ACOUSTICAL F0 ANALYSIS OF CONTINUOUS CANTONESE SPEECH

Yujia LI, Tan LEE and Yao QIAN
Department of Electronic Engineering
The Chinese University of Hong Kong, Hong Kong
{yjli, tanlee, yqian}@ee.cuhk.edu.hk

ABSTRACT
This paper presents a preliminary study on acoustical analysis of fundamental frequency (F0) in continuous Cantonese speech. By understanding how the surface F0 contour is determined by many co-functioning and inter-playing linguistic or non-linguistic factors, our ultimate goal is to facilitate automatic F0 prediction for highly natural text-to-speech synthesis. A novel method of F0 normalization is proposed to effectively reduce the undesirable fluctuation of the speaker’s F0 range. Statistical analysis is performed on the normalized F0 contours. Specifically, our investigation is focused on: (1) F0 movement over intonation phrases; (2) tone contours in continuous speech; (3) effect of tonal context; and (4) co-articulated tone contours in disyllabic words.

1. INTRODUCTION
For the generation of highly natural synthetic speech, proper control on the fundamental frequency (F0) contour is of primary importance. F0 contour is one of the major acoustical manifestations of supra-segmental features such as tone, pitch accent and intonation. These features are all critical to perceptual naturalness of human speech. In tonal languages like Mandarin or Cantonese, different tones are associated with the same syllable pronunciation to express different lexical meanings. In this case, the importance of F0 contour is not only on naturalness but also on delivering the content correctly.

The surface F0 contour in natural speech is determined by many co-functioning and inter-playing linguistic or non-linguistic factors. Understanding the effect of each individual factor is very useful to establish an appropriate prosody model for text-to-speech (TTS) synthesis. The methods of F0 analysis can be divided into two major categories, namely acoustical analysis and analysis by synthesis. Acoustical analysis deals with the acoustical measurements directly, focusing on a few examples or a large corpus, so as to study how the F0 contour depends on a particular factor of interest. The approach of analysis by synthesis typically involves a parametric production model that attempts to approximate the observed contour. The optimized parameters in the best approximation reveal the underlying contributions of the respective factors. The Fujisaki model [1] and Stem-ML [2,3] are examples of well-established analysis by synthesis methods.

This paper describes a study on acoustical analysis of F0 in continuous Cantonese speech. It is the first step towards the development of a sophisticated prosody model for our Cantonese TTS system that was reported in [4].

2. TONES IN CANTONESE
2.1 Tone system
Cantonese is a major Chinese dialect. Like Mandarin, spoken Cantonese is seen as a string of monosyllabic sounds. Each Chinese character is pronounced as a monosyllable that carries a specific tone. Each syllable is divided into the Initial part and the Final part. Cantonese is said to have nine citation tones that are characterized by different pitch patterns as illustrated in Figure 1. The so-called “entering” tones occur exclusively with “checked” syllables, i.e. syllables ending in an occlusive coda /p/, /t/ or /k/. They are contrastively shorter in duration than the “non-entering” tones. In terms of pitch height, each entering tone coincides with a non-entering counterpart. In many transcription schemes, only six distinctive tones, labeled by numerals 1 to 6, are used [5].

![Figure 1. Tones in Cantonese: schematic description](Image)

2.2 Acoustical realization
In spoken Chinese, tone is manifested in the F0 movement across the voiced portion of a syllable. In Cantonese, the Final segment can be regarded as voiced while the Initial is either voiced or unvoiced. Figure 2 depicts the average F0 contours of the nine tones produced by a male subject who speaks native Cantonese. The averages were computed over 1,800 monosyllabic utterances that cover most of the tonal syllables used in today’s Cantonese. It is observed that four of the non-entering tones (Tone 1, 3, 4, 6) have flat or slightly declining F0 patterns while the other two (Tones 2 and 5) show different rates of rise of F0. Discrimination among these tones relies much more on the relative height than the shape of F0 profiles.

![Figure 2. F0 profiles of different tones uttered by a male speaker. The dashed lines are derived from the respective entering tones](Image)

3. SPEECH DATABASE
This research is done based on a large speech corpus, named CUProsody, which was developed at the Digital Signal Processing Laboratory of the Chinese University of Hong Kong. The corpus was recorded from a trained female speaker. It is a read-speech corpus that consists of 1300 continuous utterances, among which 1200 are newspaper sentences and the remaining...
contains mostly conversational content. In this research, only the
newspaper sentences are used for our investigation.

