \frac{1}{n_{i+1}})} \quad (5)$$

where point $t_{N-I,1-\alpha/2}$ satisfies

$$P[-t_{N-I,1-\alpha/2} \le T_{N-I} \le t_{N-I,1-\alpha/2}] = 1 - \alpha$$

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 the same slope ratio, and the two loci will be mergered. A tonal F0 contour may be divided into one locus, two loci or three loci, from these loci one is decided as the tone nucleus. For one-locus case, the whole Final's F0 contour is the tone-nucleus. For three-loci case, the medium locus is the tone-nucleus. For two-loci case, the tone-nucleus is chosen by the linear discriminant function explained in the following section.

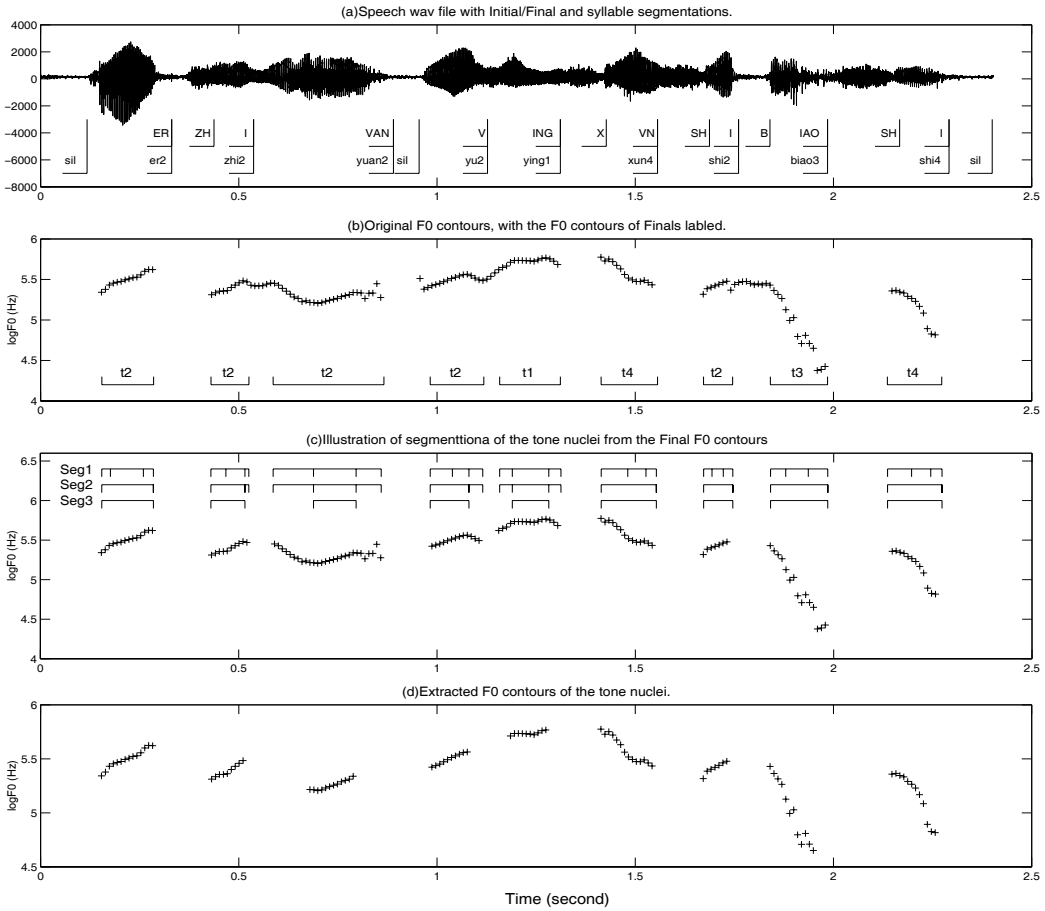

Fig. 2. Illustration of the extraction of tone nuclei.The panel (a) depicts the speech wave file, with phonetic segmentations of Initials/Finals and syllables. The panel (b) depicts the original F0 contours, with the tonality and the Finals' boundaries labled. The panel (c) illustrated the segmentation process for the tone nuclei from the Finals' F0 contours, where the "Seg1" depicts the results of segmental K-means segmentation, the "Seg2" for the results of segment merge, and the "Seg3" for the tone nuclei. The panel (d) depicts the extracted F0 contours of tone nuclei.

