MICRO-PROSODIC CONTROL IN CANTONESE TEXT-TO-SPEECH SYNTHESIS

Tan Lee[1], Helen M. Meng[2], W. Lau[1], W.K. Lo[1] and P.C. Ching[1]
[1]Department of Electronic Engineering
[2]Department of Systems Engineering & Engineering Management
The Chinese University of Hong Kong, Shatin, Hong Kong
tanlee@ee.cuhk.edu.hk http://dsp.ee.cuhk.edu.hk/speech

ABSTRACT

This paper describes a pioneer study on prosodic control for Cantonese text-to-speech synthesis. We attempt to establish a set of segment-level duration rules and context-dependent F0 profiles and apply them to a syllable-based concatenative speech synthesizer which uses TD-PSOLA as prosodic modification technique. The prosodic features are extracted by statistical characterization of a large amount of speech data. Subjective listening test shows that the micro-prosodic control results in a marginal but consistent improvement in perceptual naturalness.

Keywords: TTS, Cantonese, micro-prosody

1 INTRODUCTION

Cantonese is a major Chinese dialect spoken by over 60 million people in Southern China and Hong Kong. As the demand for human-computer speech interfaces rises within Chinese-speaking communities, Cantonese spoken language technologies have attracted increasing attention in recent years.

We have developed one of the few existing Cantonese text-to-speech systems, as previously reported in [1]. This system adopted the syllabe-based concatenative synthesis approach using TD-PSOLA technique. As having been shown in many other studies, the TD-PSOLA method can produce acoustic signal with fairly high voice quality [2],[3]. It is extremely suitable for monosyllabic and tonal language like Mandarin and Cantonese because of its great flexibility in F0 and time-scale modification [4].

Prosodic control is of critical importance for attaining high naturalness of synthetic speech. In this paper, the problem of controlling micro-prosodic parameters for Cantonese TTS is being addressed. By micro-prosody, we refer mainly to the segment-level temporal structure and F0 variation. The temporal structure includes the duration of sub-syllable segments as well as pause length between adjacent syllables. The syllable-wide F0 profile is seen as the primary control of lexical tone. Based on statistical derivation from a large speech database, a set of prosodic rules is established to improve the perceived naturalness of synthetic speech.

2 PROSODIC STRUCTURES OF CANTONESE

A spoken Cantonese sentence is a sequence of syllables. Each syllable essentially corresponds to a Chinese character which may have lexical or grammatical function. Syllable is also considered the fundamental pronunciation unit of Cantonese. Traditionally, a Cantonese syllable can be divided into an INITIAL (I) and a FINAL (F). The INITIAL is basically a consonant onset and the FINAL is typically a vowel nucleus followed by an optional consonant coda. Table 1 gives the list of INITIALs and FINALs, while Table 2 lists all phonologically valid syllable structures in Cantonese.

22 INITIALs	
Unaspirated plosives (UP)	p, t, k, k^w
Aspirated plosives (AP)	p, t, k, k^w
Approximants (G)	, ,
Nasals (N)	, ,
Fricatives (F)	, , ,
Affricates (AF)	t, t, t, t
53 FINALs	
Nasal (N)	,
long vowel (LV)	, , , , , ,
Diphthong (D)	, , , , , , , , ,
long vowel + stop (LV-S)	p, t, k, k, k, p, t, k,
Short vowel + stop (SV-S)	p, t, k, t, k, k
Long vowel + nasal (LV-N)	, , , , , , , ,
Short vowel + nasal (SV-N)	, , , , ,

Table 1 : Cantonese INITIALs and FINALs

Syllable Structure	# of existing syllables	Examples
D	6	,
LV	3	,
LV-S	4	p, t
LV-N	5	,
SV-S	4	p, k
SV-N	4	, ,
N	2	,
C-D	134	p, t, , , , t
C-LV	82	t, k, , , , t
C-LV-S	117	k^wk, pt, k, k, t, tt
C-LV-N	133	k, k^w, , , , t
C-SV-S	79	pk, tk, t, k, t, tp
C-SV-N	91	t, p, , , , t

Table 2 : Different syllable structures in Cantonese

Cantonese is well known of having nine tones as depicted in Figure 1. They are numbered from 1 to 9 respectively. Tone 1 – 6 are referred as non-entering tones and tone 7 – 9 are referred as entering tones. The primary acoustic feature for Cantonese lexical tones is the syllable-wide F0 profile. Also, entering tones, which are associated exclusively with syllables with stop coda (i.e. /p/, /t/ or /k/), are much shorter than the non-entering tones.

