



F₀ analysis of perceptual distance among Cantonese level tones

Rerrario Shui-Ching Ho^{1,2}, Yoshinori Sagisaka¹

¹ Global Information and Telecommunication Institute, Waseda University, Tokyo, Japan

²Englisches Seminar, Universität Basel, Switzerland

shui-ching.ho@unibas.ch, sagisaka@giti.waseda.ac.jp

Abstract

This paper presents an acoustical analysis of the pitch height of the four level tones of Cantonese in search for a quantitative relationship of their perceptual distance. Our preliminary measurements and calculations give the first evidence that the conventional representations were mostly mistaken.

Index Terms: Cantonese, lexical tone, tone-letter notation, pitch, prosody, perceptual distance

1. Introduction

A precise quantitative characterization of the canonical F₀ contours of the lexical tones of a tone language is fundamental not only to speech recognition and synthesis but also to accounting for speech problems across a large range of language disciplines. A slight deviation in the pitch pattern of lexical tones is not only perceived as out of tune by native speakers but often leads to unintelligibility or misunderstanding. In speech technology, an inaccurate pitch will often cause unnaturalness in synthesized speech and affects the performance of automatic speech recognition. This is particularly so in Cantonese, where four out of her six lexical tones are level.

Until the past decade, there had not been much work on modeling and synthesis of the surface F₀ contour of continuous speech of Cantonese. Assumptions about the canonical tone patterns have been uncritically and mainly based on the subjective and non-native account stemming from the pre-war era [1, 2], when neither the modern computer nor speech instruments were available. On top of that, in more than half a century, Hong Kong has undergone extremely rapid development in all sectors [3]. Her population has already expanded a few folds. The Cantonese vernacular spoken there has become localized and attained its unique identity, distinct from the Guangzhou or a pan-Guangdong variety. Hence the pre-war account is too primitive and outdated to be regarded as representative of the current Hong Kong Cantonese variety. Serious acoustical investigation into the dynamical nature of the pitch height relationship among the lexical tones are crucial but are lacking. Efforts of linguists and phoneticians [4, 5, 6] have been superficial or erroneous for engineering considerations. In another work in preparation, we have discussed the problems of those impressionistic accounts in terms of musical scale. We carry on to examine the validity of the popular and traditional description based on Chao [1] by analyzing the F₀ relationship among the four level tones of Hong Kong Cantonese.

2. Cantonese Tone system

Cantonese is a major southern Chinese dialect spoken in Hong Kong, Guangdong province and many overseas Chinese communities in Southeast Asia and English-speaking western

countries. Depending on the way one defines the lexical tone, Cantonese may be seen as having six or nine lexical tones. In the tradition of Chinese linguistics, the three ‘entering’ (‘checked’ or ‘clipped’) tones, which end with an unreleased /p/, /t/ or /k/, are listed as separate tones --- T₇, T₈ and T₉, in addition to the six non-entering tones (Table 1). In the framework of western, classical phonetics [7], the six-way opposition was preferred since the three entering tones are not in minimal distinctive opposition with the others due to the additional final phoneme. We will focus on the level tones since they also define the starting and ending points for the remaining two rising tones. The issue of interest is --- how should their perceptual distances be quantitatively represented?

Chinese Character	Tone number	Phonetic transcription	Pitch pattern	Tone-letter notation [1]
詩	T ₁	/si/	high level	55
史	T ₂	/si/	high rise	35
試	T ₃	/si/	mid high level	33
時	T ₄	/si/	low level	21
市	T ₅	/si/	low rise	13
事	T ₆	/si/	mid low level	22
色	T ₇	/sikʷ/	high stop	5
not existing	T ₈	/sikʷ/	mid high stop	3
食	T ₉	/sikʷ/	mid low stop	2

Table 1: Cantonese tone system.

3. Conventional representations

Two main ways of quantitative representation [1, 7] of lexical tone patterns can be found in the literature but only one has survived till the present --- the tone-letter notation, which was later adopted by IPA to represent tone contours. When Chao introduced this five-point scale, he illustrated its use by representing the six Cantonese tones as 55 (or 53), 35, 33, 11 (or 21), 23, 22 (Table 1). To focus our discussion on the level tones only, the four levels are arranged in an ascending order of pitch --- T₄, T₆, T₃, T₁, and are denoted by a sequence of four single tone-letters in square brackets: [1 2 3 5]. To interpret Chao’s five-point scale, one has to ask how the level-tone sequence is related to the actual frequency. Since the five-point scale resembles a musical scale and the latter is defined by a strict set of discrete frequencies, it is worth looking at the second means of representation.