4. F0 NORMALIZATION

F0 is a highly variable acoustical feature. Even for the same
speaker’s voice, the actual range of F0 changes from time to time
because of a variety of physical, emotional, semantic or stylistic
factors. Therefore, the absolute F0 values of the lexical tones in
Chinese may not be directly comparable if they are produced
under different context. The goal of F0 normalization is to reduce
the undesirable variation caused by some irrelevant factors. In
the context of this research, such undesirable variation refers to
the change of the speaker’s F0 range from one utterance to another,
and the intonation movement within a long utterance.

Usually, F0 normalization is done by, at each time instant,
dividing the absolute F0 value by a normalization factor. This
factor is expected to be a good indicator of the F0 range at that
time instant. In [7], the normalization factor was computed on per
speaker basis for speaker-independent isolated tone recognition.
To deal with the change of F0 from one utterance to another,
the utterance-wide mean of F0 can be used. In [8], it was proposed to
use a moving window approach to better capture the timely
change of F0 within an utterance.

Here we propose a new method of F0 normalization based on
a properly estimated phrase curve. The following assumptions have
been made:
1) F0 movement over an intonation phrase can be approximately
described by a straight line, referred to as phrase curve [2,3];
2) Although the absolute F0 level of a particular tone may vary
greatly, its relative height with respect to each of the other
tones remains largely invariant. Such invariance is preserved
locally, i.e. between neighboring syllables, because of the
requirement of communication accuracy and the continuous
muscle movement of vocal cords.

The normalization process is divided into three steps: (1)
estimation of the relative tone ratios; (2) derivation of the phrase
curve; and (3) normalization of the absolute F0 value.

Given a pair of neighboring tones (i,j), where i and j denote
the identities of the preceding and the succeeding tones
respectively, their relative ratio of height is computed as,

\[ R(i, j) = \frac{\text{height of tone } i}{\text{height of tone } j} \]

where the height of tone is computed as the mean F0 value of the
respective tone contour. Each tone contour is represented by five
points equally spaced over the tone duration. For the computation
of the tone height, the first and the last points of the contour are
not used. The overall relative tone ratio, denoted by \( \hat{R}(i, j) \), is
obtained by averaging over all occurrences of the respective tone
pair. As a result, a six-by-six matrix \( \{\hat{R}(i, j)\} \) has been obtained
as shown in Table 2.

Table 1. A summary of the CUProsody corpus

<table>
<thead>
<tr>
<th>Overall range of F0: 140-300Hz</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of sentences = 1200; Average sentence length: 66 syllables</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. of intonation phrases = 4973; Average phrase length: 15 syllables</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. of words = 38743; Average word length: 2.05 syllables</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-syllable words 2-syllable words 3-syllable words 4-syllable words</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22.9% 58.6% 11.4% 5.5%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. of syllables = 79528; Average syllable duration: 0.192s</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tone 1 tone 2 Tone 3 Tone 4 Tone 5 Tone 6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25.5% 12.6% 16.1% 17.1% 6.4% 22.6%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Matrix of relative tone ratios

<table>
<thead>
<tr>
<th>i</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.97</td>
<td>1.39</td>
<td>1.28</td>
<td>1.60</td>
<td>1.39</td>
<td>1.35</td>
</tr>
<tr>
<td>2</td>
<td>0.71</td>
<td>0.99</td>
<td>0.92</td>
<td>1.11</td>
<td>0.95</td>
<td>0.97</td>
</tr>
<tr>
<td>3</td>
<td>0.80</td>
<td>1.07</td>
<td>1.02</td>
<td>1.32</td>
<td>1.13</td>
<td>1.13</td>
</tr>
<tr>
<td>4</td>
<td>0.65</td>
<td>0.91</td>
<td>0.83</td>
<td>1.08</td>
<td>1.00</td>
<td>0.94</td>
</tr>
<tr>
<td>5</td>
<td>0.70</td>
<td>0.99</td>
<td>0.93</td>
<td>1.06</td>
<td>1.02</td>
<td>1.01</td>
</tr>
<tr>
<td>6</td>
<td>0.73</td>
<td>1.01</td>
<td>0.95</td>
<td>1.22</td>
<td>1.07</td>
<td>1.05</td>
</tr>
</tbody>
</table>

It is noted that \( \hat{R}(i, j) = \frac{1}{R(i,j)} \). That is, the occurrence
order of the tones doesn’t affect their relative ratio of heights.

