Perceptual Equivalence of Approximated Cantonese Tone Contours

Yujia Li and Tan Lee

Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong

yjli@ee.cuhk.edu.hk, tanlee@ee.cuhk.edu.hk

Abstract

This paper describes a perceptual study on approximated Cantonese tone contours. We believe that the perception of tone contours relies mainly on the major trend of pitch movement, and is not sensitive to the exact F0 values at particular time instants. The tone contours of individual syllables and the transition between them are approximated as a small number of linear movements. The effect of such approximation is assessed by perceptual experiments. It is found that the six Cantonese tones can be represented by one or two linear movements, and the transition between tones can be represented by a single linear movement, without creating noticeable perceptual difference. Such simple approximations are desirable for perception-driven F0 modeling for text-to-speech applications. 

Index Terms: perceptual equivalence, Cantonese tones, prosody modeling.

1. Introduction

Previous research showed that the relation between speech production and speech perception is very complicated. In [1], it was demonstrated that a production-intended rising tone of Mandarin Chinese is perceived as a falling tone, under certain contextual conditions. In [2], it was shown that the change of perceived speaking rate does not mean uniform duration variation over the entire utterance. Stressed syllables are more influential than unstressed ones. Indeed, human perception system may not have a sufficiently high resolution to differentiate every minute acoustic variation. It was found that pitch movement becomes perceivable only if the rate of F0 change is over a threshold [3]. On the other hand, there exist physical constraints on human speech production [4]. It is possible that the perceptually most preferable targets are not found by acoustic measurement and analysis of human speech. In [5], it was shown that a re-synthesized fast speech sounds better than natural-fast speech.

The present study concerns lexical tones in Cantonese. Cantonese is a major Chinese dialect known to be rich in tones. Lexical tone perception is crucial to the intelligibility and naturalness of Cantonese. Phonologically, each lexical tone of Cantonese is described by a specific pitch pattern, which is acoustically represented by the F0 contour over the syllable. In real speech, the F0 contours vary greatly, but the tones can be well perceived. Thus it would be sufficient to approximate the original tone contours in a way that the perceptually most important features are retained. In this work, each Cantonese tone contour is approximated as a concatenation of line segments that describe the syllable-wide trend of F0 movement.

One application of our research is on prosody modeling for speech synthesis. Many existing approaches to prosody modeling are production-driven, i.e., they attempt to precisely regenerate the acoustic targets found in human speech. This may not lead to the best perceptual effect because satisfying human perception is not their primary goal. Our work is intended to be a first step towards perception-driven prosody modeling.

2. Cantonese Tone System

In Cantonese, each Chinese character is pronounced as a monosyllabic sound carrying a specific lexical tone. Each syllable consists of an Initial and a Final part. Cantonese is said to have nine tones described by their distinctive pitch patterns, as shown in Figure 1. “Entering tone” is a historically defined tonal category that was used as a cover term for tones that co-occur with “checked” syllables, i.e., syllables that end in an occlusive coda such as /p/, /t/, or /k/. Entering tones are contrastively shorter in duration but coincide with a non-entering counterpart in terms of pitch level [6]. Many linguistic researchers have suggested treating the three entering tones as abbreviated versions of their non-entering counterparts. In the Jyut Ping system [7] only six distinctive tone categories, which are labeled by the numerals 1 to 6 as in Figure 1, are defined. The six tones are divided into level tones (Tone 1, 3, 4, 6) and rising tones (Tone 2, 5). The pitch patterns shown in the figure are the phonological descriptions of the tones. They can be considered as the underlying targets for speech production to reach, and for speech perception to identify.

Figure 1: Pitch patterns of Cantonese tones.

Tone is acoustically represented by the F0 contour over the voiced portion of a syllable. The acoustical realization of tones in isolation can well resemble the phonological patterns. However, in continuous speech, due to many factors, the realized F0 contours have complicated variations.

3. Experiments and Results

We try to determine the simplest approximation of Cantonese tone contours that can provide equivalent perception to the tone contours measured from natural speech. Our previous studies [8] [9] showed that the pitch height and the syllable-wide pitch movement are the most useful features for Cantonese tone identification. In [10], each tone contour is represented by only three coarsely sampled values. Thus we hypothesize that a linear approximation of the tone contour would be adequate for Cantonese. In the following experiments, only linear approximations are investigated.
3.1.1. Method

Tone contours carried by syllables spoken in isolation can be regarded as the canonical pattern, since isolated syllables are the most independent and intact units on which the required acoustic targets can be approached without much contextual constraint. In this experiment, isolated tone contours are approximated by one or two line segments. Different approximation strategies are used for level tones and rising tones, as shown in Figure 2. They are designed on the basis of both the phonological descriptions in Figure 1 and acoustic observations from natural speech.

