Adapting SpeakGoodChinese to teach Cantonese

Gijs Kruitbosch, University of Amsterdam, May 2008

1 Introduction

SpeakGoodChinese (SGC) is a computer program written by a team at the University of Amsterdam and experts from other universities, that intends to help the user learn to properly pronounce words in Mandarin Chinese, the official administrative language of the People's Republic of China.[5] It is based on Praat, another piece of software that is used extensively across the world to aid research in linguistics, speech analysis and synthesis.

Cantonese Chinese is a major Chinese dialect, spoken by over 70 million people[4]. Like Mandarin Chinese, it is a tonal language, which means that the (change in) fundamental frequency of each syllable alters its semantics. Mandarin Chinese uses 4 different tones and a fifth neutral tone to accomplish this. Cantonese Chinese uses 6 main different tones, and 3 tones sometimes considered separately. The latter three tones are equal in pitch pattern to three of the six main tones, but are shorter in length. A more extensive introduction of the different tone models is offered in section 2.

In this paper, I will try to outline the approach Speak Good Chinese currently uses to teach users Mandarin (section 3), why it is difficult to adapt this approach to Cantonese Chinese (section 4), how this could be attempted anyway (section 5) and finally some conclusions regarding Speak Good Chinese and the Cantonese dialect in section **??**.

2 Mandarin and Cantonese tonal systems

Both Mandarin and Cantonese are tonal languages. This means that the tone alterations within a word change its meaning, much in the same way that changing the actual phonemes could change it. For instance, in Cantonese, the words /fan^1/, /fan^3/ and /fan^6/, even though they use the same phonemes, mean 'divide', 'sleep' and 'share', respectively. This distinction is only pronounced by the difference in tone.

In Mandarin Chinese, 4 distinct full tones and a fifth 'neutral' tone are used:

1. The level tone in Mandarin starts at a high frequency and stays level. It tends to be a little bit longer than the average tone.
2. The rising tone in Mandarin starts at a lower frequency and rises to the same level as the first tone.
3. The third tone in Mandarin starts at a halflow frequency, drops to a low frequency, and then possibly rises quickly to halfhigh. However, the precise pronunciation of this tone depends on the tone following it.
4. The falling tone in Mandarin falls quickly from a high frequency (roughly that of the first tone) to a low frequency.
5. The neutral tone is pronounced at the same frequency the last tone ended up, and relatively short.

In Cantonese Chinese, 6 main tones are used, and three additional tones, called entering tones, are often distinguished as shorter versions of three of the main tones. They occur only in syllables ending in /t/, /k/ or /p/. All of the tones are visualized in figure 1. Tones 2 and 5 are the only ones which are not level: both rise, one more than the other.

While Cantonese has more tones than Mandarin, it also has several tones which are level in pitch, as opposed to Mandarin which has just one. To further complicate matters, the fluent pronunciation of the level tones in context is often not level in pitch at all [3].

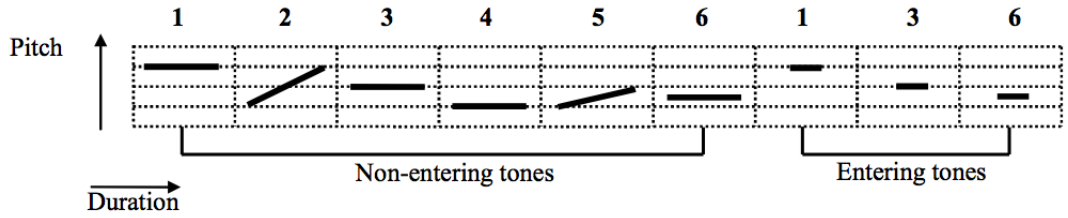

Figure 1: Pitch patterns of Cantonese tones [2]

3 Speak Good Chinese interpretation and synthesis of Mandarin

Currently, Speak Good Chinese is able to compare spoken Mandarin with its own synthesis of the same Mandarin, and deduce whether or not the speech fragment is a correct rendering of the Mandarin word. This section will go into some more detail as to the technique used to make this comparison, and how Speak Good Chinese attempts to aid the student in correctly pronouncing the word.