Figure 2 shows the acoustic waveform of a Cantonese utterance, aligned with the time-varying F0 and short-time energy (RMS). The utterance consists of two digit strings separated by a major break in the middle. It is observed that the syllable nucleus (vowel) can be roughly estimated from the peaks in the energy plot. Also, each syllable is made up of an optional unvoiced segment and a voiced segment. If the coda is a stop, syllable duration tends to be short and a closure will follow. Just like in English, sentence-final lengthening is noticeable in Cantonese.

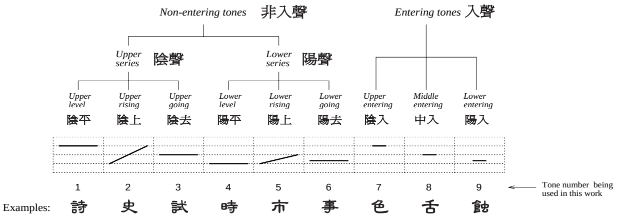

Figure 1. The nine Cantonese lexical tones

It is also obvious that the F0 profile is heavily affected by tonal context. For example, digit “2” (tone 6) occurs four times (labeled as case A-D) in the utterances, and the observed F0 patterns differ greatly among the cases. In case A, F0 keeps rising from a low level. This is because its left context is the lower level tone. In case B and D, where the left context is the upper rising tone, a declining F0 pattern can be observed. Lastly in case C, the slight declination of F0 is caused by its right context which is the lower rising tone. In addition, there exists a long-term and slow declination of F0 across the whole utterance.

3 THE BASELINE TTS SYSTEM

3.1 *The Use of TD-PSOLA*

As described in [1], the baseline system produces synthetic speech by concatenating pre-recorded syllables which have been modified using TD-PSOLA technique to match the prescribed duration or F0 targets. Only the voiced segment of the syllable is subject to PSOLA modification while the unvoiced segment is concatenated as it is.

3.2 *Syllable Inventory*

Undoubtedly prosodic modification by TD-PSOLA would distort the original signal. For the audible distortion to be kept at a low level, the degree of modification should be as small as possible. Therefore tonal syllables have been chosen as the basic templates for synthesis.

We are using the CUSYL database which is designed specially for syllable-based synthesis [5]. It has a large coverage of about 1,800 Cantonese tonal syllables, which include many colloquials and alternative pronunciations. All syllables were recorded from a female native Cantonese speaker.

3.3 *Prosodic Control*

Fixed syllable duration was assumed in the baseline system. The voiced segment of all syllables with non-entering tones were assumed to be 180 msec in length, regardless of their difference in syllabic structure. For all syllables with entering tones (i.e. with coda /p/, /t/ or /k/), a duration of 90 msec was assigned.

For each of the nine lexical tones, a fixed F0 profile was used regardless of any contextual effect.

The baseline system allowed adjustment of duration and F0 at utterance level. That is, speaking rate and F0 dynamic range can be varied by linearly and uniformly scaling the nominal syllable duration and F0 profile.

4 DURATION AND PAUSE CONTROL

Obviously the duration of a Cantonese syllable depends very much on its phonetic content. For example, the voiced segment of a C-LV-N syllable (e.g. /kan/) is longer than that of a C-LV or C-D syllable (e.g. /ka/, /ka /). In this work, we try to obtain:

- nominal duration of the voiced and unvoiced segments in each Cantonese base syllable;
- nominal length of inter-syllable pause between each pair of syllable coda and onset.

4.1 *Speech Database*

We use part of CUSENT, a newly developed Cantonese speech database, for duration measurement. The speech data includes a total of 13,800 continuous sentences from 46 different speakers. The sentence length ranges from 4 – 30 syllables and the average is 10 syllables.

4.2 *Segmental Duration*

Syllable-level time alignment is carried out using HMM forced alignment method. The length of inter-syllable pause is also available from this time alignment. Afterwards voiced/unvoiced detection is performed using the “get_f0” program in the ESPS waves+ software package [6]. The “get_f0” program essentially implements a robust algorithm for pitch tracking (RAPT) base on normalized cross-correlation function [7]. In this way, duration the of voiced and unvoiced segments are derived.

