Speech Prosody 2004
Nara, Japan
March 23-26, 2004

Tonal Contextual F0 Variations and Anchoring Based Discrimination

Jin-Song Zhang[1], *Satoshi Nakamura*[1] *and Keikichi Hirose*[2]

[1]ATR Spoken Language Translation Research Laboratories
[2]Dept. of Frontier Informatics, Univ. of Tokyo
e-mail: {jinsong.zhang,satoshi.nakamura}@atr.jp, hirose@gavo.t.u-tokyo.ac.jp

Abstract

We investigated the problem of discrimination of some substantial tonal contextual F0 variations in Chinese continuous speech. We proposed that anchoring discrimination hypothesis might serve as an important cue for human beings to discriminate those tones. Experimental results from statistic distributional analyses and tone recognition experiments provided strong supports for this proposal.

10.21437/SpeechProsody.2004-121

1. Introduction

The four Chinese basic lexical tones (referred to as Tones 1, 2, 3 and 4) are usually characterized according to their different patterns of fundamental frequency (hence F0) contours, i.e., Tone 1 with a high-level, Tone 2 with a mid-rising, Tone 3 with a low-dipping and Tone 4 with a high-falling F0 contour [1]. For the discrimination between them, researchers reported in the previous studies that F0 height is critical for discriminating Tone 1 (high F0) and Tone 3 (low F0) [2], Tone 3 (low F0) and Tone 4 (high F0) [3]. F0 slope coefficients are helpful for discriminating between tones with different dynamic directions, such as Tone 1 (flat), Tone 2 (rising) and Tone 4 (falling) [2].

The tonal F0 contours may vary substantially in continuous speech compared with those in isolated syllables. Not only the F0 heights but also the slope coefficients may change so much that the underlying tonalities can not identified from the surface F0 contours. On the other hand, perception experiments showed human beings are able to perceive the purported underlying lexical tones with high consistency despite of the substantial F0 variations, provided the tonal context [4]. This indicates that there exist other discriminating cues in the tonal context besides the F0 height and F0 slope coefficients.

By investigating two well-known tonal contextual F0 variations and one tonal perception result [4], we suggested that anchoring based features [6, 5] can provide a consistent account for the fact that human beings can discriminate those tones in contextual F0 variations. Also, experimental results from statistic distributional analyses [5] and tone recognition [6] can partially serve as sound evidences of our suggestion.

2. Tonal Contextual F0 Variations Phenomena

The contextual variations we investigated here include two well-known phenomena, *downstep lowering* and *contextual assimilation*. Strictly speaking, the downstep lowering phenomenon is a kind of contextual assimilation in the specific context of a high pitch following a low pitch target.

2.1. Downstep Lowering Effect

Downstep lowering effect is known as the phenomenon that in a "HLH" tone sequence the second "H" tone is lower than the first "H" in F0 height due to the existence of the "L" tone. If it acts successively on an alternate H and L sequence, it may result in a step-like function in F0 contours [7]. F0 height of an H tone in a latter position of an utterance may turn out to be lower than that of an L tone in a former position within the utterance. Three of the four basic lexical tones in Chinese own low target, either in onset (Tone 2), offset (Tone 4) or both onset and offset (Tone 3), as in Table 1. Fig. 1 is an example of the downstep phenomenon. From it, we can see two interesting phenomena:

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

Table 1: *Onset and offset pitch values of the four lexical tones. "H","L" depict high and low targets respectively. "L" is suggested to be the trigger of downstep effect.*

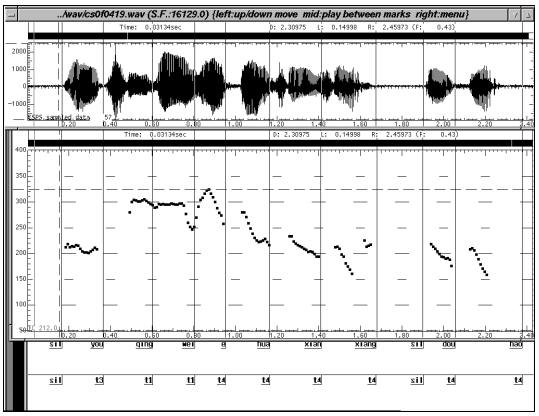

Figure 1: *An example of downstep lowering effect. The utterance text is: "you3 qing1 wei1 e4 hua4 xian4 xiang4" (There are phenomena of slightly worsening).*

1. Downstep effects took place successively three times on the later 3 Tone 4s, making them into a descending staircase.
2. We also can see H onset of the final Tone-4 is with the same level F0 height of the initial Tone-3 of the utterance.

