CANTONESE TEXT-TO-SPEECH SYNTHESIS
USING SUB-SYLLABLE UNITS

K. M. Law, Tan Lee and Wai Lau
Department of Electronic Engineering
The Chinese University of Hong Kong
{kmlaw, tanlee, wlau}@ee.cuhk.edu.hk

ABSTRACT
This paper describes our recent investigation on the use of both intra-syllable and cross-syllable acoustic units for Cantonese text-to-speech synthesis. In our previous work, isolated monosyllable units were used for concatenative speech synthesis of Cantonese. The synthetic speech was considered to be unnatural in such a way that there was an obvious lack of perceptual continuity. The proposed system adopts an acoustic inventory that covers all legitimate intra-syllable and cross-syllable acoustic units. Synthetic speech produced via concatenation of such sub-syllable units better captures the pertinent transitory effects that are crucial to perceived naturalness. Different strategies are used to concatenate speech segments with different acoustic-phonetic properties. Subjective listening test shows a noticeable performance improvement that is accounted for mainly by smoother transition between sonorant segments.

1. INTRODUCTION
Concatenation of pre-recorded speech units has become a widely accepted approach in text-to-speech (TTS) synthesis for various languages [1]-[5]. It strives for the automatic generation of synthetic speech with a higher level of perceived intelligibility and naturalness. One of the key research issues concerns the selection of appropriate speech units to be concatenated. On one hand, small-size units like context-independent monophones are preferable because there are only a limited number of them, making the acoustic inventory compact and generic. On the other hand, to better capture the co-articulation effect in fluent speech, large-size units that cover cross-phone transitions are needed.

The selection of concatenation units is language dependent. While many TTS systems for Western languages rely on diphone units or equivalent, monosyllable units have been commonly used for Chinese dialects like Putonghua and Cantonese [4]-[5]. Syllables are considered as independent pronunciation units for Chinese. Each syllable corresponds to a written Chinese character, which is the smallest non-separable lexical item that can be pronounced. It is a fairly straightforward approach to synthesize Chinese speech via the concatenation of isolated syllables.

In our previous work, a Cantonese TTS system was developed based on an acoustic inventory of 1,800 isolated syllables [6]. Time-domain PSOLA was employed for the modification of F0 trajectories and segmental duration. This system was capable of producing highly intelligible Cantonese speech. However, the synthetic speech was considered unnatural in such a way that there was an obvious lack of perceptual continuity. This is due to the fact that the cross-syllable transitions were missing in the synthetic speech. In natural human speech, adjacent syllables are closely coupled with each other. They exhibit different spectral behaviors from the case that they are uttered in isolation. The co-articulation effect is especially severe when the syllable juncture involves voiced speech. Figure 1 shows such a continuant transition in a pair of co-articulated syllables.

![Figure 1](image-url)

(a) syllables uttered in isolation
(b) syllables uttered continuously

This paper presents our recent investigation on Cantonese speech synthesis using sub-syllable acoustic units. In particular, it aims at improving the cross-syllable transitions, which are crucial to perceived naturalness of the synthetic speech. In the next section, the Cantonese dialect will be briefly introduced. An overview of the proposed TTS system will be given in Section 3. In Section 4, the design of an acoustic inventory of 5,800 Initial-Final and Final-Initial units will be described. The recording and data post-processing procedures will also be explained. In Section 5, specific concatenation techniques for different phonetic classes will be discussed. Subjective listening test will be presented in Section 6.

2. THE CANTONESE DIALECT
Cantonese is a major Chinese dialect spoken by over 60 million people in Southern China and Hong Kong. Like other Chinese dialects, spoken Cantonese is seen as a string of monosyllabic sounds. Each Chinese character is pronounced
as a single syllable that carries a specific tone. However, a
character may have multiple syllable pronunciations, and on
the contrary, a syllable typically corresponds to a number of
different characters.

Phonologically, a Cantonese syllable is composed of an Initial
and a Final, as shown in Figure 2. There are 19 Initials and
53 Finals in Cantonese. Some of the Initials are semi-vowels
or nasals. Other categories of Initials include liquids, glides,
fricatives, affricates and plosives. The 53 Cantonese Finals
are divided into five categories: vowel (long), diphthong,
vowel with nasal coda, vowel with stop coda and syllabic
nasal. Except for the syllabic nasal, each Final contains at
least one vowel element.

