Transcribing Tone – A likelihood-based quantitative evaluation of Chao's *tone letters*

Phil Rose

Division of Humanities, Hong Kong University of Science and Technology
College of Asia and the Pacific, Australian National University
philjohn.rose@gmail.com

Abstract

The accuracy of the widely used and International Phonetic Association-sanctioned Chao five-point scale of tonal transcription is examined quantitatively. Perceptually transformed acoustic data are used from two Chinese dialects with complex tone systems, and a measure derived of the conformability of the data using their likelihoods. It is shown that some tones conform well to the model, but others do not, with tonal pitch targets lying equidistant between the Chao integers. It is concluded that the Chao model is probably not an accurate reflection of the distribution of tonal pitch targets.

Index Terms: Chao tone letters, tonal F0, tonal pitch, likelihood, Bayes' theorem, Cantonese, Wencheng

1. Introduction

The study of linguistic diversity is logically underpinned by what has been dubbed the *Indispensible Foundation* of Phonetics [1]. It is the key that unlocks the door to all the other linguistic levels in which languages differ – phonology, morphology, syntax, semantics. Logically prior to phonology and acoustic quantification is a phonetic transcription which can accurately represent the range of allophonic variation realising contrastive differences between speech sounds. Although the phonetic transcription of segments is arguably well catered-for by the categories of the International Phonetic Alphabet, the phonetic transcription of tone – a major phonological category – is not. The International Phonetic Alphabet [2] provides only ten separate symbols for pitch transcription: [˥ ˦ ˧ ˨ ˩] representing five level pitches (*extra high*, *high*, *mid*, *low, extra low*), and five contour pitches: [˩˥ ˥˩] (*falling*, *rising*), [˧˥ ˨˦] (*high* and *low rising*), and *rising-falling*. Augmentation is provided by the method of pitch notation developed by the Chinese phonetician and polymath Chao Yuenren [3]. Sanctioned by the International Phonetic Association in their 1989 revision [4], this well-known method using iconic five-point 'tone letters' remains the preferred method for transcribing tonal pitch in Asian languages. I attempt to evaluate it quantitatively in this paper. Claimed to be "the only well-developed model of speech description" [5], the IPA alphabet has thus theoretical import, and, as with all hypotheses, it is therefore important to test it. The idea is to take an appropriate perceptual transform of the main acoustic correlate of tonal pitch (fundamental frequency) from two Chinese varieties with complex tone systems, and evaluate how well the tonal pitch targets derived from the perceptually transformed acoustics conform to Chao's model with five equally spaced tonal pitch levels. The term *pitch* is now often used to refer to its acoustical correlate of *fundamental frequency*, so it is probably important to point out that the nature of this paper requires the two terms be carefully distinguished [6].

2. Chao's *tone letters*

Chao originally proposed his five-level *tone letters* as an accurate, elegant and convenient method of transcribing both phonetic and phonemic tonal pitch. The idea is simple: normal pitch range is divided into four equal parts, the boundaries being labeled with the integers 1 through 5, named as *low*, *half-low, medium*, *half-high* and *high* pitch respectively. Thus {13} represents a pitch rising from the lowest point in the pitch range to the middle and {513} represents a pitch falling from the top of the pitch range to the bottom and then rising to the middle of the range. {55} would indicate a level pitch at the top of the normal range. (To avoid possible confusion with this paper's reference format, e.g. "[11]", I have enclosed the tone letters in curly braces, instead of the conventional square brackets appropriate to phonetic transcription). Chao does not equate his intervals with any absolute value, although he suggests that, for learning the method, the range from 1 to 5 could correspond to an augmented fifth, i.e. that each interval would be equivalent to a musical tone.

How accurate is this five-level representation as a model for tonal pitch? Do languages really have tones that can be exhaustively and accurately represented phonetically in terms of these discrete pitch values? Equivalently, do tone systems make use of just five phonetic pitch height targets? This paper is an initial attempt to empirically and quantitatively investigate these questions.