First, the student records their pronunciation. The pitch range of this pronunciation is used to determine a base frequency, F_h, which is scaled up or down from standard frequencies for two different male, a female or a child's voice (150, 200, 300 and 450Hz, respectively). Scaling up is allowed to a greater extent than scaling down, as students' natural reaction to first studying Chinese seems to be to pronounce tones higher than required, which resulted in the students being given more clearance in that direction.

Using this base pitch (F_h), SGC generates reference pitch contours for all possible combinations of tones for this sample. The length of these samples depends somewhat on what tones are used. The exact multipliers used to determine the length of each syllable can be found in [5]. So, if the word that is pronounced has two syllables, there are a total of 36 ($6 \cdot 6$) combinations. However, the first tone cannot be neutral (which removes 6 possibilities) and the 2-3 and 3-3 combinations of tones are identical, leaving 29 possible pronunciations of the word in question.

The pitch contours of the student's speech recording are then compared to all the different generated bits of speech, using Dynamic Time Warping (DTW). The most likely candidate is chosen, assuming that there was one that was within some range of any of the tried combinations of tones. Some bias towards the correct choice is then applied, picking the correct one even if the student's pronunciation is closest to another pronunciation. If the tone combination that is recognized is still not the correct one, Speak Good Chinese tries to adjust for specific "mistakes" that DTW makes. If even these variations do not result in recognizing the correct tone pattern, these adjustments are reverted.

Finally, some analysis is done on the speech provided by the student, in order to give advice to the student so they can improve their pronunciation. Currently, SGC is able to advise if the speech is either too high or too low, if the changes in pitch for non-level tones are too small or too big, or if the wrong tone was recognized.

4 Problems in extending Speak Good Chinese to support Cantonese

Because of the difference in the tones used in Cantonese, as well as the way Cantonese listeners identify tones, and the way SGC currently recognizes Mandarin tones, extending SGC to support Cantonese will be quite complicated. In this section, I will attempt to outline precisely how and why this is the case.

First of all, the pronunciation of the different tones in continuous speech by a native speaker varies quite significantly from the canonical patterns described in section 2, as shown in [3]. Even the traditional model described earlier has been questioned by some researchers [1]. Furthermore, literature on the interactions between different tones seems to be very scarce, making it particularly difficult to generate a natural correct pronunciation to compare a student's attempts with.

Second, using an adaptive reference pitch and DTW in a tonal language with many level tones is highly error-prone. Because there are fewer distinctive tone movements, the pitch range of a speaker is much harder to determine from just one word. Consequently, if someone pronounces only one level tone at 200Hz, we cannot be entirely sure if this is tone 1, 3, 4 or 6 for this speaker, as there is no other tone present to compare it to, and there is no hard-set pitch level at which to pronounce these tones. So, the reference pitch from just one syllable is going to be inaccurate. For multiple syllables, the situation is a little bit better, but still not ideal due to the large number of level tones.

Then, the use of DTW further complicates this, as the way tones are normally perceived by native speakers depends on context and phoneme categories [6], rather than the distance to a reference pitch, which is what DTW would use to classify them. In other words, whereas DTW might classify a sound as tone 4 when it appears after tone 1, because it is a tone that is significantly lower given the reference pitch deduced from the entire sample, a native speaker might classify it as tone 3 or 6, based on their experience with the speaker which indicates that they pronounce tone 4 even lower. This experience and categorization of phonemes cannot be encoded in plain DTW, and a completely novel approach would have to be followed in order to do so.

Finally, while Mandarin has been officially standardized by the Chinese government, Cantonese has seen no such standardization. There is considerable disagreement, even within the linguistic community, about the definition of the tones used in Cantonese. Notable problems include whether or not the three entering tones should be considered separate tones, whether IPA's tone-letters can be considered accurate representations of the tones used and if tone 3 is really located in the center of the range of tones used in Cantonese [1].

5 Implementing Cantonese support in Speak Good Chinese

If, despite the problems outlined in the previous section, we try to implement Cantonese tone recognition and generation in SGC, I feel the following steps should be taken to do so. Note that these steps assume the addition of a Cantonese "language pack". The specific tones and rules of each new language are separated from the overal structure of the program into separate packs. Indicating this in each of the following steps was felt to be overly repetitive, so these caveats are omitted from the remainder of this section.

