Syllable Fusion and Speech Rate in Hong Kong Cantonese

Wai Yi Peggy Wong

Department of Linguistics
The Ohio State University
pwong@ling.osu.edu

Abstract

Syllable fusion is a Hong Kong Cantonese connected speech process, whereby edges of syllables are obscured by consonant lenition or deletion, and vowel reduction. More extreme fusion can simplify contour tones and merge the qualities of vowels that would be separated by an onset or coda consonant at more normal degrees of disjuncture between words. This paper investigates the influence of speech rate on syllable fusion. An experiment tested the prediction that faster speech rate would give rise to more occurrences of fusion forms and a greater degree of fusion. Subjects repeated word groups in two conditions: at normal rate and at fastest possible speech rates. Results show that speech rate is a reliable predictor for the amount and for the degree of fusion. Implications for incorporating prosody in speech synthesis systems are discussed.

1. Introduction

Traditional descriptions of Cantonese give the impression that Cantonese word structure is very simple. In Cantonese the majority of native morphemes are single syllables, and syllable structure in Cantonese is very simple. In addition to the obligatory tone, each syllable contains one to three segments. A single-segment syllable can be just a syllabic nasal (m̩ or ŋ) or a long vowel. If the nucleus is a long vowel, there can be a following coda consisting of a voiceless stop, a nasal, or a glide. (This coda is obligatory following a short vowel.) A vowel can also be preceded by any of the 19 onset consonants of the language.

Polysyllabic words are traditionally described as compound forms created by stringing together two or more monosyllabic morphemes. However, in connected speech Cantonese polysyllabic words can be shortened, as in the alternation between bisyllabic /mɐt^{5} jɛː^{23}/ 'what' and the "contracted" monosyllabic [mɛː^{55}]. In this paper, I will first describe this phenomenon of "contraction" or "syllable fusion" in more detail, comparing it to phonological phenomena that are associated with (compound) word formation in other languages such as English, and suggesting that syllable fusion is indicative of an intermediate level of prosodic organization between the syllable and the intonational phrase [9] — i.e. the "foot". I will then list factors that should be correlated with incidence of syllable fusion, and describe an experiment in which I tested one of these factors — speech rate. Results of this study will be presented followed by a discussion of the implications of the results for synthesizing prosody for Cantonese.

2. What is syllable fusion?

Past literature had already noted the phenomenon. Yuan et al. called it 'shrink-reduce' [10], which Cheung translated as "contracted form" [2]. Independent of Yuan et al. perhaps, Hashimoto used the term "contraction" to describe the phenomenon [4]. The use of the terminology was followed by Li [6]. The term "contraction" suggests an affinity to English contracted forms such as the alternations between *I am* and *I'm*, *do not* and *don't*, *would not* and *wouldn't* and so on (cf. [1]). This is misleading on several grounds as the examples in the above literature show.

Yuan et al. already seem to have noted there are intermediate degrees of bisyllabic reduction. The form [miː^{55} ɛː^{23}] is a form between bisyllabic [mɐt^{5} jɛː^{23}] and monosyllabic [mɛː^{23}] for /mɐt^{5} jɛː^{23}/ 'what' [10]. (Note that in modern Hong Kong Cantonese the alternation for /mɐt^{5} jɛː^{23}/ 'what' is [mɛː^{55}]. The tonal difference in the "contracted" form [mɛː^{23}] given in Yuan et al. could be dialectal, as their example is based on Guangzhou Cantonese.) Hashimoto also notes cases of bisyllabic reduction when the numeral 'ten' occurs after another numeral (except 'one') to form a digit number, e.g. /paːt^{3} sɐp^{2}/ 'eighty' → [paːʔ^{3} aː^{22}] → [pa^{33} aː^{22}]. "Contraction" has been characterized as the following: in the numeral that precedes 'ten' "the ending stop consonant may become a glottal stop or be completely lost, or the main vowel may be shortened, or both" [4]; "Some contractions involve the deletion of a non-final coda, some involve the deletion of non-marginal coda and onset." [2]; "When strings of words are pronounced casually, some of the juxtaposed syllables pair up and merge." [6]. Cheung distinguishes two kinds of "contracted forms", the less extreme "plain form" (e.g. /pɐt^{5} jyː^{21}/ 'it'd better' → [pɐ^{5} yː^{21}]) and the more extreme "coerced form" (e.g. /pɐt^{5} jyː^{21}/ → [pyː$^{5+21}$]). In reduplicated syllables, "contraction" fuses two syllables into a single syllable with the vowel length prolonged (e.g. /ts^{h}eŋ^{55} ts^{h}eŋ^{55} ts^{h}ɔː^{35} ts^{h}ɔː^{35}/ 'clearly' → [ts^{h}eŋ^{55} ts^{h}ɔː^{35}]; both syllables lengthened) [6]. Thus, there are intermediate degrees of "contraction" when the term is used to describe the phenomenon (see more examples in (1)). In fact, spectrographic evidence shows that the examples in (1) are just a few tokens of an extremely fine-grained continuum, with many intermediate degrees of reduction (see Figure 1). Contrastively, "contraction" as the term is used in English does not indicate the finer-grained intermediate effects.