3. Wencheng tones

Wencheng 文成, a Chinese Wu dialect of S.E. Zhejiang province, has seven contrasting isolation tones [7]. Named after their pitch, these are as follows. The **upper-mid level tone**, with a level pitch contour just above the middle of the speaker's pitch range. The **depressed upper-mid level tone**, with the same level pitch as the upper-mid level tone, but a depressed onset so that pitch over the initial third to half of duration is rising. The **lower-mid level** tone, with level pitch a little lower than the upper-mid level tone. The **high rising** tone, with a rising pitch from the middle into the upper pitch range, short length, and optional truncation by a glottal-stop. The **low rising** tone, with pitch which rises from low in the speaker's pitch range to mid. The **low fall-rise** tone, with a long dipping contour, first falling within the speaker's low pitch range to their lowest pitch, and then rising into the mid pitch range. The **mid fall-rise** tone also has a long dipping pitch, but its onset is higher, in the mid pitch range. It often has a prolonged initial level component, and it does not fall as low as the low fall-rise tone.

Mean tonal acoustic values (F0 plotted as a function of duration) for the seven Wencheng tones of a single male are shown in figure 1. Details of the corpus and measurement are in [7]. Means were calculated over between five and ten

replicates per tone. The F0 shapes have been plotted in two panels, as their complex configuration would have made it difficult to identify their shapes otherwise. Upper- and lower-mid level and depressed upper-mid level tones are shown in the left panel, with least squares lines fitted in anticipation of the following section. The remaining contour tones are in the right panel. Figure 1 shows the bottom-heavy F0 configuration typically found in complex tone systems [8], [9].

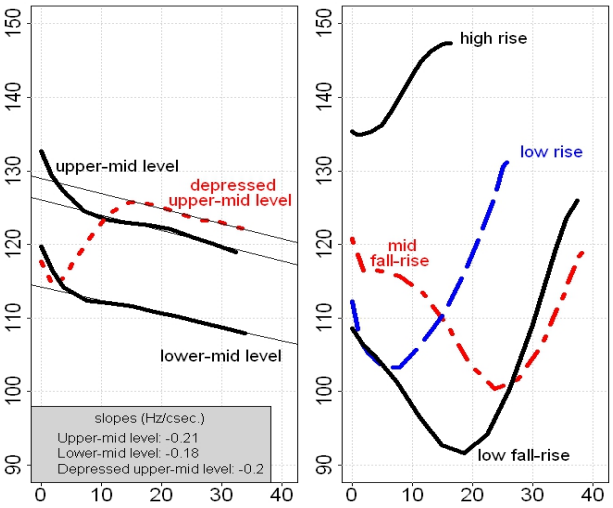

Figure 1: *Mean F0 (Hz) of the seven Wencheng tones plotted against mean duration (csec.). Least-squares lines have been fitted to the* **upper-** *and* **lower-mid**, *and* **depressed upper-mid level** *tones. Corresponding slopes are given in box.*

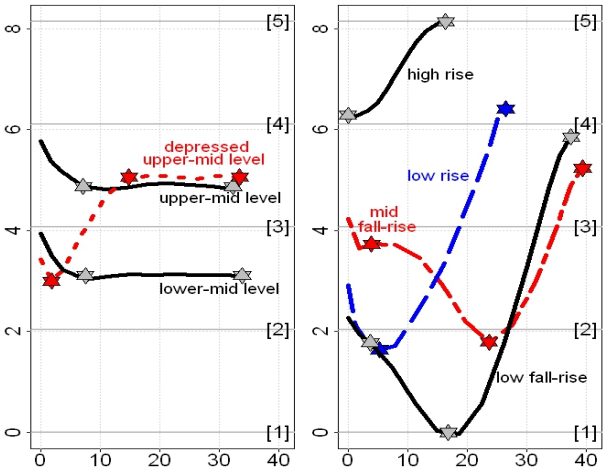

Figure 2: *Perceptual transform of Wencheng tonal F0 showing pitch targets (stars). Vertical scale at left of panels = semitones, horizontal scale = duration (csec.).*

Tones can never be acoustically observed independent of their segmental realisation, due to the well-known assumed intrinsic effects from concomitant segmental articulation, like intrinsic vowel and consonantal F0. The Wencheng tone corpus was actually well controlled for intrinsic vowel F0, but consonantal effects can still be observed in the tonal F0 in figure 1. For example, the abruptly falling F0 over about the first five centiseconds of the upper-mid level tone probably reflects [+spread] vocal cord onset consonants like voiceless fricatives and aspirated stops; and shorter perturbatory effects can be seen in some other tones. It is best therefore to understand figure 1 as showing the mean tonal acoustics plus the effect of the syllable-initial consonant.