4.3 *Speaking Rate Normalization*

Speaking rate normalization is performed to reduce undesirable variation of segmental duration from utterance to utterance. For each syllable S in an utterance, its local rate of speaking is evaluated as [8],

$$SROS(S) = \frac{DUR(S)}{\mu_{DUR}(S)}$$

where $\mu_{DUR}(S)$ is the mean duration for all occurences of S and $DUR(S)$ denotes the duration in this particular utterance. Then the utterance-level rate of speaking is estimated as the average over all syllabes, i.e.

$$UROS = average_S\left[SROS(S)\right]$$

Both the absolute segmental duration and inter-syllable pause legnth are normalized using the $UROS$.

4.4 *Nominal Duration and Pause Length*

For each of thr 664 base syllables in CUSYL, the nominal duration of its voiced and unvoiced segments are estimated as described above. The results are shown as in Figure 3 and 4. For easy visualization, syllables which similar phonetic structure are grouped together.

Indeed, segmental duration varies greatly from on syllable to another. As shown in Figure 3, the duration of voiced segment in (C)-LV-S syllable is much shorted than those in a (C)-LVor (C)-D syllable. The duration difference between syllables with long vowel and short vowel as nuclei is also quite noticeable.

Figure 5 shows the nominal pause length for different coda-onset combinations. As expected, a short pause needs to be inserted whenever there is a closure between the syllables. This pause may be up to 90 msec if the coda is a stop and the following onset is an unaspirated plosives.

5 CONTEXT-DEPENDENT F0 PROFILE

In this work, we focus on how the F0 profile of a Cantonese syllable may be affected by its left tonal context. Speech materials used for analysis are obtained from a female native Cantonese speaker and make up a total of 4,000 polysyllabic words.

F0 extraction is performed using the “get_f0” program in the ESPS software package, with the syllable boundaries given by HMM forced alignment. All of the F0 patterns are linearly re-sampled to have the same length of 24.

There are 10 possible kinds of left tonal context for each syllable, i.e. tone 1 – 9 and utterance-beginning. An averaged F0 profile is calculated for each context. As an illustrative example, the context-independent and context-dependent F0 profiles for tone 6 are plotted in Figure 6. Overall speaking, tone 6 is featured by a slowly declining F0 pattern. At the utterance-beginning position, the whole F0 profile tends to shift upwards. It also seems that F0 keeps good continuity even across syllable boundaries. As shown in Figure 6, a relatively high F0 is observed when the left context is tone 1 or tone 2 both of which conclude with high F0 level.

6 PERCEPTUAL TEST

6.1 Design of the Test

Subjects are required to listen to pairs of utterances and to grade the utterances in a scale from 1 to 5 (1 being the worst and 5 the best). In each pair, one utterance is generated by the baseline system and the another is the result of either one of the following prosodic controls:

1) Duration and pause only;
2) Context-dependent F0 only;
3) Both duration and F0.

The reference is arbitrarily placed in the first or the second position. A total of 30 sentences have been selected as the synthesis materials. Therefore, each subject has to listen to 90 pairs of synthetic utterances (which are randomly ordered) and give 180 grades. Fifteen subjects participated in the test.

6.2 Results Analysis

For each trial in the listening test, a pair of grades is obtained. Let G_p bet the grade for the utterance with prosodic control and G_b be the grade for the reference utterance. Then the difference G_p - G_b would be a good indication of the relative improvement (or degradation) resulted from the prosodic control. In Figure 7, the histograms of G_p - G_b are plotted separately for the 3 types of prosodic control. It can be observed that there is a marginal but consistent improvement after applying either of the prosodic modification. It is also observed that the effect of duration modification is more prominent than the F0 modification.

7 DISCUSSION & CONCLUSION

Indeed, the improvement attained is marginal. But this is expected for several reasons. Firstly, the overall perceptual naturalness of synthetic speech is affected by many factors which include minor or major breaks at word, phrase or sentence level, stress, intonation, etc. It might be possible that, in fluent speech, the macro-prosodic factors overwhelm the contribution of the segment-level duration and F0 adjustment. Secondly, our duration rules are derived from speech data which are all read newspaper sentences. They usually carry much more than the micro-prosodic effects. Thirdly, the HMM forced alignment method is known to be erroneous. This may affect to certain extent the accuracy of estimated nominal duration. For more reliable prosodic rules, manually labelled speech materials are most desirable. Fourthly, we only consider left tonal context at this stage. This is certainly inadequate as evidenced by the case C in Figure 2. After all, it is our belief that the segment-level duration and F0 control is the first essential step towards natural speech synthesis. In the near future, we will proceed to investigate the long-term prosodic phenomena and properly incorporate them for the betterment of Cantonese TTS technology.