Although F0 heights of the lexical tones undergoing downstep lowering effects may be rather substantial, they

have been well known to exert no interference in human pitch perception, i.e., those Tone 4s in Fig.1 are perceived not only as the same tonality, but also of nearly same heights in onsets. Furthermore, in speech synthesis studies, if two neighboring tones which are supposed to undergo downstep lowering effect are assigned with the same F0 values at high pitch point, the synthesized speech will be perceived as an unusual stress in the second lexical tone [8].

2.2. Contextual Assimilation Tonal Variations

The concept of contextual assimilation tonal variations is used here to indicate the phenomena that the assimilation variation effect is so severe that the tonal F0 slope directions are even changed to those of other tonalities. Fig. 2 gives an example in which the circled F0 contour is Tone 2 whose distinctive pattern is a rising slope. But here it changes to a flat F0 contour, which is of Tone 1.

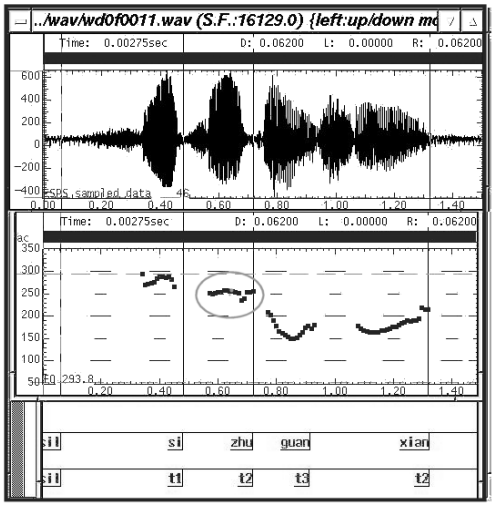

Figure 2: *An example of contextual assimilation. The utterance text is: "si1 zhu2 guan3 xian2" (traditional stringed and woodwind instruments). The tone in gray circle, supposed to be rising, changed into a flat shape due to preceding Tone 1.*

One important question is that whether the tone undergoing F0 contour variation has changed into another tone or not. The question is so confusing that once rather different opinions were proposed. In [1, pp.27], a phonological tone sandhi rule was even proposed to describe such a case as that in Fig. 2. It suggests that a Tone 2 sandwiched between a preceding tone with high offset and a succeeding tone with one of the four basic tones would change into the Tone 1 for speech at conventional speed. However, recent studies revealed that as long as the tonal context leading to the contextual F0 variations is present, listeners still perceive the tone as the original tonality [4]. This means that the Tone 2 subject to F0 variation in Fig.2 is still perceived as Tone 2 rather than Tone 1. Our listening to the utterance confirmed this conclusion.

2.3. Two Findings of A Perception Experiment

Xu gave a more systematic investigation into the tonal contextual F0 variations and their influences on human pitch perception in [4]. One interesting experiment he carried out is to examine how human perceive a sandwiched tone in a tri-syllable sequence and the same sandwiched tone in a swapped context of the first and third syllables. Since swapping of the first and the third tones may result in significantly different tonal environment for the sandwiched tone, it is hoped to offer a special way to test the tonal contextual influences on pitch perception. Fig. 3 illustrates two findings from Xu's experiments:

1. When the original tonal context was compatible and the swapped context became conflicting, the sandwiched tone tended to be perceived as the same tone with a sharper slop in the swapped context than in the original context (see the upper flow chart).
2. When the original tonal context was conflicting and the swapped context became compatible, the sandwiched tone tended to be perceived as a tone with an opposite dynamic contour in the swapped context to that in the original context(see the lower flow chart).