Already back in 1912, Jones used the musical notation to indicate the pitch height of Cantonese tones. The four level tones fit into a major scale of [d r m s] in the solfège syllable. If one is ‘deaf’ to the accidentals (sharps and flats), Jones’ musical representation can be transformed to [1 2 3 5]. It is not clear whether it was merely a coincidence or whether Chao simply took over Jones’ description. But at a point, he

mentioned that the pitch range between the highest and lowest tone is approximated better by an augmented fifth [2], which is one semi-tone higher, than a major fifth. Different cultures use different musical scales and develop different sense of musical distance. It is not possible to talk about perceptual distance objectively without a universal scale. For this purpose, we have to adopt a well-tempered scale. Unlike a musical scale (which is a static scale), however, we should not take the scale steps as absolute because the pitch range between any two level tones is never fixed. To illustrate, consider the frequency of an instance of our level-tone sequence $[f_1, f_2, f_3, f_4]$. If they represent a static scale, f_2 can be defined after f_1 is fixed. Chao's five-point notation or Jones' musical notation are definitely wrong if interpreted as a static scale. Mathematically, they make sense only when we treat their illustrations as instances of a more general, dynamic scale. With reference to a well-tempered dynamic scale, the pitch range is not pre-defined to fall between 1 and 5. Jones' level-tone sequence can be re-adjusted to [1 2 3 4.5]. Hashimoto [8] gave both representations but there were inconsistencies between them. Anyway, her impression of the level tones was rendered [1 3 4 5]. According to our native Cantonese author, the two steps between 3 and 5 in Chao's version sound unnecessarily wide or too melodious, like some foreigners learning Cantonese overdo or over-idealize the scale step. In his view, [1 2 3 4] would suffice. It is the aim of the present acoustical analysis to examine which of the various versions of subjective representation serves as the closest approximation or whether the conventional representation has to be revised and replaced by a new scheme.

4. Mathematical formulation

The frequency of the successive notes of a well-tempered scale can be derived from the formula $f = AR^n$, whereby A is an arbitrary pitch, n is an integer and R the m -th root of 2 for m -th equal temperament. Across pitches generated by consecutive integers n 's, equal pitch distances are perceived. In western music, m is equal to 12. To generalize the mathematical form to represent a dynamic scale, both A and R are arbitrary. Let the frequencies for any utterance of T_4, T_6, T_3 and T_1 be f_1, f_2, f_3 and f_4 respectively. They can be expressed as

$$f_1 = ar^{n_1} \quad (1a)$$

$$f_2 = ar^{n_2} \quad (1b)$$

$$f_3 = ar^{n_3} \quad (1c)$$

$$f_4 = ar^{n_4} \quad (1d),$$

where a, r and the n 's are arbitrary. Taking the ratio of each successive pair of equations and then the logarithm, the following differences in exponents of r between successive equations are obtained:

$$n_2 - n_1 = \log(f_2/f_1)/\log(r) \quad (2a)$$

$$n_3 - n_2 = \log(f_3/f_2)/\log(r) \quad (2b)$$

$$n_4 - n_3 = \log(f_4/f_3)/\log(r) \quad (2c)$$

Hence, the ratios of native perceptual distance between consecutive tone levels can be expressed as $(n_2-n_1):(n_3-n_2):(n_4-n_3)$. Since r is arbitrary, it can be chosen in such a way that

$(n_2-n_1)=1$. The perceptual distance ratios of Cantonese level-tones become

$$1 : \frac{\log(f_3/f_2)}{\log(f_2/f_1)} : \frac{\log(f_4/f_3)}{\log(f_2/f_1)} \quad (3)$$