8 RERFERENCES

[1] Min Chu and P.C. Ching. “A Cantonese Synthesizer Based on TD-PSOLA Method”, in *Proceedings of ISMIP-97*, pp. 262–7, Taipei.

[2] E. Moulines *et al*, “A Real-Time French Text-to-Speech System Generating High-Quality Synthetic Speech”, in *Proceedings of ICASSP-90*, Vol.1, pp.309-12.

[3] D. Bigorgne *et al*, “Multi-lingual PSOLA Text-to-Speech System”, in *Proceedings of ICASSP-93*, Vol.2, pp.187-90.

[4] Min Chu and Shinan Lu, “A Text-to-Speech System with High Intelligibility and High for Chinese”, *Chinese Journal of Acoustics*, Vol.15, No.1, pp.81-90, 1996.

[5] W.K. Lo, Tan Lee and P.C. Ching, “Development of Cantonese Spoken Language Corpora for Speech Applications”, in *Proceedings of ISCSLP-98*, pp.102–7, Singapore.

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

[7] D. Talkin (1995), A Robust Algorithm for Pitch Tracking (RAPT), in *Speech Coding and Synthesis* (W.B. Kleijn and K.K. Paliwal eds.), pp.495–518, Elsevier Science B.V., Amsterdam.

[8] Tan Lee, R. Carlson and B. Granstrom, Context-Dependent Duration Modeling for Continuous Speech Recognition, in *Proceedings of ICSLP-98*, Vol.7, pp. 2955–8, Syndey.