2.2. Tone Nucleus Discriminant Function

For a two-loci F0 contour, we extracted a number of prosodic features such as, timing, power and F0 related ones, and denoted by a vector x. We use a linear discriminant function as the automatic classifier to decide which of the two loci is the tone nucleus.

$$y = w^T x + w_0 \tag{6}$$

$$x \in \begin{cases} \text{Group } -1 & y < 0 \\ \text{Group } 1 & y > 0 \end{cases} \tag{7}$$

With the training samples $(x_{-1,1}, \ldots, x_{-1,N_{-1}})$ of Group -1, and $(x_{1,1}, \ldots, x_{1,N_1})$ of Group 1, we use Fisher ratio J_F [9, pp.104-105] to find the w of equation 6.

$$J_F = \frac{|w^T(\mu_{-1} - \mu_1)|^2}{w^T S_w w} \tag{8}$$

where,

$$\mu_i = \frac{1}{N_i} \sum_{j=1}^{N_i} x_{i,j} \text{ where } i = -1, 1. \tag{9}$$

$$S_w = \frac{1}{N-2}(N_{-1}\hat{\Sigma}_{-1} + N_1\hat{\Sigma}_1) \text{ and } N = N_{-1} + N_1.$$

where $\hat{\Sigma}_{-1}$ and $\hat{\Sigma}_1$ represent the maximum likelihood estimates of the convariance matrices of Group -1 and 1 respectively. Through maximizing the ratio J_F, the optimal weight w is calculated as,

$$w = S_w^{-1}(\mu_{-1} - \mu_1)$$

Furthermore, through the sequential forward selection (SFS) [9, pp.222-223], a subset of prosodic features can be chosen to form an efficient discriminant function. The chosen features can reveal the statistic distributional characteristics of the tone nuclei. The SFS works like this:

1. Initialization: let X, the objective feature set, be null.
2. J_F estimation: for each x_i in the candidate feature set, form a new feature set (X, x_i), and calculate the Fisher ratio $J_F(X, x_i)$ in equation 8.
3. Best candidate: the feature $x^* = \arg\max J_F(X, x_i)$.
4. Significance check: $(J_F(X, x^*) - J_F(X))$ is checked to see if the improvement is statistically significant. If yes, include x^* into X, and delete x^* from the candidate feature set, then go back to 2. Otherwise, output X, and End.

3. EXPERIMENTS

3.1. Tone Nuclei Detection Experiment

Experiments were carried out on data of a female speaker (0f) in the published corpus HKU96. 250 utterances from cs0f0001 to cs0f0250 were used to train the LDA function of tone nuclei, and another 50 utterances from cs0f0251 to cs0f30300 used as the test data. F0 was extracted by integrated F0 tracking algorithm (IFTA) with a frame shift of 10ms. Phonetic segmentation were assumed available, and achieved by force alignment of the database using Initial and Final acoustic models. After the iterative segmentation and almalgamation, there are 664 two loci samples in the training data. We then manually label them to either Group -1 or 1. Among them 289 samples belong to Group -1, and the other 375 ones to Group 1. The estimated LDA function had an accuracy of 95.2% for discriminting the these 664 training samples into two groups. Fig. 2 illustrates the whole procedure to segment and locate the tone nuclei.