4. Modeling Wencheng tones with Chao's *tone letters*

How well can the Chao five-point scale model the pitch percept for Wencheng tones? For example, the pitch of the Wencheng upper-mid and lower-mid level tones brackets the speaker's mid pitch range, which means that they have to be represented by {44} and {22}. How accurate is such a representation?

An important point is that the Chao five-point scale is a pitch scale, whereas the data in figure 1 are acoustic, not perceptual. So an attempt to evaluate in terms of acoustics would be misguided, and two things were done to transform the acoustics into appropriate perceptual values. Since it has been shown that linguistic pitch perception involves taking into account F0 declination [10], [11], the mild declination observable in the F0 of the tones with level pitch components (upper-mid and lower-mid level, and depressed upper-mid level) was modeled with linear regression and the F0 of all tones adjusted by the resulting mean slope (-0.198 Hz/csec. – individual slopes are given in figure 1). Secondly, since the semitone scale has been shown to give the best approximation for native listener perception of intonational pitch [12], the resulting F0 values were then transformed into semitones relative to the lowest tonal F0 value (the trough in the low fall-rise tone). The resulting quasi perceptual transform is shown in figure 2, where it can be seen that the transformed F0 range covers about 8 semitones. This semitone scale is then mapped onto five equidistant points representing the five levels of the Chao scale, with the highest level {5} set equal to the highest point in the short high rising tone, and the lowest level {1} set equal to the trough of the low falling-rising tone. The five-point scale is shown on the right of each panel in figure 2. (Note that the range as defined here is exceedingly close to the eight semitones noted by Chao as useful intervals for pitch practice. This is not always the case: speakers can differ considerably in their semitonal range.) It can be seen that the level tones now have effectively level pitch, and the slope of the contour tones is also not as steep as with their F0 in figure 1. Putative pitch targets are indicated with stars. They were assigned according to the following principles. Initial targets were taken after discounting any initial consonantal perturbation, e.g. in the upper-mid level, mid fall-rise or low rise tones. In tones with a level pitch component, e.g. the upper-mid level tone, the target was the mean of its level pitch values, represented at two points to mirror the binomial Chao representation. The offset target in contour tones was taken as their final pitch value; and in complex tones, e.g. the mid fall-rise or depressed upper-mid level tones, the intermediate target was taken at the inflection point.

It can be appreciated from figure 2 that most of the contour tones in the right panel can be rather well represented by Chao tone letters, in the sense that their pitch targets are usually located close to one of the Chao integers. Thus the high and low rising tones could be non-controversially represented as {45} and {24} respectively, and the low fall-rise tone as {214}. The mid fall-rise tone, however, although its onset and turning points are close to {3} and {2}, has an offset pretty much intermediate between {4} and {3}.

Representing it as either {4} or {3} will therefore be procrustean (or as Chao might have written 削足適履 *cutting the foot to fit the sandal*).

The tones with level pitch components in the left panel of figure 2 are not well represented with the five-point scale. It can be seen that, in keeping with their pitch description, they do indeed fall either side of the mid pitch range point, but that their values are neither high enough nor low enough to be represented as {44} or {22}. It is also clear that the pitch height of the upper-mid level tone is the same as the offset of the mid fall-rise tone, and that this, at least for this speaker, represents a separate pitch height target in his range. The same applies to the pitch of the lower-mid level tone and the onset of the depressed upper-mid tone. If anything, this Wencheng tonal configuration would therefore appear to need seven separate points for accurate modeling.

5. Numerical evaluation

In order to progress from a visual to a quantified evaluation of how well the data fit the model, it was first transformed into a continuous probability density function. This is necessary because its conventional, discrete, form would not permit any useful evaluation, any non-integer pitch target value having a likelihood of zero, given the model.

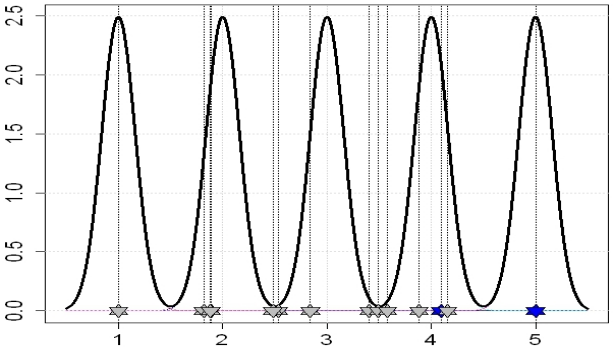

Figure 3: *GMM for evaluating Chao five-point model. X-axis = pitch target, y-axis = probability density. Stars and vertical lines indicate location of Wencheng tonal pitch targets shown in figure 2.*

