Using Cross-syllable Units for Cantonese Speech Synthesis

K.M. Law and Tan Lee

Department of Electronic Engineering
The Chinese University of Hong Kong
{kmlaw, tanlee}@ee.cuhk.edu.hk

ABSTRACT

Monosyllables have been widely accepted as the basic units for concatenative speech synthesis of Chinese dialects. However, concatenating individual syllables is not adequate to produce highly natural synthetic speech because of the improper coupling at syllable boundaries. This paper describes a preliminary research of using cross-syllable units for Cantonese speech synthesis. The acoustic inventory contains 1,725 cross-syllable units, which are excised from properly selected and recorded carrier words. TD-PSOLA is employed for prosodic modification of synthetic speech. The results of subjective listening tests reveal that the proposed use of cross-syllable units has potential in producing highly natural synthetic speech, although the currently achieved performance is only fair. Substantial improvement is anticipated with better smoothing technique for waveform concatenation and greater coverage of context-dependent variation of the acoustic units.

1. INTRODUCTION

The ultimate goal of text-to-speech synthesis is high intelligibility and naturalness. In this regard, the concatenative approach, i.e. synthesizing speech by connecting pre-stored acoustic units, has been proven very effective for various languages. The types of synthesis units being used are language specific — di-phones for many Western languages, and monosyllables for Chinese dialects like Putonghua and Cantonese [1]-[4]. Syllables are considered as independent pronunciation units for Chinese. Each syllable corresponds to a written Chinese character, which is the smallest non-separable lexical term that can be pronounced. It is quite straightforward to synthesize Chinese speech via concatenation of syllables.

Our previous work reported a Cantonese text-to-speech system with an acoustic inventory of about 1,800 isolated tonal syllables [4]. With appropriate control of segmental duration and syllable-wide F0 profile, this system can generate highly intelligible speech but the naturalness is not quite satisfactory [5]. The synthetic speech is often commented as “one syllable jumps out in a sudden after another”. The perceived transition between adjacent syllables is rough since cross-syllable transition was not considered in the concatenation process. In many circumstances, consecutive syllables are closely co-articulated in naturally spoken Cantonese. To produce highly “smooth” speech, such cross-syllable effects need to be considered. This can be done by modifying the speech signal in spectral domain, so as to produce the desirable transitory effect. However, owing to the difficulty in establishing precise relationship between acoustic features and phonetic transition, this approach has not been very successful in synthesizing highly natural speech.

Another way of capturing cross-syllable effect is through the use of speech units other than isolated syllables, e.g. word, in which the contextual variation of interest are included implicitly. In this paper, we propose to use cross-syllable phonemic units for concatenation in Cantonese speech synthesis.

In the next section, a brief introduction to the Cantonese dialect is given. The selection of cross-syllable units and the preparation of acoustic data are discussed in Section 3 and 4 respectively. The overall processes of Cantonese text-to-speech synthesis with cross-syllable units and monosyllable units are presented and compared in Section 5 and the results of perceptual test are provided in Section 6. In Section 7, the limitations of the existing approach are discussed, with a number of suggestions for improvement.

2. PROPERTIES OF CANTONESE

Phonologically, a Cantonese syllable is divided into an INITIAL and a FINAL. The INITIAL is an optional consonant while the FINAL consists of a vowel nucleus (or diphthong) followed by an optional coda. There are 19 INITIALs and 53 FINALs in Cantonese. Given the phonological constraints, the total number of legitimate INITIAL-FINAL combinations is around 620, which are commonly referred as base syllables [6].

Cantonese is well known of being rich in tones. There are totally nine citation tones in Cantonese. As a result, the number of so-called tonal syllables in Cantonese is about 1800. Basically, tone is a feature that describes the trajectory of the fundamental frequency (F0) across a syllable. F0 is defined only for voiced speech. In a Cantonese syllable, the FINAL part is always voiced while the INITIAL can be either voiced or unvoiced.

![Figure 1. The nine tones of Cantonese](image)

The nine tones in Cantonese can be characterized by the F0 patterns as shown in Figure 1. They are generally categorized as entering tones and non-entering tones. The entering tones are shorter in duration but have similar F0 level with some non-entering tone counterparts. Entering
tones are always associate with stop codas (/p/, /t/ or /k/). If tones are being classified only in terms of F0 level, a 6-tone system is resulted. It consists of the six non-entering tones in which two are rising (tone 2 and 5) and the rest are level (tone 1, 3, 4 and 6). In other words, tone 7, 8 and 9 are merged with tone 1, 3 and 6 respectively.

