Three-sectional-staff characterization of Cantonese level tones

Rerrario Shui-Ching Ho1,2, Yoshinori Sagisaka1
1 Global Information and Telecommunication Institute, Waseda University, Tokyo, Japan
2 Enghelbs Seminar, Universität Basel, Switzerland

Abstract

We have introduced a new approach of acoustical characterization of the level tones of a tone language [1] and tested it with Cantonese [2]. To examine the validity of any perceptual account by analysing its correlation with acoustic data, our new approach requires it to be expressed in terms of the pitch distance ratios between the successive levels instead of representing the pitch levels directly by any system of notation. In this paper, we continued to apply the new approach to search for the existence of a general quantitative relationship governing the pitch distances between each Cantonese level tone. We gathered more recordings and modified the elicitation process of the pitch height of all the four level tones [2] by breaking down the single process for all tones into two sub-processes to avoid the problem of pitch reset.

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

1. Introduction

Despite the strong parallels between music and speech [1], reading the ‘key’ and ‘notes’ of speech melody remains underdeveloped. Whereas description of intonation patterns has been largely primitive (compared to the objectivity achieved in music), accounts of canonical tone patterns have been predominantly non-native, out-dated and subjective [1, 2].

The tone-letter notation [3], which is popularly employed by linguists and on which not few engineering assumptions are still based, looks like a musical short-hand but is actually musically ill-defined and mathematically unformulated [1, 2]. As fundamental frequency is the main acoustic correlate of speech melody, it is of primary importance to pin down the F0 contour pattern mathematically, not only for implementation of automated recognition and synthesis of natural speech, but also for an objective description and characterization for linguists and on which not few engineering assumptions are still based, looks like a musical short-hand but is actually musically ill-defined and mathematically unformulated [1, 2]. As fundamental frequency is the main acoustic correlate of speech melody, it is of primary importance to pin down the F0 contour pattern mathematically, not only for implementation of automated recognition and synthesis of natural speech, but also for an objective description and characterization for linguistic application, for instance, to explain why and how foreigners’ accented speech is perceived as out of tune by native speakers. In our recent work [1, 2], we took up this issue with an aim of developing a more general speech melody scale. We began by considering lexical tones because they are highly constrained both in the pitch and temporal dimension.

Level tones received our first attention because, being distinguished essentially by pitch height, they are less complex than contour tones. Cantonese is of particular interest because it is rich in level tones — four out of her six tones are essentially level. Drawing on native reflections on the basic mechanics of the specific level-tone scale, a new approach of characterization of a level-tone scale system was proposed and demonstrated. However, as further experiments also showed, the pitch contrast elicited through recitation in succession tended to suffer from pitch reset. In this paper, we modified the design of our initial experiment and gathered more realistic data in order to show that the new approach is feasible, and to search for the mathematical relationship governing the pitch distances among the four level tones.

2. Theory

2.1. 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 [4]. Not only is it rich in level tones but also, the level tones are more level than those found in other tone languages which are regarded as level. Depending on definition, 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 --- T5, T6 and T9 (Table 1 of [2]), in addition to the six non-entering tones (Table 1). Since the conventional tone numbers do not correspond to their natural order of pitch levels, for convenience of discussion, we rename the four level tones T6, T9, T3 & T1, in their ascending order, as T1, T6, Tii & Ti, respectively. Since the Cantonese level tones are essentially level and probably more so than many found in other languages and Chinese dialects, they form a system comparable to a musical scale — henceforth Cantonese level-tone scale (CLTS). The issue of interest here is --- how should their perceptual pitch distances be quantitatively represented?

<table>
<thead>
<tr>
<th>Chinese Character</th>
<th>Tone number</th>
<th>Phonetic transcription</th>
<th>Pitch pattern</th>
<th>Reordered tone number</th>
</tr>
</thead>
<tbody>
<tr>
<td>譁</td>
<td>T1</td>
<td>/si/</td>
<td>high level</td>
<td>Tiv</td>
</tr>
<tr>
<td>譴</td>
<td>T2</td>
<td>/si/</td>
<td>high rise</td>
<td>Tii</td>
</tr>
<tr>
<td>譞</td>
<td>T3</td>
<td>/si/</td>
<td>mid high level</td>
<td>Tiii</td>
</tr>
<tr>
<td>譸</td>
<td>T4</td>
<td>/si/</td>
<td>low level</td>
<td>Ti</td>
</tr>
<tr>
<td>譲</td>
<td>T5</td>
<td>/si/</td>
<td>low rise</td>
<td>Tii</td>
</tr>
<tr>
<td>譳</td>
<td>T6</td>
<td>/si/</td>
<td>mid low level</td>
<td>Ti</td>
</tr>
</tbody>
</table>

Table 1: Cantonese tone system.