For the level tones, two types of approximation are used: level (denoted by $L_a$) and falling (denoted by $F_a$). $L_a$ is the simplest form, for which only one parameter is required. $F_a$ is included in the study because many of the level tones exhibit $F_0$ falling in natural speech. For the rising tones, three different approximations are defined, namely rising ($R_a$), level-rising ($L R_a$) and falling-rising ($F R_a$). Since there are four level tones and two rising tones, a total of 14 approximations will be tested.

3.1.2. Material

Three syllables with voiced initial consonant are used as carriers. There are a total of 18 tonal syllables, since each syllable may carry 6 different tones, as shown in Table 1. A female native speaker uttered each of the syllables three times. Among the three samples, the best one was selected for our experiment. The $F_0$ contour of each syllable sample was extracted. For each applicable type of approximation in Figure 2, the best approximation was determined based on subjective perceptual judgment. During this process, the original speech sample is modified such that its $F_0$ contour follows the approximated version. The approximated contour is considered to be the best if the modified speech gives the closest perception to the natural speech. The above analysis and modification of speech were carried out by using the PRAAT software [11]. For each syllable carrying a level tone, two approximations, i.e., $L_a$ and $F_a$, were generated. For rising-tone syllables, there are three approximations, $R_a$, $L R_a$ and $F R_a$.

3.1.3. Test Procedures

Pair comparison was used in our perceptual tests. Each pair of stimuli contains a modified sample and its corresponding natural sample. There were a total of 42 (14 approximations x 3 carriers) test pairs. They were presented to human subjects through an interactive computer interface. The subject was asked to wear a headphone and listen to each pair of stimuli for up to three times. He/she was then asked to determine the better stimulus. Or if they could not perceive any difference, a choice of “same” was allowed. The 42 stimuli pairs were randomly sequenced, and in each pair, natural and modified sounds also appeared in random order. Ten subjects, who are all university students, participated in the test. They were paid for the workload. For each of the 14 approximations, 30 responses were obtained.

3.1.4. Results

The test results are shown as in Figure 3. Each vertical bar represents the 30 responses for a particular approximation. It consists of three parts that represent the percentage of different perceptual preferences: “preference to modified speech”, “perceptually same” and “preference to natural speech”. For Tone 1 and 3, both $L_a$ and $F_a$ can reach equivalent perception to natural speech. For Tone 4 and 6, only falling approximation seems proper. As for the rising tones, $R_a$ is not appropriate; $F R_a$ gives the most competitive perception. To summarize, to reach equivalent perception, three kinds of linear approximation are needed. For high level tones, tone 1 and 3, $L_a$ is enough; for low level tones, tone 4 and 6, $F_a$ is necessary; for rising tones, tone 2 and tone 5, $F R_a$ seems the best approximation. More generally, all the level tones should be approximated by a single linear movement, while all the rising tones need to be approximated by two linear movements.

3.1. Experiment 1: Perception of approximated tone contours of isolated syllables

Table 1: List of 18 carrier syllables.

<table>
<thead>
<tr>
<th>Tone</th>
<th>Base syllable</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>wai</td>
</tr>
<tr>
<td>2</td>
<td>ji</td>
</tr>
<tr>
<td>3</td>
<td>jing</td>
</tr>
<tr>
<td>4</td>
<td>面</td>
</tr>
<tr>
<td>5</td>
<td>形</td>
</tr>
<tr>
<td>6</td>
<td>祐</td>
</tr>
</tbody>
</table>

Figure 2: Approximations of isolated tone contours.

Figure 3: Result for perception of approximated tone contours of isolated syllables.
3.2. Experiment 2: Perception of approximated tone contours of disyllabic words

In continuous speech, there exists co-articulation between neighboring tones [1]. That is, the F0 contour of a naturally spoken disyllabic word is not a simple concatenation between the tone contours of individual syllables as they appear in isolation. If the tone contours are at different pitch levels, a transition period is introduced in between. This transition is crucial for perception and identification of tones, as reported in [12]. In [3], it was shown that the perceived naturalness of speech is highly sensitive to abrupt F0 change. There is a necessity to ensure a “continuous” F0 contour at the juncture of syllables. In this experiment, we investigate on the approximation of tone transitions. Disyllabic words are used as the carriers.