We manually checked the extracted tone nuclei of the 671 tones in the 50 test utterances to evaluate the performance of the segmentation and location of tone nuclei. We regard those tone nuclei as errors only when they obviously missed the targets. So, if the whole F0 of a Final is chosen as the tone nucleus, it is regarded as correct no matter whether it includes transitory courses or not. Table 1 shows the number of samples in different segmentation groups and the correct rate of extracted tone nuclei.

	1 locus	2 loci	3 loci	Total
# of tones	247	161	263	671
# of correct nuclei	247	152	255	654
Correct rate	100%	94.4%	97.0%	97.5%

Table 1. Detection performance of tone nuclei for an open set of 50 utterances.

3.2. Tone Recognition Experiments

The application of the automatic tone nuclei detection is shown through tone recognition experiment. In this experiment, 500 utterances from cs0f0001 to cs040500 were used as training set, while 200 utterances from cs0f0501 to cs0f700 were used as testing set. Continuous density HMMs with left-to-right configuration were used as lexical tone models. The number of states for the four basic tones is 5, and that for the neutral tone is 3. Mixture number is 6 per state. The standard feature vector has $\log F0$, frame energy and their 1st, 2nd order time derivatives.

Comparison recognition experiments have been made with respect to the use of acoustic features of either full syllable or tone nucleus, and the use either 5 context independent (CI) tonal HMMs or 176 context dependent (CD) tonal HMMs.

Tonal HMMs	Recognition correct rates (%)	
	Full syllable	Nucleus
CI	75.3	81.5
CD	76.2	83.1

Table 2. Average correct rates for the four basic and the neutral tones.

3.3. Discussions

Table 2 gives recognition results, we see,

- Using tone nuclei to recognize tones improved performances significantly. In the approaches of both the CI and CD HMMs, the technique brought by more than 6% absolute improvements when compared with the standard approach using full syllabic features.
- The significant recognition improvements by the tone nuclei also indicates that the tone nuclei are robustly segmented and located by the methods presented.

4. CONCLUSION

This paper presents an efficient method for detecting tone-nuclei from speech signals. The method uses iterative Segmental K-means segmentation method and T-test almalgamation to robustly get linear F0 loci in a syllable F0 contour, and an LDA function to discriminate tone nuclei among the two-loci cases. The selected features for the LDA function by the SFS algorithm reveal the statistic distributional characteristics of the tone nuclei. Tone nuclei detection experiment and tone recognition experiments showed the presented method has detected the tone nuclei appropriately. In the next step, we will study the problem arising from applying the method to speaker independent tasks.

Acknowledgement

This research was supported in part by the Telecommunications Advancement Organization of Japan.

5. REFERENCES

[1] M. Ostendorf, "Evaluating the use of prosodic information in speech recognition and understanding", Final report to NSF and ARPA, May, 1997

[2] H. Fujisaki, "Prosody, Models, and Spontaneous Speech",In Y. Sagisaka, N. Campbell and N. Higuchi, editors, Computing Prosody: computational models for processing spontaneous speech, New York: Springer-Verlag, 1997. pp.27-42.

[3] B. Granstrom, "Applications of Intonation - An overview", ESCA workshop on Intonation: Theory, Models and Applications, Athens Greece, Sep. 1997, pp.21-24.

[4] J.-S. Zhang and K. Hirose, "Tone nucleus modeling for Chinese lexical tone recognition", Speech Communication, forthcoming.

[5] Y. Xu, "Effects of tone and focus on the formation and alignment of F0 contours", Journal of Phonetics, Vol.27, No.1, 1999, pp.55-105.

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

[7] J.-S. Zhang, K. Hirose and S. Nakamura, "A multilevel framework to model the inherently confounding nature of sentential F0 contours for recognizing Chinese lexical tones", ICASSP2003, Vol. I, pp.776-779.

[8] E. Garding and J.-L. Zhang, "Tempo effects in Chinese prosodic patterns", ESCA workshop on Intonations: Theory, Models and Applications. Athens Greece, Sep. 1997, pp.145-148.

[9] A. Webb, "Statistical pattern recognition", Arnold press, London, 1999.