**Table 1. Examples of sub-syllable units**

<table>
<thead>
<tr>
<th>Word</th>
<th>I-F units</th>
<th>F-I units</th>
</tr>
</thead>
<tbody>
<tr>
<td>hoeng1 gong2 (香港)</td>
<td>h-oeng1, g-ong2</td>
<td>oeng1-g</td>
</tr>
<tr>
<td>zung1 gwok3 (中囯)</td>
<td>z-ung1, gw-ok3</td>
<td>ung1-gw</td>
</tr>
<tr>
<td>jat6 bun2 (日本)</td>
<td>j-at6, b-un2</td>
<td>at6-b</td>
</tr>
</tbody>
</table>

3. OVERVIEW OF THE TTS SYSTEM

Figure 4 illustrates the general architecture of a Cantonese
text-to-speech system. It consists of three key modules,
namely text analysis, acoustic synthesis and prosodic
modification. The text analysis module converts Chinese
character sequence into a string of sub-syllable units that
include both intra-syllable and cross-syllable junctures.
Subsequently, the acoustic synthesis module identifies the
required speech segments and concatenates them properly.
These concatenated segments are subject to prosodic
modification by TD-PSOLA, based on a set of pre-determined
prosodic rules.

This study is focused primarily on the design of acoustic unit
inventory and the concatenation strategies for different speech
units. The text analysis and the prosodic modification
modules are kept as simple and fundamental as possible.
4.2. Design of the Acoustic Inventory

We attempted to collect a speech database that covers all legitimate sub-syllable units defined as above. In a Cantonese syllable, the Final is regarded as the primary carrier of tone. The number of tonal Finals is about 300 and the number of tone dependent I-F units is around 1,800. As stated earlier in Section 2, the Initial segment is optional in a Cantonese syllable. If the so-called null Initial is not considered, the number of tonal F-I units is approximately $300 \times 19 = 5,700$. In the case of null Initial, an F-I unit essentially becomes an F-F unit, where the second "F" refers mainly to the vowel nucleus segment. As a result, the total number of sub-syllable units is about 16,000.

Given that the pre-stored acoustic units are subject to F0 modification at a later stage of synthesis, it becomes less critical for us to include all tonal variants in the acoustic inventory. Taking advantages of this flexibility and the acoustic similarities among the Cantonese tones, the number of required sub-syllable units could be much reduced. This is done by categorizing the six tones into two groups, namely rising-tone group (tone 2 and 5) and level-tone group (tone 1, 3, 4 and 6). For the level tone group, tone 3 is chosen as the representative because it locates at the middle of a speaker’s pitch range. When an acoustic unit carrying tone 3 is prosodically modified to produce either a higher or lower tone variant, the required degree of F0 modification can be kept minimal. For the same reason, tone 5 is selected to represent the rising tone group. Indeed, the F0 level of tone 2 sometimes goes extremely high.

In summary, the acoustic inventory was designed to include sub-syllable units carrying the two representative tones only. The resultant number of sub-syllable units is about 5,800.

4.3. Recording

Each target sub-syllable unit was recorded as an embedded body in a carrier word. The carrier words were created based on the following criteria:

- **Popularity** – words that are frequently used by general speakers are preferred so that a relatively high level of naturalness can be maintained.

- **Word length** – short words (i.e. 2 to 5 syllables) are preferred in order to minimize the sentential effect.

To cover the 5,800 intended units, about 5,700 carrier words were designed. There were a few carrier words each containing more than one target unit.

About 40% of the created carrier words satisfy the "popularity" criterion while the others are more or less "non-sense". Most of these words were actually for F-F units. Though the coverage of these "non-sense" words is high, their occurrence in a sentence is comparatively small. Thus, the overall performance would not be degraded too much.

A female speaker was asked to read the 5,700 carrier words in a natural and fluent manner. In addition to the audio signal, a laryngeal signal was captured simultaneously.

4.4. Data Post-processing

The recorded utterances words were automatically segmented using the technique of HMM forced alignment with the HTK toolkit [8]. The Hidden Markov models (HMM) were context-independent phone models trained with the same set of utterances.