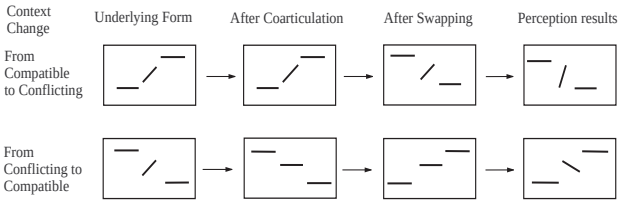

Figure 3: *Swapping contextual influences on tone perception of sandwiched lexical tones (adopted from [4]).*

Lexical Tones	Onset gap	Offset gap
Tone 1	≥ 0	≥ 0
Tone 2	≤ 0	≥ 0
Tone 3	≤ 0	≤ 0
Tone 4	≥ 0	≤ 0

Table 2: *Anchoring based feature patterns for the four basic lexical tones in continuous speech.*

3. Anchoring Hypothesis and Anchoring F0 Features

Aiming at finding a more efficient feature for discriminating the lexical tones, we adopted the psycho-acoustic perception findings [9, 10] to make the following anchoring hypothesis [6, 5]:

- Relative F0 difference between the offset point of the first lexical tone and the onset of the second lexical tone may be an important discriminating cue for high or low pitch, besides the direct cue of a gliding F0 contour.
- There should be a timing allocation mechanism for the competition effects [10].

Based on this hypothesis, a lexical tone in continuous speech can also be acoustically characterized using the patterns given in Table 2, besides using the flat, rising, dipping or lowering F0 patterns.

- Onset gap: the difference between the onset F0 and the offset F0 of preceding lexical tone.
- Offset gap: the difference between the offset F0 and the onset f0 of succeeding lexical tone.

4. Anchoring Discrimination of Tonal Contextual Variations

We found that we can exploit the anchoring hypothesis to predict consistently those tones with tonal variations above-mentioned.

4.1. Anchoring Discrimination of Downstep Tones

Fig. 4 illustrates the estimation methods for the onset and offset gaps. The r features stand for the onset and offset gaps, and the d features for the F0 slope coefficients as duration is normalized. The onset and offset points are those points corresponding to the tone nuclei [11].

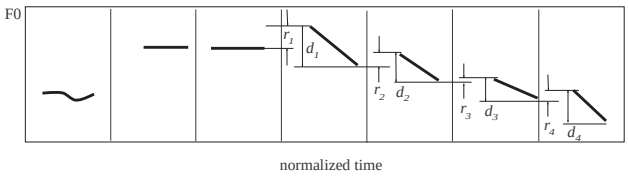

Figure 4: *Illustrations of the estimation of onset and offset gaps for the downstep tones in Fig.1. Thin vertical lines depict syllable boundary.*

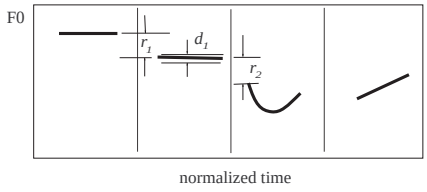

Figure 5: *Illustrations for the estimation of the onset and offset gaps for the second tone in Fig.2.*

Tone 4 has a pitch pattern of "HL". According to the anchoring tone discrimination hypothesis, satisfying the following conditions will lead to the identification of Tone 4.

1. Onset gap is positive, i.e., $r \geq 0$. Since Tone 4 has a H onset, if preceding tone's offset is L, then $r > 0$; whereas preceding tone's offset is H, then $r \approx 0$.
2. Offset gap is negative, due to the similar mechanism of the above one.
3. F0 contour slope is negative, $d < 0$. (Gliding pitch).

From the Fig.4, we can see: although the four Tone 4s differs greatly in their absolute F0 heights, they have small differences in satisfying the above conditions. Every Tone 4 in the four Tone 4s has a similar level of positive onset gap, a similar level of negative offset gap, and a similar F0 contour falling slope. Since these three features are assumed by the pitch anchoring hypothesis to be of great importance for tone discrimination, it is reasonable to predict that the four Tone 4s will have similar tone perceptions, as evidenced by listening.

4.2. Anchoring Discrimination of Contextual Assimilated Tones

Fig. 5 illustrates the estimation of onset and offset gaps r and the F0 slope coefficient d for the second tone.

From the illustration, we can see

1. Onset gap is rather negative, $r_1 < 0$.
2. Offset gap is rather positive, $r_2 > 0$.
3. F0 contour slope is flat $d \approx 0$.