5.1 Pitch detection

In order to be able to recognize different speakers' utterances, which may have very different frequencies, SGC uses a reference pitch (partly) deduced from the speakers' own utterances. This pitch is deduced from the recording, by observing the highest pitch (register) used, and the range of tones in the recording. These are adjusted for within limits of 3 semitones down and 6 semitones up [5]. However, for Cantonese, this approach does not work well, as outlined in the previous section: if only level tones occur in a word (which is quite likely given the number of level tones and the fact that we only use words with a maximum of two syllables), it is not possible to properly determine if these tones are generated correctly by the student, as they may use the same frequency for all the level tones.

Several solutions can be considered in order to solve this problem:

- One could use the current approach anyway. This would cause a severe bias towards correct pronunciation, because there are so few tones which move significantly in pitch (i.e. which are not level). This approach will frustrate students the least, but will also teach them the worst, given that it will be so lenient.

- One could reuse the reference pitch from previous words. In order to do this, we should keep the pitch 'top' and range determined in `SGC_ToneProt.praat` for determining consistency with the pitch of the next word, possibly allowing a small adjustment of one semitone or similar in order to compensate for natural changes in repeated fragmented speech (i.e., speech that is not a continuous utterance, such as a complete sentence). However, the possibility for errors is very large in this approach, and students may also find themselves lost as to how to improve: if their previous pronunciation was wrong, the reference pitch would still influence how the next one was judged, so finding the 'correct' intonation considering this reference pitch will be very hard. In order to

mitigate this, one could choose not to save the reference pitch for words which were determined to be wrong, but even this would help comparatively little.

- One could also reuse the current approach, but allow for much less variation (e.g. 2 semitones down and 4 semitones up). This would make the recognition a little bit more strict, but it is not clear if this is enough, considering the fine distinctions between the Cantonese tones 3, 4 and 6.

This problem is probably the worst obstruction to implementing a Cantonese tone recognizer in an useful, reliable and educative manner.

5.2 Individual tone implementation

First, the proper tone duration relations should be implemented in the `ToneRules.praat` script. Processed data on the proper length relations of these tones seems to be scarce to non-existant, and it would therefore be best to attempt to extract such data from a natural speech corpus (such as the one used in [3], though preferably with more than one speaker), or have them set by expert native speakers of the language.

Similarly, the same script should be adapted to contain the main Cantonese tones, in the procedure `toneRules` referenced by 1 through 6, with the Dutch neutral sound being referenced as 7. The levels of the Cantonese tones are quite different from those of Mandarin, however. I will attempt to outline the relation between the tones in terms of pitch height.

Literature seems to agree that tones 1, 3, and 6 are level tones. Not all literature considers tone 4 as being level [6]. However, we shall assume this is the case for now. Traditional models have suggested that tone 1 is twice as far from tone 3 in terms of pitch as tone 4 is from tone 6 [2]. In this model, tone 1 is the highest tone, roughly one scale step up from tone 3, which is half a step from tone 6, which is half a step from tone 4 (see also figure 1). However, recent research has indicated that this model may be wrong, and that in natural speech, tone 1 is in fact also half a scale step up from tone 3 [1]. Again, careful analysis of a speech corpus in Cantonese by multiple speakers would most likely provide the best way of resolving such conflicting ideas.

For all these level tones, the first point (`startPoint`) should be set in accordance with the above, and the last point (`endPoint`) ought to be the same, or only slightly lower. Here, I must note that in continuous speech, these tones may not always be exactly level, which would require the addition of at least one extra point (`midPoint`), which would be located slightly lower than the first and last points [3].

The two remaining tones, 2 and 5, are rising tones, with the former rising twice as much as the latter. Per [2], the start point for tone 2 should be located at roughly the same level as that of tone 4, and the end at roughly that of tone 1. Tone 5 should also start at the same pitch as tone 4, but end at the pitch of tone 3. It is not clear how moving the level of tone 1 down as per [1] would influence the height of the `endPoint` of tone 2. For now, I have assumed this point moves with that of tone 1.