A simple Gaussian Mixture Model was used, comprising five Gaussians with means equal to the five integers and a uniform standard deviation of 0.16 (chosen to make 99% of a component's distribution lie in the interval between +/- 0.5 of an integer). This is shown in figure 3. Between values of 1 and 5 the distribution has maximum probability densities at the five Chao integer values, and minimal densities at values intermediate between the integers. Stars indicate the location of the Wencheng pitch targets, from which vertical lines help to indicate the probability density of the target, given the model. Taking the high rising tone as an example, it can be seen in the right hand panel of figure 2 that its two pitch targets are close to the Chao integers {4} and {5} (the latter target artificially so because of its forced congruence with the Chao {5} value). These two pitch targets can be seen in figure 3 as the highest star on the right at {5}, intersecting the model at its maximum density, and the third highest star from the right. (The second highest star is the second pitch target of the low rising tone.)

The high rising tone's two pitch targets have densities of ca. 2.110 and 2.493, the latter being the maximum density of the model. Normalised with reference to the models' minimum density value (ca. 0.037), these become 84.5% and 100% respectively, with their mean value providing a *Chao tonal conformability index* (CCI_{tone}) of ca. 92%, indicating that it conforms well to Chao's five-point scale. In contrast, the depressed upper-mid level tone has densities of ca. 0.04 at all three pitch targets, giving a CCI_{tone} of effectively 0% – no part of it conforms to the Chao model. The CCI_{tone} values for the seven Wencheng tones are listed in table 1, as well as the conformability indices of each tone's individual pitch targets. It can be seen that the tones with values close to the Chao integers are resolved with higher CCIs. One can also of course, should one wish to do so, use the CCI_{target} values to automatically assign a Chao integer to a tonal pitch target, the sense being that the nearer the CCI approaches 0% the greater the uncertainly as to the Chao integer. The mid falling tone, for example, has CCI_{target} values of 60%, 78% and 1.7%, indicating that the first two targets can be assigned with confidence to {3} and {2}, but that there is almost total uncertainty as to whether the third target is either {3} or {4}.

	CCI_{target}			**CCI_{tone}**
tone	**Pitch Target 1**	**Pitch Target 2**	**Pitch Target 3**	
upper-mid level	3	3	-	3
lower-mid level	0	0	-	0
high rise	84	100	-	92
dep. u.-mi. level	0	0	0	0
mid fall-rise	60	78	2	47
low rise	51	62	-	57
low fall-rise	76	100	76	84

Table 1. *Chao conformability profile (%) for Wencheng tones.*

6. Application to multispeaker tonal data

Chao's tone letters are commonly used to describe the tones of a language or dialect, for example the high level tone of Hong Kong Cantonese is frequently described as {55}. Although it is never made clear how such a representation is to be arrived at from the Chao representations of the tones of individual speakers, it is possible to extend the procedure described above for one speaker's tones to evaluate the CCIs for the tones of a group of speakers (after all, it may be objected that one or more of a single speaker's tones are distributed far from the mean of the variety). Hong Kong Cantonese is used for the demonstration as it has a fairly complex tone system, with many existing Chao transcriptions (conveniently summarised in [13]); and multispeaker acoustic tonal data exist.

On syllables ending in a sonorant, conservative Hong Kong Cantonese contrasts three level, two rising and one falling pitched tones. The three level tones are located at the top, in the middle and just below the middle of the speaker's pitch range and are usually transcribed {55}, {33} and {22} respectively. Both rising tones start low in the pitch range, with one rising to high – transcribed as {35} or {25} – and one rising to mid, transcribed as {23}, {13} or {24}. The falling tone starts low and falls still lower, such that its phonation type usually becomes non-modal (breathy or creaky) as it falls below a speaker's normal pitch range. It is has been transcribed {21} or {22}. Data were taken from the first study on the normalisation of Cantonese tones [14], which used recordings

from five young male and five young female speakers, controlled as well as Cantonese phonotactics would allow for intrinsic vowel and consonantal effects on tonal F0.