Now that we have assigned 1 to the step distance between the lowest and the second lowest level, assigning the lowest level as 1 ($n_1=1$) will make the second lowest level become 2 ($n_2=2$), thus conforming to the first two steps of Chao's scale. From the above ratios, the remaining two level tones can be easily deduced. The level-tone sequence can be obtained as

$$\left[1 \quad 2 \quad \left(2 + \frac{\log(f_3/f_2)}{\log(f_2/f_1)} \right) \quad \left(2 + \frac{\log(f_3/f_2)}{\log(f_2/f_1)} + \frac{\log(f_4/f_3)}{\log(f_2/f_1)} \right) \right] \quad (4)$$

We can check to which representation the level-tone sequence conforms to --- the notorious and the most concerned [1 2 3 5] of Chao [1], [1 2 3 4.5] generalized from Jones [7], [1 3 4 5] of Hashimoto [8] or [1 2 3 4] of the present author. To visualize the comparison of different theories, they are plotted as two-dimensional paths in Figure 1. Measurement results can be plotted against such a background for comparison.

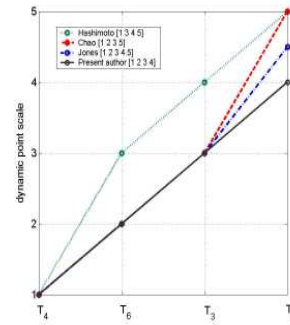


Figure 1: Visualization of perceptual distance of level-tone steps.

5. Experiment

The experiment aimed at extracting the pitch height of the four level tones of Cantonese through recording repetitive production by native speakers.

5.1. Methodology

The horizontal nature of the level tones manifests itself more clearly in certain speech contexts when a native speaker tends to lengthen the syllables, for instance, when one shows a small child or foreigner what a character should be pronounced or when one corrects the wrong pronunciation uttered by a learner. In our experiment, we tried to determine the fundamental height of level tones by asking native speakers to intentionally lengthen their utterance of target characters.

5.2. Materials

The target Chinese characters of the four level-tone classes were taken from the /si/ series (Table 1) and written horizontally from left to right on two cards --- 時 事 試 詩 on Card 1 and 詩 試 事 時 on Card 2. They correspond to the ascending order and descending order of pitch levels --- $T_4, T_6,$

T_3 , T_1 and T_1 , T_3 , T_6 , T_4 respectively, which resemble musical chords or scales.

5.3. Subjects

Two native speakers of Cantonese --- one female and one male (henceforth subject F and subject M respectively), aged between 30 and 40 years, participated in the recording. They were born and have lived their whole life in Hong Kong. The two subjects were both advanced, half-professional musicians and have an accurate ear for musical scales. A preliminary aptitude test was carried out to make sure that they have a good aptitude for lexical tone recognition. This is not trivial because Hong Kong Chinese generally never learn about the tonal aspect of their own vernacular at school.

5.4. Recording

The subjects were asked to read aloud the sequence of four Chinese characters on both cards repetitively for 15 cycles. Each cycle consisted of one up-scale and one down-scale run of citation of the four level tones corresponding to card 1 and card 2 respectively. The citation was requested to be loud and clear in a very slow rate. Each run of up-scale and down-scale citation, and vice versa, was separated by a short pause of about one to two seconds before the subjects were signaled to resume reading. Subjects were instructed to lengthen the syllable as much as they could, as if they were teaching a learner or a child to read. They were asked to pay attention to the pitch level of each Chinese character and try to sustain it. For each run of citation, the speakers were free to choose any starting pitch for the first character of the level-tone scale. The subjects were also asked to try to maintain a uniform intensity. Before recordings started, a few trials were practiced till the speakers felt confident to do so. The recording took place in a normal apartment, in a relaxed environment so that it would not sound too formal.

5.5. Data

From each subject, 15 up-scale and 15 down-scale runs of citation were obtained. By using of a computer program written in Matlab, the four tokens of each run was segmented and then their F_0 and intensity were extracted. A total of 120 tokens of all the four tones were obtained. For our analysis, only those runs with tokens which exhibited broad and horizontal plateaux or terraces of F_0 contour were selected. The selection criterion was based on the requirement that the frequency level of the representative pitch level, which was determined visually and manually by setting a horizontal line against the contour, should not deviate by more than a semi-tone over half of the whole duration. This criterion is only preliminary for the present work and has to be improved in the future. Consequently, a total of 21 runs of level-tone sequence --- 10 up-scale and 11 down-scale, were selected from subject M whereas a total of 22 runs of level-tone sequence --- 11 up-scale and 11 down-scale, were chosen from subject F. Altogether, 84 and 88 tokens of citation tones were obtained.