Being located in the middle of the tone space, the height of
Tone 3 is selected as the reference for the speaker’s F0 range.
Given an occurrence of Tone \( k \ (k = 1, 2, \ldots, 6) \), we can convert
its F0 height to an equivalent height as if it were Tone 3, by
multiplying with the conversion ratio \( \hat{R}(3,k) \). For example, if the
height of an occurrence of Tone 4 is 150 Hz, the equivalent height
of Tone 3 would be equal to 150Hz \times 1.32 = 198Hz . In this way,
all tones in the utterance are transformed into Tone 3 regardless of
their original identities. The phrase curve is then obtained by
performing linear regression over these converted tone heights.
Figure 3 shows an example of phrase curve.

Lastly, F0 normalization is done by dividing the original F0
contours by the corresponding F0 value on the phrase curve.
intonation phrase. In the previous section, the phrase curve was computed as a straight line that approximates the movement of F0 level of a particular tone class, i.e. Tone 3. The phrase curves computed from all of the 1,200 utterances were analyzed and the results are summarized as in Table 3.

<table>
<thead>
<tr>
<th>Average initial value</th>
<th>Average slope</th>
<th>Average length</th>
</tr>
</thead>
<tbody>
<tr>
<td>218.65Hz</td>
<td>-2.13Hz/syllable</td>
<td>16 syllables</td>
</tr>
<tr>
<td>No. of declining cases</td>
<td>No. of inclining cases</td>
<td></td>
</tr>
<tr>
<td>4128 (83%)</td>
<td>845 (17%)</td>
<td></td>
</tr>
</tbody>
</table>

**Table 3. A summary of the phrase curves**

An utterance is the largest independent unit that the speaker attempted to complete in continuation. It may consist of one or more intonation phrases whose contents are inter-related and structured. About 18.5% of the utterances consist of two phrases, 20.8% consist of three phrases and 15.1% consist of 4 phrases. It is observed that the phrase curve depends on its position in the utterance. Our observations include:

1. Most of the phrases show declining F0. The average slope of phrase curve is -2.13 Hz/syllable. This perfectly agrees with the results that we attained with the Stem-ML approach [9].
2. The first phrase consistently exhibits a greater slope of F0 downshift than the succeeding phrases. For example, for the utterances that contain three intonation phrases, the average slopes of phrase curves are -2.6Hz/syllable, -1.6Hz/syllable and -2.3Hz/syllable respectively.
3. The initial value of the first phrase is significantly higher than that of the others. The difference is about 15-20Hz.

Figure 4 shows the average phrase curve pattern of the utterances that contain four phrases. The reset of F0 is seen clearly at the phrase boundaries.

**6. ANALYSIS OF TONE CONTOURS**

Figure 2 gives the canonical patterns of the Cantonese tones in isolation. In continuous speech, the tone contours become much more complicated and cannot be described sweepingly by these canonical patterns. In this section, we analyze the contextual variation of tone contours.

**6.1 Single-tone contours**

For each of the six tones, an average contour is computed from the normalized five-point contours of all occurrences of this tone. The results are shown as in Figure 5.

Obviously, tone contours in continuous speech deviate greatly from their canonical patterns. In particular, the beginning section of Tone 1 is substantially lowered and those of Tone 2, 4 and 5 are lifted up. A reasonable explanation is the co-articulation caused by the neighboring tones. In Figure 2, Tone 1 is realized with the top height while tone 2, 4 and 5 are initially at the lowest level. In order to connect with the neighboring tones smoothly, the tones change their canonical patterns to compromise.