As with Wencheng, the speakers' tonal F0 was first perceptually transformed by conversion to semitones and declination slope adjustment. All tones except the mid-level were adjusted by the slope of the speaker's high level tone; the mid-level tone was adjusted by its own slope. The high level tone was taken to mark the top of the Cantonese tonal range. The lowest part of the tonal range is not clear, because at least some of the F0 of the low falling tone occurs on non-modal phonation and would therefore be outside the normal range. Therefore the lowest part of the range was initially taken to be represented by the lowest part of the low-to-high rising tone. The resulting semitone values were then z-score normalised [15] to remove as much speaker-dependant information as possible, and the mean normalised values calculated for each of the six tones.

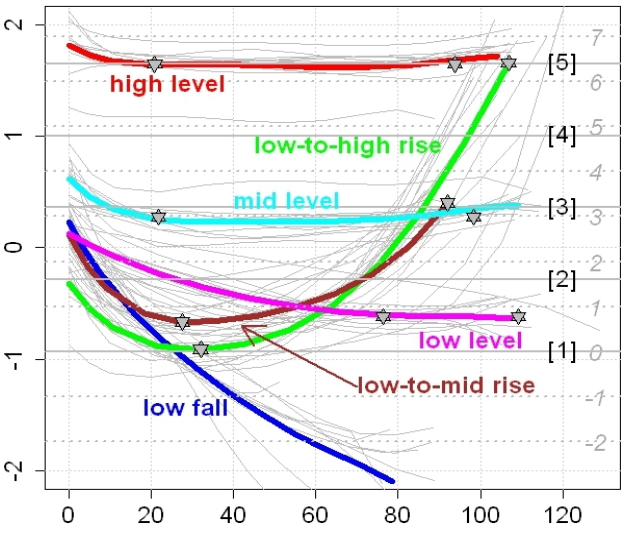

Figure 4: *Z-score normalisation of perceptually transformed tonal F0 for 10 Cantonese speakers. Stars indicate location of tonal pitch targets. X-axis = normalized duration (%), y-axis = standard deviations around mean. Dotted lines = semitones, solid lines = Chao pitch levels.*

Figure 4 shows the 10 speakers' perceptually transformed mean normalised tonal F0 as a function of normalised duration. The individual speakers' normalised tones are plotted with thinner lines. Pitch targets are again shown with stars on the mean normalised tone contours, and have been calculated and located according to the same principles as with Wencheng. I was not confident in determining pitch target(s) for the **low falling** tone, so I didn't: there is little indication of the extent of an onset perturbation and where phonation becomes non-modal. Chao integer values are shown on the right in brackets, as well as the semitone values corresponding to the normalised units in grey italics. Thus the lowest value of the **low-to-high rise** tone has a semitone value of 0 as it was chosen as the reference, and a Chao value of {1}.

As can be seen, the **high level** and **low-to-high rise** tones conform perfectly to the Chao model, as all but one of their pitch targets were forced to be congruent. These data could be accurately transcribed as {55} and {15} respectively. The **mid level** tone also shows high conformability (its CCI_{tone} was 77%) and might also be represented as {33}. The **low level** tone, however, lies mid-way between the Chao integers {1} and {2} and its CCI_{tone} value of 1.5% means that it can only be procrusteanly represented by either {1} or {2}. The second target of the **low-to-mid rising** tone has good conformability (CCI_{target} = 88%). As was pointed out in [13], its first target may still incorporate the effect of consonantal perturbation, and so its poor conformability (3.6%) should probably be discounted.

Given the problems of determining the lowest point of the tonal range, it would be necessary to optimise the system on an overall conformability measure, with the lowest point as variable. It is clear from figure 4, however, that, given the configuration, a value for the range lower than the lowest point of the low-to-high rising tone would be unlikely to improve on the overall conformability.

7. Summary and discussion

This paper has used perceptually transformed tonal acoustics from two Chinese varieties with complex tone systems to investigate the accuracy of Chao's five-point method of tonal transcription. It has shown how the degree to which the tonal data fit the model can be quantified with measures based on the likelihoods of the pitch targets, given a continuous GMM version of the five-point model. The results show that, although some tones conform well to the model, not all do; and there are examples of tonal pitch targets lying equidistant between Chao integers. Neither is it the case that the minimal spacing between tones relates clearly to the magnitude of a Chao interval.

But this is talking about how well the data fits the model. Ultimately, of course, we want to know how good the model is, given the data, and the proper way of evaluating this is with Bayes' Theorem [16]:

$$p(\text{model} \mid \text{data}) = p(\text{data} \mid \text{model}) * p(\text{model}) / p(\text{data}) \quad (1)$$

Even if we assumed non-informative priors as to the distribution of pitch accents to inform our model probability p(model) – which is unlikely, given the bottom-heavyness of tone systems – the likelihoods demonstrated in this paper (i.e. the p(data | model) would mean a very low probability of the model, given the data. It would be trivial to show that it would be much lower than, for example, a model specifying targets not at five, but at nine equidistant levels. Such considerations must cast doubt upon the accuracy of model.