5.6. Results

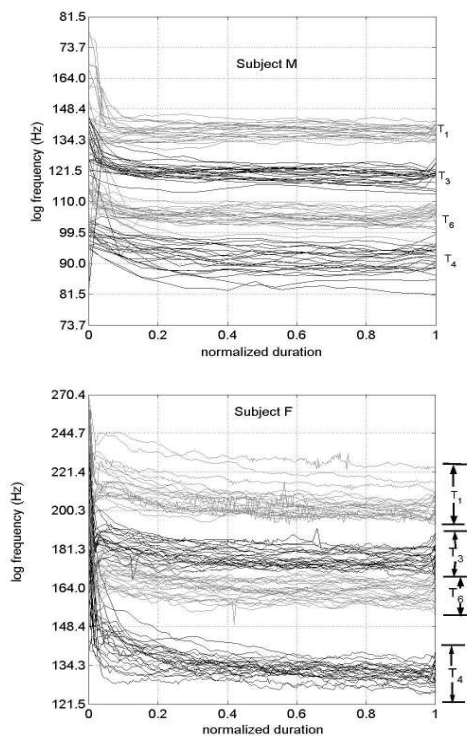


Figure 2: Pitch contours of all level-tone citations.

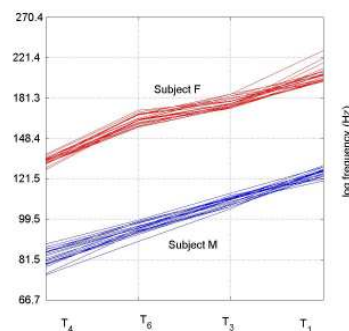


Figure 3: Pitch height development of all citations.

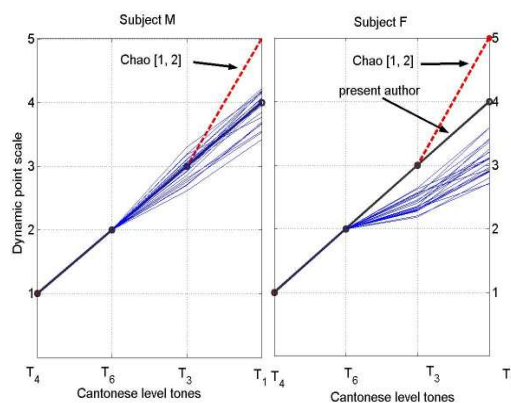


Figure 4: Visualization of tone-level sequence.

To compare the pitch contours of all the tokens, duration is normalized to unity, from 0 to 1. Figure 2 shows that all contours of the same tone class concentrate in distinct bands, which are represented by alternating grey and black curves. To differentiate the contribution by individual runs of tone sequence, the representative frequency of each tone contour is plotted and linked to the consecutive tones of the same run by straight lines, as collectively shown in Figure 3, for both subjects. These curves correlate closely with equation (2). The results corresponding to equation (4) are shown in Figure 4.

5.6. Analysis

For subject M, all the bands are sharply separated by gaps of uniform width whereas for subject F, T_6 - and T_3 -band almost touch one another but can still be visually discriminated. Theoretically, the bands need not be non-overlapping, as the subjects were not explicitly asked to start each run of recording with similar frequency of previous runs. It was not clear whether the subjects had deliberately chosen or were inclined to do so. However, extensive repetition of up and down-scale citation was probably one of the most monotonous tasks that would automatically discourage speakers from experimenting with change of degrees of formality and the associated pitch range. Thanks to that, the non-overlapping band structure allows us to observe relatively easily the average and dispersion of tone contour. One can see that in both cases, the T_1 - T_3 band gap is comparable to that between T_6 - and T_4 -band, which disagrees with most impressionistic accounts in the literature. More importantly, one can tell immediately that the average distance between T_1 - and T_3 -band is by no means twice as that between T_6 - and T_4 -band, as Chao's impression [1, 2] suggested. The three-sectional staffs in figure 3 tell us several things. First, they show the pitch ranges from T_4 to T_1 of the subjects: 75 to 130 Hz for subject M and 127 to 230 Hz for subject F. The steepness of each staff corresponds to the perceptual distance between successive tones. For subject M, the pitch distance between successive tones was maintained uniformly. For subject F, the pitch distance between T_6 and T_3 drops markedly. By normalizing the first staff to 1, those in figure 4 show that the male subject's level-tone sequence path concentrates around [1 2 3 4] whereas those of the female subject deflect to lower values. The perceptual ratios (equation (3)) for both subjects fall short of 1: 1: 2 considerably. Nor are they closer to 1: 1: 1.5 than to 1: 1: 1. However, her T_3 - T_1 slopes are near to that of T_4 - T_6 . A reset and lowering of key signature could have occurred at the T_3 - T_1 boundary whereby the subject tended to divide the whole phrase into two units. Depending on the degree of formality, subjects may over- or under-discriminate the pitch distance of level-tone scale. More experiments are needed to look into such phenomena and disturbances.