2.2. New approach

Drawing on native reflections on the basic mechanics of the specific level-tone scale and contrasting it with the mathematical relationship governing the set of notes defining a musical scale, we introduced the key notions of ‘second-order relativity’ to capture the dynamic nature --- hence ‘dynamic scale’, of the level-tone scale. By contrast, a musical scale is regarded as a ‘static scale’ by virtue of its fixed pitch intervals. Both a musical scale and level-tone scale share the basic property that the starting note is arbitrary. Unlike that in music, where the second note is fixed after the first one is, a level-tone scale is ‘dynamic’ in the sense that the second tone
cannot be pinned down even the first one. In other words, the interval between the first two tones is to a large extent arbitrary. The simplest mechanics which can be conceived of is that the third tone (the second interval) depends on the first two tones (the first interval). To capture this simplest dynamics of a level-tone scale, a mathematical framework has been formulated by generalizing the mathematical form governing the mechanics of a musical scale [1, 2]. With this mathematical framework, a new approach of characterization of a level-tone scale system was proposed. In general, for a tone language system with \( n \) level tones, the approach consists in describing the ratios of the \( n-1 \) level triad pitch distances between the successive levels instead of giving the absolute values of pitch intervals, let alone pitch height. An \( n \)-level system is generated and hence characterized by \( n-1 \) instead of \( n \) parameters. Graphically, an \( n \)-level system can be represented by an \( (n-1) \)-sectional staff.

Applying it to Cantonese, the pitch height of the four level tones is generated by the ratios of their three pitch distances, represented graphically by a three-sectional staff [1, 2]. Given the frequencies for any native utterance of \( [\text{T}_{i}, \text{T}_{ii}, \text{T}_{iii}, \text{T}_{iv}] \), the pitch height of the four level tones can be denoted by the Chao’s scale, the remaining two level tones can be deduced.

Assigning 1 (\( n = 1 \)) to the lowest level and 2 (\( n = 2 \)) to the second lowest level, thus conforming to the first two steps of Chao’s scale, the remaining two level tones can be deduced. The pitch height of the four level tones can be denoted by the sequence

\[
1: \log \left( \frac{f_a}{f_{a+a}} \right), \log \left( \frac{f_a}{f_{a+b}} \right)
\]

Within the new mathematical framework, the conventional linguistic impressionistic accounts based on both the musical notations and tone-letter notations [3, 5, 6] can be re-interpreted. With the new approach of characterization, we can proceed to conduct acoustical measurement and search for, if there is any, the mathematical relationship governing the pitch distances among the four level tones and test the validity of the conventional accounts.

### 2.3. Problem

The most straightforward way to elicit the pitch height of the four levels tones in question is from native informants’ simple recitation of four corresponding syllables in succession. In actual recording, there are always imperfections. In our last experiment, alternating down-scale \( ([\text{T}_{i}, \text{T}_{ii}, \text{T}_{iii}, \text{T}_{iv}]) \) and up-scale \( ([\text{T}_{i}, \text{T}_{ii}, \text{T}_{iii}, \text{T}_{iv}]) \) recitations were intended to neutralize down-drift. However, the problem of pitch reset cropped up, as confirmed by additional experiments with some other informants. This happens very often at the starting syllable in the second half of a sequence, namely that of \( \text{T}_{i} \) in a down-scale and \( \text{T}_{ii} \) in an up-scale recitation. This could be probably due to an unconscious imposition of a 2-2 rhythm to the four-syllable sequence, as it often happens in the act of counting and enumerating. This led to magnification or shrinkage of the pitch interval between \( \text{T}_{ii} \) and \( \text{T}_{iv} \). In addition, we have tested our new approach with only two musically trained native speakers. It is desirable to test it with more normal speakers and with slight modification of the experiment.