By checking the Table 2, we find that Tone 2 is the most appropriate one for the above features, hence we predict the tone should be discriminated as Tone 2. It is the minus onset gap and the positive offset gap that determined the Tone 2 prediction based on the anchoring hypothesis.

4.3. Tone Predictions for the Swapped Context

Based on anchoring tone discrimination hypothesis, we can also easily get correct predictions of tonality perceived in the above-mentioned swapped context experiments. Figure 6 illustrates our prediction procedure and results for the sandwiched tone in swapped context. A target is first predicted as a high (+) or a low(-) target based on the onset, offset gaps and gliding pitch, then a tonality can be predicted based on the two targets in its onset and offset. We can see the predictions are consistent with the reported two findings.

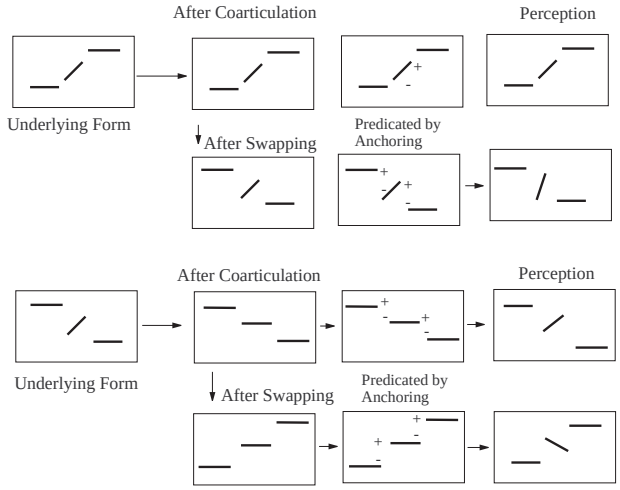

Figure 6: *Predictions of tonality changes in the swapping context perception experiments of [4] by the "Anchoring" hypothesis. The "+" implies a possible high pitch target and "-" implies a possible low pitch target.*

5. Evidences from Statistic Distributional Analyses and Tone Recognition

We have had evidences from the statistic distributional analyses [5] and tone recognition experiment [6].

5.1. Statistic Distributional Analyses

We randomly selected 10 continuous utterances per speaker for both the 10 male and 10 female speakers in the data corpus HKU96 and got 2147 samples for the four basic lexical tones. Then collapsed the data of each speaker to each sample each tone by taking the mean values of the same tone, and got a tone-balanced set consisting of the 80 collapsed tones from the 20 speakers. F0 contours in logarithmic scale were normalized with respect to each speaker's mean log F0 and standard deviation to remove F0 range differences of inter-speakers.

$$z = \log F0_{norm} = \frac{\log F0 - \overline{\log F0}}{\hat{\sigma}_{\log F0}}$$

For each tone, two z values are collected:

- z_0 for tone onset,
- z_1 for the tone offset.

And four anchoring features are calculated.

- $z_i^L = z_i - z_1$ of the preceding tone: indicate the left-to-right anchoring effect, $i = 0, 1$.
- $z_i^R = z_i - z_0$ of the succeeding tone: indicate the right-to-left anchoring effect, $i = 0, 1$.

Table 3: *Tone based group mean values of the collected features.*

Tone	z_0	z_1	z_0^L	z_1^L	z_0^R	z_1^R
Tone 1	.912	.874	.913	.876	.496	.458
Tone 2	-.384	.386	-.475	.295	-.866	-.009
Tone 3	-.322	-1.306	-.686	-1.670	-.433	-1.418
Tone 4	.596	-.8108	.850	-.557	.916	-.491

Table 4: *Tone onset based group statistics.*

Tone onset		z_0	z_1	z_0^L	z_1^L	z_0^R	z_1^R
mean value	L	-.353	-.460	-.580	-.688	-.650	-.757
	H	.754	0.003	.882	.159	.706	-.002
F(1,78)		477.7	6.2	652.5	17.3	283.1	26.6
P		.000	.015	.000	.000	.000	.000

Table 5: *Tone offset based group statistics.*

Tone offset		z_0	z_1	z_0^L	z_1^L	z_0^R	z_1^R
mean value	L	.137	-1.059	0.008	-1.114	.242	-.954
	H	.264	.630	.219	.586	-.185	.181
F(1,78)		.886	486.6	.62	213.4	6.6	115.7
P		.349	.000	.434	.000	.012	.000