(1) Examples of syllable fusion forms.

a. 識唔(識) 'know NEG (know)'
   sek^{5} m̩^{21} (sek^{5}) → seʔ^{5} m̩^{21} (sek^{5}) → se^{5} m̩^{21} (sek^{5})

b. 其實 'in fact'
   k^{h}ei^{21}sɐt^{2} → k^{h}e^{21}ɐt^{2} → k^{h}ɐt^{21+2}

c. 知道 'know'
   tsiː^{55}tou^{33} → tsiː^{55}ou^{33} → tsiːu^{55+33}

d. 朝頭(早) 'morning'
   tsiːu^{55}t^{h}ɐu^{21} (tsou^{35}) → tsiː^{55}ɐu^{21} (tsou^{35}) → tsiːu^{55+21} (tsou^{35})

Note also that some of these intermediate forms do not conform to Cantonese phonological structure constraints. As Cheung already noted when he observed "coercion" forms that are clearly monosyllabic in terms of the consonant-vowel structure but bisyllabic in terms of having two different tones specified (e.g. /pɐt^{5} jyː^{21}/ 'it'd be better' → [pyː$^{5+21}$]) [2]. In addition, we can also see that the "coerced" form can violate the constraint against the co-occurrence of a labial onset and a high front rounded vowel within a syllable in the language. By contrast, the English forms *don't* and *wouldn't* are all "legal" forms phonologically, with *don't* having the same general structure as the monomorphemic *point* and *wouldn't* having the same general structure as *bottled*.

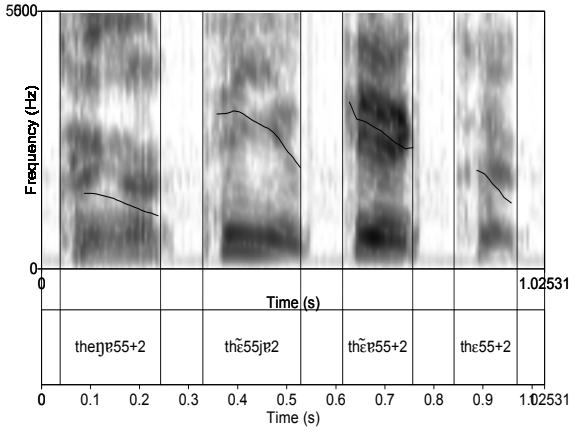

Figure 1: *Four degrees of fusion for the word /tʰeŋ^{55} jɐt^{2}/ 'tomorrow' produced by four speakers. Notice that the word becomes progressively shorter as more segments are deleted and the vowels are merged. Nevertheless, the tones are not deleted. ([h] denotes aspiration in the figure.)*

Finally, fusion seems to be an extremely productive process that can affect *any* sequence of morphemes if the sequence becomes frequent enough. For example, when I went home for a visit after being away in the U.S. where I could not keep up with the local morning traffic news, it took me quite a while to recognize the fused form [fyːŋ$^{5+21}$ (kaːi^{55})] of the street name /faː^{55} jyːn^{21} (kaːi^{55})/ 'Garden Street'. Note also that syllable fusion can occur across word boundary (see the middle case in (1a)). Thus, the fusion process seems to be a productive prosodically driven process that can affect any frequently co-occurring words, rather than a categorical alternation between a longer and shorter form of the most common function words.