### 2.4. Methodology

If pitch reset is caused by insufficient capacity of temporary memory or focus, reducing the number of tones to be recited in a row to three might help avoid the problem. Also, three is the smallest working number since two pitch levels do not pin down the tonal mathematics of a minimal tonal opposition. We stayed with sequences of level tones in pitch order, believing that it would be less demanding for informants to focus on two tones of consecutive pitch levels rather than jumping across a large pitch distance. Out of four level tones, only two possible consecutive triads can be constructed --- the upper triad \( [\text{T}_{iv}, \text{T}_{iii}, \text{T}_{ii}] \) and the lower triad \( [\text{T}_{iii}, \text{T}_{ii}, \text{T}_{i}] \). We kept working with alternating forward (up-scale) and backward (down-scale) recitation to minimize undesirable pitch down-shift or up-shift. Instead of eliciting the canonical pitch distance ratios (1) directly from alternating recitation --- \( [\text{T}_{iv}, \text{T}_{iii}, \text{T}_{ii}, \text{T}_{i}] \) and \( [\text{T}_{i}, \text{T}_{ii}, \text{T}_{iii}, \text{T}_{iv}] \), we decomposed the elicitation into two recitation tasks. First, the pitch distance ratio

\[
r_{i,i,i} = \frac{\log(f_{a+a})}{\log(f_{a})}
\]

was elicited from alternating recitation of \( [\text{T}_{iii}, \text{T}_{ii}, \text{T}_{i}] \) and \( [\text{T}_{i}, \text{T}_{ii}, \text{T}_{iii}] \). Second, the pitch distance ratio

\[
r_{i,i,i} = \frac{\log(f_{a+a})}{\log(f_{a})}
\]

was elicited from alternating recitation of \( [\text{T}_{iv}, \text{T}_{iii}, \text{T}_{ii}] \) and \( [\text{T}_{i}, \text{T}_{ii}, \text{T}_{iii}] \). Subsequently, \( d_{i,i} : d_{i,i} : d_{i,i} \) can be derived from the two ratios given by (3) and (4), as thus

\[
1 : r_{i,i,i} : r_{i,i,i} : r_{i,i} \quad (5)
\]

whereas the level-tone sequence (2) can be expressed as

\[
1, 2, \left(2 + r_{i,i} \right), \left(2 + r_{i,i} + \frac{r_{i,i}}{r_{i,i}} \right) \quad (6)
\]

### 3. Experiment

The experiment aimed at eliciting the simple mathematical relationship, if any, governing the pitch intervals between successive levels of the four Cantonese level tones.

#### 3.1. Materials

Four target Chinese characters --- 詩 試 事 時 --- corresponding to the four level tones --- \( \text{T}_{i}, \text{T}_{ii}, \text{T}_{iii}, \text{T}_{iv} \) were taken from the /si/ series (Table 1). Two prompt cards were prepared. On each card, three characters were written horizontally from left to right --- 詩 試 事 on Card 1 and 詩 事 時 on Card 2. They correspond to the upper triad --- \( \text{T}_{iv}, \text{T}_{iii}, \text{T}_{ii} \) and the lower triad --- \( \text{T}_{iii}, \text{T}_{ii}, \text{T}_{i} \) respectively, both in descending order of pitch.
3.2. Informants

6 native speakers of Cantonese in their forties --- 3 females (F1, F2 & F3) and 3 males (M1, M2, M3), were invited to participate voluntarily in the recording. They were born, educated and have lived their whole life in Hong Kong except for occasional short-term vacations. None of them were known to have any speech and hearing defect, nor do they carry any noticeable accent, nor display significant personal deviation in speech style.

3.3. Recording

The basic task for the informants were to utter aloud the sequence of three Chinese characters on a card, alternating in forward and backward order. Informants were to repeat the cycle for four to six times until the investigator gave gesture to stop. This constituted one block and a few blocks were repeated. The recording for each informant was divided into two sessions --- session 1 and session 2. Card 1 and 2 were used for session 1 and 2 respectively. Each session consisted of 4 to 6 blocks interrupted by pauses of around 30 seconds. In each block, recitation consisted of four to six uninterrupted cycles. Each cycle comprised a down-scale (descending) run and an up-scale ascending run of the corresponding triad --- [Tiv Tiii Tii] followed by [Tii Tiii Tiv] in session 1 and [Tii Tiii Ti] followed by [T i Tii Tiii] in session 2. The recitation was not intended to be as slow as that in the last experiment [2]. Short demonstration was given at the beginning to show the desired tempo. The informants were asked to try to maintain a uniform loudness. They were also instructed to sustain each syllable rather than to utter a staccato form in order to avoid significant drop in pitch level. But during the course of recording, the informants were not interrupted even their performance did not meet the expectation of the investigator. The informants were free to choose any starting pitch to start a block or cycle. Before recording started, a few trials were practiced till the speakers felt confident to do so. The recording took place in a casual room in a normal apartment in Hong Kong.