3. THE ACOUSTIC INVENTORY

Our goal is to select a set of speech units that cover all possible cross-syllable contexts in naturally spoken Cantonese. Obviously, di-syllable units would not be a good choice simply because there are too many of them, about 3 millions in theory. In this work, we propose to use cross-syllable FINAL-INITIAL combinations as the synthesis units. Suppose the two consecutive syllables, S1 and S2 have sub-syllable compositions of I1-F1 and I2-F2 respectively. The juncture unit F1-F2 is expected to cover most of the intersyllable co-articulation between S1 and S2. If the phonetic identities of I1 and F1 are not considered, the unit F1-F2 is said to be context-independent. On the other hand, the FINAL part in a syllable is regarded as the primary carrier of tone while the INITIAL is generally less relevant to the tone identity. In view of this, the tone identity of S1 (or F1) is considered to be an important contextual factor for the unit F1-F2. Consequently, we aim at a set of context-independent and tone-dependent cross-syllable F-I units. Theoretically, the total number of such units is about 4940.

A simple example is given in Table 1, where the syllabic units are labeled using the “Jyut Ping” transcription scheme [6]. In this example, F1 is /oeng/ while I2 is /g/. The targeted F-I unit is /oeng1-g/.

<table>
<thead>
<tr>
<th>Targeted F-I unit</th>
<th>Carrier words</th>
</tr>
</thead>
<tbody>
<tr>
<td>ai3-g</td>
<td>sai3-gaii3 (世界)</td>
</tr>
<tr>
<td>aau3-p</td>
<td>gaau3-pai3 (教派)</td>
</tr>
<tr>
<td>eoi5-s</td>
<td>neoii5-sing3 (女性)</td>
</tr>
</tbody>
</table>

Table 2. A partial list of carrier words

The selected carrier words were read out by a female speaker of 31 years old in normal speaking rate. All recordings were completed within 2 consecutive days so that consistency of voice quality is guaranteed.

4. DATA PREPARATION

Each of the required F-I units was recorded as an embedded body in a carrier word. The carrier words were selected carefully based on the following criteria:

- high readability – words that are commonly used in daily speech are preferred;
- match of tone identity – words made up of syllables with tone equals to 3 and 5 are preferred;
- ease of segmentation – F-I units that are easily excisable from the carrier word are preferred. It is desirable that the preceding INITIAL is a stationary and unvoiced consonant like plosive, fricative or affricate. Examples such as /b/, /p/, /d/, /s/ and /f/ etc are favorable;
- word length – words with shorter length would be preferred in order to minimize the sentential effect.

In general, the F-I units are subject to prosodic modification when being used for speech synthesis. That is, its F0 profile needs to be changed according to some prescribed targets. Time-domain Pitch-Synchronous Overlap-Add (TD-PSOLA) is one of the major signal processing techniques for pitch modification [7]. TD-PSOLA would not introduce significant signal distortion if the degree of modification is moderate. Taking advantage of this flexibility and considering the similarities among the Cantonese tones, we can further reduce the number of required F-I units. As shown in Figure 1, the six non-entering tones can be classified as rising tones (tone 2 and 5) and level tones (tone 1, 3, 4 and 6). Within each category, only one representative tone is covered in the F-I inventory. For the level tone category, tone 3 is chosen as the representative. Tone 3 is located in the middle of the dynamic range of F0. It can be modified to both higher tone (tone 1) and lower tone (tone 4 and 6) without much signal distortion. And for the rising tone category, tone 5 is chosen as the representative since its change of F0 level is more mild than tone 2. With these 2 representative tones being used, the number of required F-I units drops to 1,653, which is about one-third of the original number. In addition, the 19 INITIALs and 53 FINALS should also be included as themselves, for the use in sentence-beginning and sentence-ending positions. As a result, the acoustic inventory consists of 1,725 acoustic units.
The excised F-I units are processed in the way that the amplitude peaks of pitch cycles in all voiced speech segments are marked. The pitch-marking process is done by the "get_f0" program in the ESPS package, followed by the manual inspection [8]. The pitch marks are used by TD-PSOLA for the modification of F0 and segmental duration.

5. THE TTS SYSTEM

Figure 3 depicts the typical architecture of a Chinese text-to-speech system. As this study is focused on the aspect of the acoustic inventory, other modules in the system are kept simple and fundamental. The text-processing module serves to convert the input Chinese text into a pronunciation representation. This representation is essentially a string of acoustic units that are available in the acoustic inventory. The prosody module specifies the segmental duration and F0 movement for each of the acoustic units being involved [5]. The PSOLA module completes the synthesis process by modifying the retrieved waveforms and concatenating them.

Figure 3. The architecture of a Chinese TTS system

5.1 The Baseline System: I-F TTS

As reported in [5], our baseline system is based on 1,800 I-F units (tone-dependent). In this system, the prosody module generates the duration of voiced part and unvoiced part in each I-F unit (or syllable), as well as the F0 profile for it. The application of PSOLA algorithm is straightforward in this case.