**6.2 Contextual variation**

To further investigate the tone co-articulation effect, we divide the syllables into four classes, according to their left or right tonal context: (1) Left tone ends at high level (Tone 1 or 2); (2) Left tone ends at low level (Tone 3, 4, 5 or 6); (3) Right tone begins at high level (Tone 1); (4) Right tone begins at low level (Tone 2, 3, 4, 5 or 6). If the syllable lies by a phrase boundary, it is considered to have no either left or right neighbor.

Figure 6 shows the variation of the normalized F0 values at the five sections over the contours. There are several interesting observations that can be made:

1. Tone 1 has consistently greater variation than the other tones at all sections. It is due to that as the highest tone, the acoustical realization of Tone 1 has relatively more freedom (to go upwards) than the other tones.
2. Almost for all tones, the beginning section generally shows much greater variation than the ending section. This may suggest that the tone co-articulation effect from the left context is more significant than that from the right context.
3. Section 3 & 4 have the smallest variation. They are considered to be the most stable parts of a tone contour.

**Figure 5. Single-tone contours in continuous Cantonese speech**

**Figure 6. F0 variation at different sections of tone contours**

**Figure 7. Context dependent tone contours**

Figure 7 shows that the left tonal context introduces great variation to the beginning section of the tone contour. The difference between the high and low contexts is remarkable, from 30 Hz up to 50Hz in terms of absolute F0 value. On the other
hand, with different right contexts, the tone contours look similar in their final sections. The difference is only about 12 Hz.

As a conclusion, Cantonese seems to have a left to right control pattern of F0. Nevertheless, when the right context is a low tone, the contours of Tone 1 and 2 would be slightly lifted up at the end. This is also related to the so-called anticipatory rising in Mandarin [10].

6.3 Co-articulated tone contours of disyllabic words

In this section, we expand our analysis unit to lexical words. In particular, we focus on disyllabic words, which form the majority of the lexical words in Chinese and cover up to 60% of the words in our database. All the disyllabic words with the same tone combination are grouped together and an average F0 contour is computed. Figure 8 shows several examples of co-articulated tone contours of disyllabic words. In Figure 8 (a), the three contours are for the tone combinations 4-1, 4-4 and 4-5 respectively. That is, they differ from each other in terms of the identity of the second tone. Figure 8 (b) shows the tone combinations 1-4, 4-4, 5-4, i.e., having different identities of the first tone. It can be seen that the contour corresponding to the first tone tends to resemble the context-independent single-tone case, regardless of the identity of the succeeding tone. The second tone shows a much severely co-articulated contour. It starts by following the height of the preceding tone and then gradually resumes its own position.

This paper describes a preliminary study on acoustical F0 analysis of continuous Cantonese utterance based on a large read-speech corpus. The two important components of an F0 contour, namely phrase curve and tone contour, are investigated. Declining phrase intonation is consistently observed in our analysis. The average rate of declination is 2.13 Hz per syllable. Tone contours in continuous speech vary substantially with their tonal context. The left context has significantly greater effect than the right context. In a disyllabic word, the realization of the first tone tends to be independent of its succeeding tone, while the F0 contour of the second tone changes a lot with its preceding tone. The continuity of voicing between the two syllables also affects the realization of the second tone.

F0 is computed only for voiced speech. In a Cantonese syllable, the Initial is either voiced or unvoiced. Voiced Initials include /j/, /w/, /l/, /m/, /n/, /ng/ and /h/ and unvoiced Initials include /t/, /d/, /k/, /l/, /m/, /n/, /ng/ and /l/. The Finals are mostly voiced except that some of them end with a stop coda /p/, /t/ or /k/, resulting in the presence of a closure. Therefore, within a disyllabic word, the two tones are either joined directly or separated by an unvoiced segment and/or a closure. We divide the co-articulated tone contours into two groups according to the Final of the first syllable (F1) and the Initial of the second syllable (I2):

Group A – F1 ends with a vowel or nasal coda and I2 is voiced

Group B – F1 ends with a stop coda or I2 is unvoiced

Figure 9 compares the resulted contours from group A and B. Disyllabic words with tone combination 1-4 and 4-1 are selected as the subjects for comparison.

8. REFERENCES

[5] Linguistic Society of Hong Kong (LSHK), Hong Kong Jyut Ping Characters Table ( 粵語拼音字表), Linguistic Society of Hong Kong Press (香港語言學會出版), 1997.