Table 3 listed the mean z values. Table 4 is ANOVA analyses with respect the onset target (L/H) while pitch values of offset are collapased, i.e., the L group includes samples from tone 2 and 3, while H group includes tone 1 and 4. Table 5 is ANOVA analyses with respect to the offset target. The major findings from these analyses includes:

- The mean values of z_0^L and z_1^R represent the onset and offset gaps in Table 2, and they conformed to the hypothesis very well.
- The statistics showed that it is very evidential that the proposed anchoring features are discriminative for the H and L targets, either on the onset or offset.
- The discriminating efficiencies of the six features for H and L onset targets can be given based on the F ratio of between-group variance and within-group variance:

$$z_0^L > z_0 > z_0^R > z_1^R > z_1^L > z_1$$

- The discriminating efficiencies of the features for the H and L offset targets can be ordered as:

$$z_1 > z_1^L > z_1^R > z_0^R > z_0 > z_0^L$$

5.2. Results from Tone Recognition Experiments

Experiments of tone recognition of continuous speech have also been carried out to give a comparison between the anchoring features and the conventional F0 features [6]. Table 6 summarizes the tone recognition accuracies in the experiments using context independent tone HMMs. The significant accuracy improvement of 12.2% proved the efficiency of the anchoring features for tone discrimination.

	T1	T2	T3	T4	T5
T1	88/69.2	6.3/16.2	0/1.4	5.2/12.7	0.5/0.5
T2	5.9/8.5	88/76.4	2.5/8.7	2.2/4.2	1.4/2.2
T3	0.3/0	3.9/10.4	84.6/70	4.7/17.2	6.5/2.4
T4	4.4/4.5	0.7/3.5	2.3/5.7	91/85.3	1.5/1.0
T5	7.3/2.4	5.7/11	25/22.6	27/33.5	35/30.5

Table 6: *Confusion matrix indicating tone recognition accuracy in percentages for the testing set of two systems. T1 -T4 represent the four basic lexical tones and T5 the neutral tone. Left digit in each cell depicts the result of the system using anchoring features and the right one using the conventional F0 features. Total improvement by our method is above 10% in accuracy (85.5%, improved from 75.3%).*

6. Conclusion

We presented here that anchoring hypothesis might be an important cue for discriminating those tones subject to substantial contextual F0 variations. Evidences from statistical distributional analyses and tone recognition experiments proved that anchoring based features are efficient to reduce the confusions between the four tones in continuous speech.

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

7. References

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

[2] D. Whalen and Y. Xu, "Information for Mandarin tones in the amplitude contour and in brief segments", Phonetica 49, 1992, pp.25-47.

[3] E. Garding et al, "Tone 4 and tone 3 discrimination in modern standard Chinese", Language andspeech, Vol.29, Part 3,1986, pp.281-293.

[4] Y. Xu, "Production and perception of coarticulated tones", J.A.S.A. (4), 1994, pp.2240-2253.

[5] J.-S. Zhang, S. Nakamura and K. Hirose, "Discriminating Chinese lexical tones by anchoring F0 Features", Proc. of ICSLP2000, Vol.II, pp.87-90.

[6] J.-S. Zhang, K. Hirose, "Anchoring hypothesis and its application to tone recognition of Chinese continuous speech", ICASSP2000, Vol.III, pp.1419-1422.

[7] Ch.-L. Shih, "Declination in Mandarin", Proc. from the ESCA Workshop on Intonation, Athens Greece, 1997.

[8] P. J. Rose, "On the non-equivalence of Fundamental frequency and pitch in tonal description", D. Bradley, E. J. A. Henderson and M. Mazaudon eds, Prosodic analysis and Asian linguistics: to honour R. K. Sprigg, Pacific Linguistics, C-104, 1988, pp. 55-82.

[9] T. Tsumura et al,"Auditory detection of frequency transition", J.A.S.A. Vol. 53, No. 1, 1973, pp.17-25.

[10] S. Shigeno, H. Fujisaki, "Effect of a preceding anchor upon the categorical judgment of speech and non-speech stimuli", Japanese Psychological Research, 1979, Vol. 21, No. 4, pp.165-173.

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