3.2.1. Method

The upper row of Figure 4 gives the F0 contours of two disyllabic words. The gradual transition between individual tones is clearly seen. Based on the observations in Experiment 1, we assume that a level tone (Tone 1, 3, 4 and 6) can be approximated by a single linear movement and a rising tone needs two movements. We are interested to know if the transition between the two tones can be represented by a simple linear approximation without loss of perceived naturalness. The lower row of Figure 4 shows how the F0 contours of disyllabic words are approximated by linear segments. In the case that both syllables carry level tones, three linear movements are used. If both tones are rising tones, we need five linear movement, two for each rising tone, and one for transition.

3.2.2. Material and test procedures

Two tests were carried out. Test 1 concerns only level tones. In Test 2, at least one of the syllables in the word carries rising tone. Only those tone combinations with conflict boundary are included because the transition is particularly critical in such cases. That is, the two tones must have significantly different pitch level at the junction, e.g., Tone 4 followed by Tone 1. The number of disyllabic words selected for Test 1 and Test 2 are 16 and 28, respectively. A native female speaker recorded these words. For each of the natural word samples, the F0 contour was approximated as shown in Figure 4 to generate a modified stimulus.

The test procedures were similar to Experiment 1. Subjects were asked to select the preferred stimulus from a pair of natural sample and modified sample. Ten subjects participated in Test 1 and another ten subjects participated in Test 2.

3.2.3. Results

Figure 5 shows the results of Test 1 and Test 2. In Test 1, the approximated contour can reach equivalent perception to natural speech in most cases. There are individual differences among the 16 words. For Word 15, a large perceptual difference is observed between approximated and natural contours. For Word 12, a significant portion of responses favor the approximated contour. Figure 6 shows the contours of these words. In Test 2, the preferences are not as unanimous as Test 1. Nevertheless, almost all of the approximated contours attained equivalent or more preferable perception in the majority (>50%) of the responses.

![Figure 4: Linear approximations of tone contours of disyllabic words](image)

![Figure 5: Results for perception of approximated tone contours of disyllabic words](image)

![Figure 6: Examples of approximations for tone contours of disyllabic words in Test 1](image)
3.3. Experiment 3: Perception of approximated tone contours of sentences

This experiment is designed to assess perception of linearly approximated F0 contours at sentence level, which is the goal of text-to-speech synthesis. The testing materials include five sentences of about 20 syllables in length. The utterances are obtained from the CUProsody speech database [13]. The speaker is a female native Cantonese speaker (different from the one in Experiment 1 and 2). The F0 contour of each natural utterance is linearly approximated to generate a modified stimulus. The approximation patterns used in Experiment 1 and 2 were adopted. Eight subjects participated in the test, with the same testing procedures as the previous experiments.

3.3.1. Results

As shown in Figure 7, all modified utterances attained at least equivalent perception to the natural speech. In the cases of sentence 3 and 5, the natural utterance was not preferred at all. This indicates that perceptually most preferable target may not be found in human speech.

![Figure 7: Result for perception of approximated tone contours of sentences.](image)

4. Discussion & Conclusion

Our experimental results clearly show that Cantonese speech perception is not sensitive to the exact acoustic F0 values at individual time instants. It is the major trend of pitch movement that most affects perception. We also showed that linear approximation is sufficient to describe the pitch movement. The six Cantonese tones can be represented by one or two linear movements, and the transition between tones can be represented by a single linear movement, without creating any noticeable perceptual difference. Such simple approximations are desirable for F0 modeling because it requires only a small number of parameters.

In the current study, the best linear approximations of the F0 contours were determined subjectively based on perceptual comparison. This may lead to a debate on the validity of the experimental results, since the modified stimuli had been pre-judged to be the “closest” to the natural one before they were presented to the subjects. However, it must be noted that the goal of our study is to show that there exists such a linear approximation that can reach equivalent perception to the natural contour. The way how we can obtain such an approximation is not relevant.

Towards the direction of building a perception-driven F0 model, this current work is only a small step. We have demonstrated the feasibility of using linearly approximated tone contours to capture the perceptually most distinctive features in Cantonese. The next step will be a quantitative investigation by which the approximated contours can be parameterized, e.g., the determination of the slope, the height and the duration of a single linear movement for level tone, and the beginning and the end of the transition period.

5. Acknowledgements

This research is partially supported by two Earmarked Research Grants (Ref: CUHK 4227/04E and CUHK 413405) from the Hong Kong Research Grants Council.

6. References

[7] 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.