3.3. Data processing

From each informant, recordings consisting of 10 to 18 cycles of down-scale and up-scale recitation of the upper and lower triads were collected (Table 2). The triads were first segmented into separate tone tokens. For each token, the initial 10% and final 10% of the whole segment was cut off. The mean pitch and standard deviation was calculated for each truncated segment in the log scale. The truncated segments were then subjected to ‘levelness’ assessment based on two criteria. They were selected for pitch distance computation only when both criteria were fulfilled. The first criterion required that the standard deviation should not exceed 0.5 whereas the second one required that the pitch range between the maximum and minimum should not exceed 2 semitones.

Table 2 shows the total number of raw and accepted triads for informants F1, F2, M1 and M2. Because a greater proportion of the tokens of informant F3 and M3 were unusable according to the levelness requirement, their data were not shown. The normalized F0 contours of all the accepted truncated patterns are plotted in Figure 1. To compare the pitch contours of all the tokens, truncated segments are normalized to unity, from 0 to 1.

3.4. Analysis

Only the tokens which were regarded as level (based on the two criteria) enough were passed on for the last step of computational analysis. Pitch distance ratios for the upper triads and lower triads were calculated according to (3) and (4). The average values of the pitch distance ratios were shown in Table 3. The pitch distance ratios for the CLTS were derived using (5). The four three-sectional staffs were visualized in Figure 2.

Table 3: Pitch distance ratios calculated from the recordings of four speakers (F1, F2, M1, M2)

Figure 1: Pitch contours of all level-tone tokens.

Figure 2: Three-sectional staff visualization of the tone-level sequence.
4. Discussion

It is perhaps instinctual for us human beings to wish that the pitch distances between four successive level tones, other things being equal, are governed by simple integer ratios. Among all literature accounts, re-interpreted under our new framework, \( r_{iiii} = 1:1 \) is perhaps one of the least contested ratio of two pitch distances. However, our results for the lower triad (Table 3) are significantly far from this ideal. The nearest one, that of \( F_1 \) (the value 0.91 in Table 3), differs by almost 10% whereas the others differ from 17% to 33%. Nevertheless, 1:1 is still the nearest simple ratio compared with 1:2, 2:1, 3:2, and 2:3. Since we have to allow for at least one level for the sake of \( T_n \), consideration of further integer ratios will lead to division of our perceptual pitch range into more than 7 levels, which is impractical. Concerning the upper triad, the results --- 1.76, 1.79, 1.49, 1.35) spread out roughly around 1.6. In a similar fashion, 2:3 would be the best but very rough simple-ratio approximation for \( r_{iiii} \). Neither the most popular 1:2 nor our speculated 1:1 are real rivals. On the whole the results of the present experiment do not point to the emergence of any underlying simple ratios. We need to analyse a large set of data to search for it.

At this state of our knowledge, we are still quite ignorant about how level a level tone must be to be perceived as 'natively level'. Yet we can justify our present levelness criteria by looking at some deviation figures. In terms of stability and consistency, with its deviations lying between 0.1 and 0.3 semitone, \( M_1 \) showed the best performance. However, \( M_2 \) was the one who produced most level tokens consistently. The deviations of his accepted tokens were among the lowest. Both the 8 down-scale runs and the 8 up-scale runs rejected from the lower triad (Table 2) were only due to the use of the low falling variant of \( T_1 \) and his personal style of using a rapid falling pitch for \( T_n \) located in the end of the up-scale triad \([T_1 \ T_2 \ T_3 \ T_4] \) which is also the end of a cycle \([T_n \ T_1 \ T_2 \ T_3 \ T_4] \). The existence of such rejected data of his, as well as others', were actually a good sign rather than a bad one because they suggested that the utterance of the accepted level tones were not unrealistically level.

Despite the lack of calibration of the rejection criteria, much can be learnt at the moment from pondering on the parallels found in music. It is not difficult to imagine that the deviation associated with singing a musical note has to be as demanding as not to blur the distinctiveness of two notes separated by just a semi-tone. Like music, a lexical tone will parallels found in music. It is not difficult to imagine that the deviation associated with singing a musical note has to be as demanding as not to blur the distinctiveness of two notes separated by just a semi-tone. Like music, a lexical tone will need to be realistically expressed in such a way that permits acoustical verification based on production. Based on this principle, our new approach requires a perceptual account to be expressed in terms of the pitch distances ratios between the successive tone levels instead of denoting the pitch levels directly. Thus, it queries the validity of linguists' conventional accounts and the general usability of the tone-letter notation [3], which the IPA has to review seriously.

5. 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. 18300603 of JSPS.

6. References