5.2 The F-I TTS System

To make use of the same prosody module as described above, the synthesis process for cross-syllable F-I units needs some modification. PSOLA cannot be applied directly to each individual F-I unit. Instead, its left and right neighbors should be considered simultaneously. This is illustrated by an example as shown in Figure 4. First of all, the speech waveforms are simply concatenated together without any prosodic modifications. The monosyllable units are then identified from this speech segment with reference to the time alignment of INITIALs and FINALS. Finally, F0 and duration of these syllable units are modified by PSOLA in the same way as the I-F TTS.

Figure 4. An example illustrating how PSOLA is applied in the F-I TTS system.

6. PERCEPTUAL TEST

The aim of perceptual test is to compare the naturalness and intelligibility of synthetic speech produced by the TTS system using cross-syllable F-I units and the one using monosyllable I-F units. Fifteen test sentences are selected randomly from a newspaper. They cover topics of news, finance, sports and entertainment. Their average length is about 8 to 9 syllables. Each test sentence was processed in parallel by the two systems and two synthetic sentences were obtained. The participating subjects were required to listen to both versions in succession (with randomly shuffled order) and then score them in a scale of 5 grades - 5 being the excellent while 1 being the poor. Table 3 shows the Mean Opinion Scores (MOS) of the tests [9]:

<table>
<thead>
<tr>
<th></th>
<th>I-F TTS</th>
<th>F-I TTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naturalness</td>
<td>3.47</td>
<td>3.12</td>
</tr>
<tr>
<td>Intelligibility</td>
<td>3.62</td>
<td>3.33</td>
</tr>
<tr>
<td>No. of Wins</td>
<td>9</td>
<td>6</td>
</tr>
</tbody>
</table>

Table 3. Results of the Perceptual Test

7. DISCUSSIONS

The MOS results are saying that both the newly developed F-I TTS and the baseline I-F TTS reach a performance level between “acceptable” and “good”, in terms of both naturalness and intelligibility. Examining the scores of individual sentences, we find that the F-I TTS system performs better for only 6 out of the 15 sentences (see Table 3).

Although the results seem to be unfavorable, they are not out of our expectation. Reasons are as follows:

1. The 1,725 selected units provide a fairly complete coverage of cross-syllable variation. However, as the units were extracted from some specific word context, they are in fact context-dependent units in terms of intra-syllable context, i.e. I_t-F_i and F_i-F_j. When these
units are used in other context, spectral mismatch becomes inevitable.

2. Some of the F-I units were not well separated from their carrier words simply because it is difficult to separate them. As said in Section 4, for a syllable with voiced INITIAL, the determination of INITIAL-FINAL boundary would be extremely hard. This may also result in strong mismatch of spectral properties at the juncture.

Currently, waveforms are connected by hard concatenation without smoothing in any aspect. This certainly introduces unpleasant audio effect.

<table>
<thead>
<tr>
<th>I-F TTS</th>
<th>F-I TTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naturalness</td>
<td>2.89</td>
</tr>
<tr>
<td>Intelligibility</td>
<td>3.25</td>
</tr>
</tbody>
</table>

Table 4. Scores for those sentences that F-I TTS wins

By analyzing those sentences that F-I TTS system wins, some points are worth notice:

1. Naturalness is affected by intelligibility. If the synthetic speech is clear, its naturalness is relatively high.

2. Context-dependency is critical. These winning sentences are mostly composed of context-dependent F-I units for voiced INITIAL syllables. That means the contents of testing sentences are the same as the carrier words being recorded. As a result, the additional FINAL speech segments carried in the previous F-I unit are matched with the following one in terms of phonetic identity.

In particular, for syllables with voiced INITIALs, it seems more appropriate to incorporate context-dependent F-I unit rather than context-independent one.

This work was originally motivated by the importance of cross-syllable acoustic variation. To realize such importance, the use of F-I units has been proposed and evaluated in this paper. The results of listening tests reveal that this is a feasible approach with not too bad performance. But again, this approach, as its current version, is inadequate because it introduces some other problems. It seems that both the monosyllable TTS and the F-I TTS don't work very well themselves. In this regard, we consider the following directions for possible improvement:

1. combine the use of I-F and F-I units to take care of both intra-syllable and cross-syllable transition;
2. develop appropriate soft concatenation technique for smooth spectral transition from one unit to another;
3. develop a better prosody module.

8. ACKNOWLEDGEMENT

The presented work is partially supported by a research grant from the Hong Kong Research Grants Council.

9. REFERENCES


[6] Linguistic Society of Hong Kong (LSHK), Hong Kong Jyut Ping Characters Table, Linguistic Society of Hong Kong Press, 1997.


[8] ESPS Programs Version 5.0, Entropic Research Laboratory, Inc.