Tone	Traditional		Ho et al., 2007	
	Start	End	Start	End
1	F_h	F_h	F_h	F_h
2	$F_h - 4\delta$	F_h	$F_h - 3\delta$	F_h
3	$F_h - 2\delta$	$F_h - 2\delta$	$F_h - \delta$	$F_h - \delta$
4	$F_h - 4\delta$	$F_h - 4\delta$	$F_h - 3\delta$	$F_h - 3\delta$
5	$F_h - 4\delta$	$F_h - 2\delta$	$F_h - 3\delta$	$F_h - \delta$
6	$F_h - 3\delta$	$F_h - 3\delta$	$F_h - 2\delta$	$F_h - 2\delta$

Table 1: Rules for Cantonese tone simulation. F_h is the reference frequency, the height of tone 1, and δ is the distance in pitch between tone 3 and tone 4 and tone 6.

The frequencies of these tones are summarized in table 1. An important problem with this representation is that it does not take into account the transitions between tones: tone 1 and 2 following each other may sound very different from the individual tones. However, no comprehensive result in literature

as to the influence of this kind of context has been found. In [3], there is some discussion on the matter, but the approach there focuses only on average heights of the tones, not on the complete pitch contours, which is the type of data necessary to properly model tones in context. In summary, there is currently insufficient information available to do a full model of tones and how they interact with each other in this report. However, as with tone durations, this information is vital to be able to teach students to speak Cantonese in a natural manner, and it would be highly preferable to research and implement this information in SGC in order to teach proper Cantonese.

5.3 Syllable parsing

In order to properly pronounce words, the length of syllables and their voiced and unvoiced parts must be taken into account. Extensive data as to what phonemes are voiced and unvoiced is available in [3], and for the sake of brevity has not been reproduced here. The regular expressions in the `convertVoicing` procedure in `ToneScript.praat` should be updated using this data.

5.4 Constructing tone movements

For the proper generation of tone movements, some of the last-time adjustments for Mandarin should first be removed. An example is line 116 to 118 of `ToneScript.praat`:

```
if toneSyllable = 3 and nextTone = 3
    toneSyllable = 2
endif
```

It is probable that it will be necessary to implement similar adjustments for Cantonese, but again literature insight in this problem is very limited, and a detailed consult with a native speaker, or extensive field testing, would be required to make such adjustments.

Then the table of different tone combinations needs to be made larger. The `ToneList` table needs to consist of 49 entries (7 times 7), and both the inner and the outer tone generation loop should always loop from 1 to 7 (there is no more neutral tone which cannot be used in the first syllable of a word, so this logic should disappear).

5.5 Recognition and feedback

Using the reference pitch and the generated tones, SGC attempts to use DTW to recognize whether the student's utterance was correct. In earlier subsections, I have explained how the code generating tones and determining the reference pitch should be adapted, so now we can focus on the actual recognition.

DTW could still be used but it is likely that, combined with the bias for the 'correct' recognition (lines 89 through 104 in `ToneRecognition.praat`), and the adaptive reference pitch detection (refer to subsection 5.1) it will generate more false positives than the current Mandarin implementation does.

On the other hand, the current implementation features several rather ad-hoc rules to compensate for common recognition mistakes. For instance, a recognized tone 6 (Dutch intonation) at the end of the word may in fact be a neutral tone (lines 333 through 338 in `SGC_ToneProt.praat`). SGC compensates for these mistakes, and similar adjustments must likely be made for Cantonese. However, it is once more not entirely clear what such problematic cases would be. Several purely hypothetical suggestions are made below, based on the theoretical tone contours outlined in section 2:

- 5-3 recognized as 2-1 (or vice versa): unlike Mandarin, Cantonese has not one but two rising tones, and they may be easily confused.
- 1-4 recognized as 1-7 (with the final tone being a Dutch intonation): just like Mandarin, the big drop in tone may be misrecognized – if the drop in tone was really too far, the pitch range detection would already have noticed this.
- the tones 3, 4 and 6 may be easily confused, especially after a drop in pitch from tone 1, as the differences between these tones are comparatively subtle.

Despite these educated guesses, only extensive testing and interpretations by native speakers would be able to give an accurate representation of the tones that are most likely to be confused.

5.6 The UI and word list

Of course, the word lists also need to be adapted: they should contain only Cantonese words, with tones 1 through 6, rather than the current list of Mandarin words with tones 0 through 4. In a perfect world, SGC would be able to distinguish between the two languages automatically, and switch language based on the word list, using the tones that are used in the different words.