6. Discussion

In our preliminary experiment, only subjects with a good sense of and training in music were used because producing sustained, idealized citation forms of level tones which are contrasted mainly by pitch levels requires a skill similar to singing. It is well known in speech experiment that, speakers perform better with perception than with production in general. The ordinary native speakers can speak with the right tune in continuous speech but are not used to recall their speech in an idealized way as if to 'teach' a foreigner or a child.

Despite the small number of subjects and the special criteria, this first acoustical evidence has more weight than its statistics can tell in support of our suspicion that the traditional assumption of the perceptual distance among Cantonese level tones are mistaken, since the traditional wisdom is only based on non-native impression whereas the previous acoustical investigations have used a wrong scale in talking about perceptual distance [5, 6], and engineering applications have not bothered to query its validity [9, 10]. Even before we could confirm our hypothesis with a large statistics, our theoretical consideration has served to query the general usability of tone-letters across languages and dialects. Eventually, the IPA might need to review this mathematically and musically ill-defined tool seriously. Before long, a common need is perhaps to rethink the belief that T_3 is located at the middle of the Cantonese tonal pitch scale [4, 5, 9, 10].

7. Acknowledgements

This work is jointly supported by Japan Society for the Promotion of Science (JSPS) and Swiss National Science Foundation under the project of "Experimentelle Quantenlinguistik I" through the programme "Stipendien für Angehende Forschende", and in part by Waseda Univ. RISE research project of "Analysis and modeling of human mechanism in speech and language processing" and Grant-in-Aid for Scientific Research B-2, No. 18300063 of JSPS. We also thank Cross-Culture Chinese Communications Centre for additional subsidies.

8. References

- [1] Chao, Y.-R. "A system of tone letters", *Maitre Phonétique* 45: 24-27. 1930.
- [2] Chao, Y.-R., *Cantonese Primer*, Greenwood Press, New York, 1947.
- [3] Ho, R. S.-C., "English teaching and learning in Hong Kong", *International Cooper Series on English and Literature*, Vol. 10, Schwabe Verlag, Basel, 2005.
- [4] Bauer, R. S. and Benedict, P. K., *Modern Cantonese phonology. Trends in linguistics. Studies and monographs* 102. Mouton de Gruyter, Berlin, 1997.
- [5] Fok, C.Y.-Y., *A Perceptual Study of Tones in Cantonese*. Centre of Asian Studies, University of Hong Kong, Hong Kong, 1974.
- [6] Vance, T. J. "Tonal distinction in Cantonese", *Phonetica*, 34: 93-107, 1977.
- [7] Jones, D. and Woo, K.T., *A Cantonese phonetic reader*, University of London Press, London, 1912.
- [8] Hashimoto, O. K.Y., *Studies in Yue dialects. 1. Phonology of Cantonese*, Cambridge University Press, Cambridge, 1972, p. 92, 122.
- [9] Gu, W., Hirose, K. and Fujisaki, H. "Analysis of F_0 contours of Cantonese utterances based on the command-response model", *Proc. INTERSPEECH 2004*: 781-784, 2004.
- [10] Li, Y.J., Lee, T. and Qian, Y., "Acoustical F_0 analysis of continuous Cantonese speech", *Proc. ISCSLP 2002*: 127-130, 2002.