For each utterance, pitch cycles were located based on the laryngeal waveforms. More precisely, the amplitude peak of each pitch cycle in all voiced speech segments was marked. This was automatically done by the "pitchmark" function in the Edinburgh Speech Tools Library [9]. Subsequently, the pitch marks were manually inspected and corrections were made when necessary.

5. SYNTHESIS TECHNIQUES

5.1. The Basic Procedures

In our implementation, the original utterances that contain the entire carrier words were retained in the acoustic inventory, though only the embedded sub-syllable segments would be used. Figure 6 illustrates the basic operation of concatenating sub-syllable units. The intended output is a word of two syllables $S_1S_2$ and the corresponding Initial-Final sequence is denoted by $I_1F_1I_2F_2$. The concatenation involves three sub-syllable units, namely $I_1F_1$, $F_1I_2$ and $I_2F_2$. Suppose the units are found in carrier word A, B and C respectively. There are two concatenation points to be determined, one in the region of $F_1$ and the other in $I_2$. In general, the decisions are made based on spectral distance measure. For example, the Itakura distance [10] between each pair of frames in the $F_1$ regions of word B and word A is computed. The pair with the minimum distance is identified as the concatenation point. The same process is applied to the concatenation in $I_2$.

5.2. Concatenation Strategies

Different concatenation strategies are needed for different speech units. Indeed, the spectral distance measure is obviously not applicable to transitory phonemes like fricatives. On the other hand, the Itakura distance doesn’t really reflect the perceptual difference between fricative segments.
In this study, three basic concatenation strategies are used:

**Plosives/Affricates** (*Initial*): They are featured by highly dynamic spectral properties. In this case, we simply retain the entire *initial* segment in the I₁-F₂ unit and abandon that in the preceding F₁-I₂ unit;

**Fricatives** (*Initial*): They are noise-like and long in duration. In this case, each of the I₂ segments in both F₁-I₂ and I₂-F₂ is divided into two parts of equal length. The second half in the F₁-I₂ is then directly concatenated to the first half in the I₂-F₂;

**Vowels/Semi-vowels/Nasals** (*Initials* or *Finals*): These phonemes are concatenated based on Itakura distance described as above.

Prior to the concatenation process, TD-PSOLA is applied on a syllable basis.

### 6. PERFORMANCE EVALUATION

Subjective listening test has been carried out to evaluate the performance of the proposed techniques. The test materials consist of ten Chinese sentences selected from local newspapers. The sentence length ranges from 6 to 12 syllables and the average is 8.2 syllables. For each test sentence, four different versions are made available for the test:

- **HUMAN** — natural human speech (recorded from the same speaker as the synthesis units);
- **MONOSYL** — TTS based on concatenation of isolated syllables with segmental duration control [6];
- **SUBSYL1** — TTS based on concatenation of sub-syllable units as described in this paper. Segmental duration control is applied as in **MONOSYL**;
- **SUBSYL2** — TTS based on concatenation of sub-syllable units. The output speech is modified to carry "perfect" prosody derived from the utterances in **HUMAN**.

Twenty native Cantonese speakers participated into the listening test. For each test sentence, the **HUMAN** version was played-back first and the other three followed in random order. The participant was asked to assign a score from 1 (poor) to 5 (excellent) for each version.

Table 2 shows the results of listening test. The proposed system, i.e. **SUBSYL1**, performs noticeably better than the monosyllable based system, i.e. **MONOSYL**. The **SUBSYL1** system is preferable especially when the sentence contains relatively more concatenations of vowel or vowel-like segments. The slightly higher score attained by **SUBSYL2** reveals that the proposed approach has good potential for improvement with better prosodic control.

Nonetheless, the current performance level is far from being perfect. The MOS of 3.14 is considered to be slightly above a "acceptable" level. The test participants often commented that there was a certain degree of reverberation making the synthetic sentences sound unpleasant. This may be due to possible spectral mismatch at the concatenation point, as shown by the example in Figure 7. As a matter of fact, the sub-syllable units being concatenated may carry significantly different spectral properties since they are acquired from carrier-word contexts.

![Figure 7. An example of spectral mismatch in the synthetic speech produced by the proposed TTS system](image)

### 7. ACKNOWLEDGEMENT

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

### 8. REFERENCES