Furthermore, the UI and the pronunciation hints it gives must be updated to consider Cantonese rather than, or in addition to, Mandarin. This is true in particular for the descriptions of how to pronounce a specific mispronounced tone, which are given from `SGC_ToneProt.praat` using a localization file.

6 Conclusion

Considering the problems faced when implementing Cantonese support in SGC, it is clear that this would be no easy task. Though a naive implementation may be relatively simple, the nature of Cantonese and its differences with Mandarin are almost certain to severely limit the effectiveness of such an implementation. I have tried to outline the problems that one faces when implementing Cantonese support, how one can attempt to work around them, and where and how specifically Cantonese support could be implemented in SGC.

The main problems when attempting to implement Cantonese support are as follows:

- Problematic pitch range detection due to the comparative rarity of non-level tones;
- Lack of data concerning the interaction of different tones, their pronunciation by various native speakers, the length of these different tones, and the confusion ratio between such tones;
- Considerable disagreement as to the idealized renditions of different tones and their relations;
- Problematic detection of tone intervals using DTW because of the number of level tones, and the fact that the reference pitch may be off (per the first item).

Hence, considerable future work remains to be done. One part of this work is empirical, and would consist of analysing the Cantonese language using a speech corpus, and adjusting the pronunciation models of this language as they are currently featured in literature based on the findings of such an analysis. Necessary parts of such an analysis would consider the interaction between tones, their pitch frequencies and relation to each other when spoken by native speakers, and tones which are likely to be confused, amalgamated or interchangeably used in common usage of the language. Ideally, such a corpus would not consist of speakers from just one region (such as Hong Kong), but from several regions where Cantonese is the de facto standard language of conversation.

Another part of this work is applied, and is the actual implementation of Cantonese support in SGC. This will hopefully be simplified by the extensive description given in this paper. However, the effectiveness of such an implementation is not clear, and it may be necessary to consider more elaborate measures to effectively teach Cantonese pronunciation through a computer program, such as an increased use of context.

Finally, I think that an implementation of Cantonese language support that is as effective in teaching contexts as SGC currently is for Mandarin will be tremendously hard to accomplish. The problems outlined in this report, though perhaps not insurmountable, will significantly impact the usefulness of such an eventual implementation, unless effective solutions for the problems described are found. I do hope that a version of SGC that supports Cantonese will be possible in the future, so as to enable more people to discover and speak the language. Hopefully, developing such an implementation will also, directly or indirectly, benefit research into the language itself.

7 Acknowledgements

I would like to thank dr. Rob van Son and Stefan de Konink for their help in understanding Praat, SGC and Cantonese, and making this project possible. I would also like to thank Louise Ng and Carlotta Chan for their help in understanding the Cantonese and Mandarin languages, and their romanization forms.

References

[1] Rerrario Shui-Ching Ho and Yoshinori Sagisaka. F0 analysis of perceptual distance among cantonese level tones. In *Interspeech 2007*, pages 1038–1041, Antwerp, Belgium, August 2007. ISCA.

[2] Yujia Li and Tan Lee. Perceptual equivalence of approximated cantonese tone contours. In *Interspeech 2007*, pages 2677–2680, Antwerp, Belgium, August 2007. ISCA.

[3] Yujia Li, Tan Lee, and Yao Qian. Analysis and modeling of f0 contours for cantonese text-to-speech. *ACM Transactions on Asian Language Information Processing (TALIP)*, 3:169–180, 2004.

[4] Jr. (ed.) Raymond G. Gordon. *Ethnologue: Languages of the World*. SIL International, Dallas, Texas, 15 edition, 2005.

[5] D.J.M. Weenink, G. Chen, Z. Chen, S. de Konink, D. Vierkant, E. Hagen, and R.J.J.H. van Son. Learning tone distinctions for mandarin chinese, August 2007.

[6] Hongying Zheng, Peter WM. Tsang, and William S-Y. Wang. Categorical perception of cantonese tones in context: a cross-linguistic study. In *Interspeech 2007*, pages 2745–2748, Antwerp, Belgium, August 2007. ISCA.