This is not to disparage Chao's ability to transcribe tone. After all, in the first Western-linguistic description of Chinese dialect phonetics – his pioneering 1928 survey of Wu [17] – he demonstrated a method of tonal pitch description that has never been, nor probably ever will be surpassed. Being also an accomplished musician, Chao imitated a speaker's tonal pitch on a sliding pitch pipe and notated it musically in sufficient detail to permit subsequent recovery of its F0 [18].

I suspect that many treat the Chao five-point representations not at face value, but as proxy for appropriately precise tonal pitch/contour descriptions, like *level in the speaker's upper pitch range*, or *rising from low in a speaker's range to high*, and this would be a sensible construal. It is also expositionally convenient, as Chao opined, to be able to succinctly write {244} instead of *depressed upper-mid level*. Nevertheless the use of integers easily encourages the belief that they have a real reference, which, I think, this paper has shown is likely not to be the case.

8. References

[1] Henderson, E.J.A., The Indispensible Foundation: a selection from the writings of Henry Sweet, Oxford University Press, 1971.

[2] Handbook of the International Phonetic Association, Cambridge University Press, 1999.

[3] Chao Yuen Ren, "ə ꜱistim əv "toun-letəz"" [A system of tone letters], Le Maître Phonétique, 45, 24-27, 1930.

[4] Maddieson, I., "The transcription of tone in the IPA", Journal of the International Phonetic Association, 20(2), 28-31, 1990.

[5] Laver, J., Principles of Phonetics, Cambridge University Press, 1994.

[6] Rose, P., "On the non-equivalence of fundamental frequency and pitch in tonal description", in D. Bradley, E. Henderson and M. Mazaudon [Eds], Prosodic Analysis and Asian Linguistics: to Honour R.K. Sprigg, Pacific Linguistics, 55-82, 1989.

[7] Rose, P., "The long and the short of Wencheng tones – acoustic and auditory description of tonologically challenging phenomena in an Oujiang Wu dialect of Chinese", in M. Tabain, J. Fletcher, D. Grayden, J. Hajek and A. Butcher (Eds), Proc. 10th Australasian Intl. Conf. on Speech Science & Technology, 54-57, 2010.

[8] Zhu, Sean and Rose, P., "Tonal complexity as a dialectal feature: 25 different citation tones from four Zhejiang Wu dialects", Proc. ICSLP, 919-922, 1998.

[9] Steed, W. and Rose, P., "Same Tone, Different Category: Linguistic-Tonetic Variation in the Areal Tone Acoustics of Chuqu Wu", Proc. Interspeech 2009, 2295-2298.

[10] Lieberman, P., Intonation, Perception, and Language, M.I.T. Research Monograph 38, M.I.T. Press, 1967.

[11] Pierrehumbert, J., "The perception of fundamental frequency declination", JASA 66(2), 363-369, 1979.

[12] Nolan, F., "Intonational equivalence: an experimental evaluation of pitch scales", Proc. 15th Intl. Congr. Phonetic Sciences, 771-774, 2003.

[13] 嚴至誠 Yim Chi Sing, "A Phonetic Study of Syllabic Constituents in Hong Kong Cantonese", Unpublished Ph.D. thesis, Hong Kong University of Science and Technology, 2012.

[14] Rose, P., "Hong Kong Cantonese Citation Tone Acoustics: A Linguistic-Tonetic Study", in M. Barlow [Ed], Proc. 8th Australian Intl. Conf. on Speech Science and Technology, 198-203, 2000.

[15] Rose, P., "Considerations in the normalisation of the fundamental frequency of linguistic tone", Speech Communication, 6(4), 343-352, 1987.

[16] Kruschke, J.K., Doing Bayesian Data Analysis, Academic Press, 2011.

[17] 趙元任 Chao Yuen Ren, 現代吳語的研究 [Studies in the Modern Wu-Dialects], Tsing Hua College Research Institute Monograph 4, 1928.

[18] Rose, P., "A Linguistic Phonetic Acoustic Analysis of Shanghai Tones", *Australian Journal of Linguistics,* 13, 185-219, 1993.