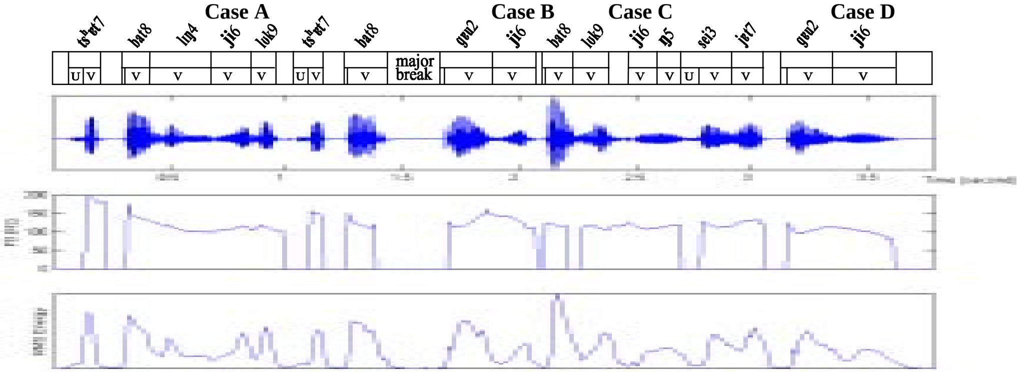

Figure 2: Prosodic structure in Cantonese speech: an example

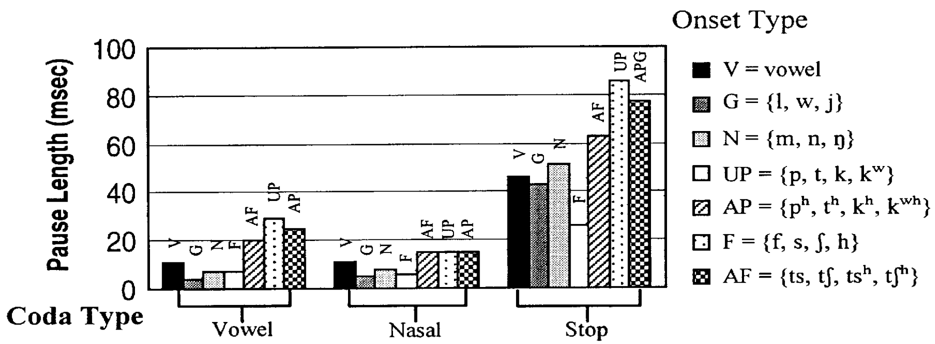

Figure 5: Inter-syllable Pause length for different coda-onset combinations

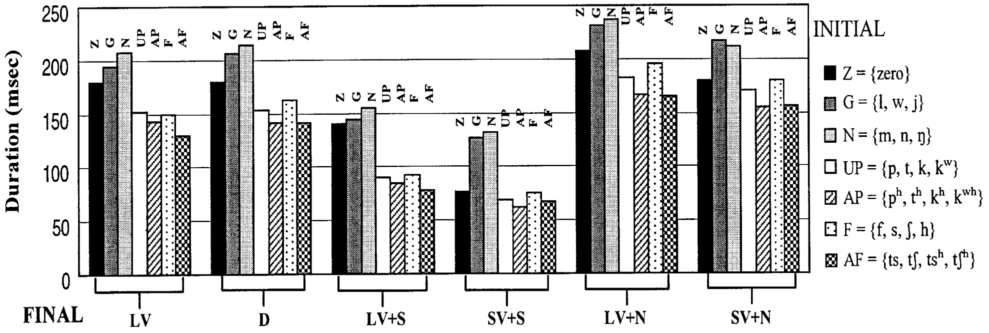

Figure 3: Duration of voiced segment in Cantonese syllables with different phonetic structures. LV: Long Vowel; SV: Short Vowel; D: Diphthong; S: Stop; N: Nasal.

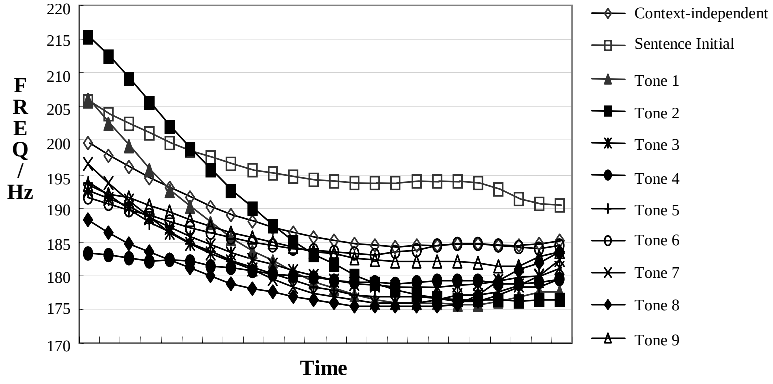

Figure 6: F0 profile of a syllable under different tonal context

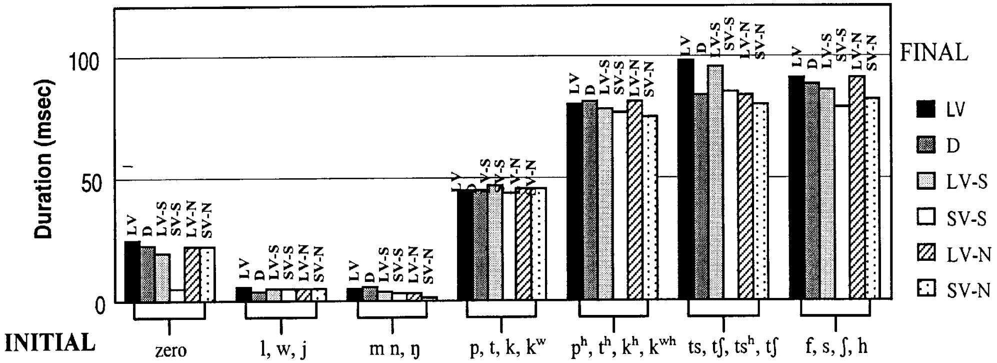

Figure 4: Duration of unvoiced segment in Cantonese syllables with different phonetic structures. LV: Long Vowel; SV: Short Vowel; D: Diphthong; S: Stop; N: Nasal.

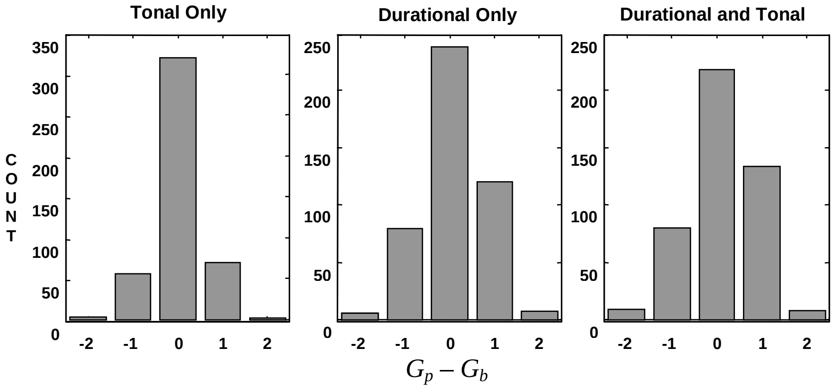

Figure 7: Results of the listening test: histograms of $G_p - G_b$ for different types of prosodic modification