If this characterization of fusion is correct, there are factors that should contribute to the likelihood of syllable fusion in present-day Hong Kong Cantonese connected speech. Some of the potential factors include speech rate, prosodic position of participating syllables, word frequency, morphosyntactic structure internal to (compound) words, and word length. Of these potential factors, this paper investigates the influence of speech rate on syllable fusion in Hong Kong Cantonese. I hypothesize that speech rate is a factor that contributes to syllable fusion. My predictions are that faster speech rate would give rise to more occurrences of fusion forms, and that faster speech rate would contribute more occurrences of fusion forms in which vowel qualities of the participating syllables are merged.

3. Method

3.1. Test materials

The stimuli were words and phrases where fusion forms had been found to occur. The selection criteria for the target words were that they (i) should be able to build the story for the story-telling task that elicited fusion forms in spontaneous speech (see [7]); (ii) were judged not to be highly lexicalized (e.g. /mɐt^{5} jɛː^{23}/ 'what' was not selected as fusion target because the realization [mɛː^{55}] or [mɛː$^{55+3}$] for /mɐt^{5} jɛː^{23}/ 'what' may be said to be lexicalized, as suggested by the fact that the Hong Kong Cantonese speakers would represent it with a single Chinese character); and (iii) should not be just occasionally noted to be fused (e.g. /fɔːŋ^{21} wɐi^{35} (wuːi^{35})/ 'The Housing Authority' was not selected, since its fused form [fɐi^{2+35} (wuːi^{35})] was just occasionally noted at the time of experiment).

It is not necessary that all fusion targets have equal likelihood of fusion. Production habits on the part of individual speakers, frequency and segmental components of the words and phrases on the part of the stimuli, and the interaction of these are all possible contributing factors to fusion. As an exploratory work, and with little prior systematic work on syllable fusion in Cantonese, I did not have reliable information to control for factors such as these in the experiment. The fusion targets were then *potential* fusion targets because their fused forms were often observed.

45 words and phrases (30 bisyllabic fusion targets; 15 fillers) were selected, with each consisting of two to four syllables. Three chunks of the two- to four-syllable words/phrases made an utterance. There were 15 utterances. While the words and phrases were meaningful, combining them made the utterances "pseudo-utterances" (i.e. these are not meaningful utterances). Each utterance consisted of seven syllables, except the last two utterances where there were eight. There were 107 syllables in total.

The experimenter read the list of utterances to a metronome at a steady rate (88 beats per minute; 1 beat = 0.68 second). Two syllables made 1 beat. The syllables were not fused. A one-beat pause separated the words/phrases within an utterance. Each utterance was followed by instructions "好快" (very fast) and "正常" (normal) to prompt the speech tempo required. A two-beat pause was provided for subjects to repeat the utterance after the prompt "very fast"; a four-beat pause followed the prompt "normal". The pacing clicks of the metronome were not heard on the test tapes.

The same set of utterances made two tapes, A and B. The two tapes differed only in the order of speech tempo (very fast-normal vs. normal-very fast) required of the subjects. The eighth utterance marked the change in the tempo order requirement. A ten-beat pause preceded the eighth utterance for the experimenter to remind the subjects of the change. Three practice utterances were inserted before the first and the eight utterances.

3.2. Subjects

32 speakers, 16 males and 16 females, participated in the experiment. All speakers were university students ranged between 18 and 26 in age at the time of the experiment. All are native speakers of Hong Kong Cantonese.

3.3. Procedures

Subjects listened over the earphones to the test materials and were recorded in a quiet, empty office in the company of the experimenter. Half of the subjects used test tape A and the other half test tape B. Subjects repeated the pseudo-utterances one by one at two tempi: at their normal speech rate and at their fastest possible rate, as instructed by the aural prompts on the test tapes. A demonstration was given to the subjects on note cards before actual listening and recording started. Subjects did simple arithmetic while repeating the utterances.

3.4. Measuring speech rate

Speech rate is defined as number of syllables per second in this study. I counted the number of "underlying" syllables, including targets and fillers, elicited for each subject in the fast and the normal rate conditions (cf. section 3.5). In each condition, speaker's rate is his/her own average rate across all utterances. Production errors were excluded because they often prolonged the overall utterance duration, obscuring the fact that the speakers produced the targets or fillers at the required rate before and after the error. Production errors excluded were marked from the onset of the errors through the onset the following target item or filler.

3.5. Counting elicited fusion targets

A set of criteria that took into account phenomena found in present-day Hong Kong Cantonese was established to count elicited tokens that were not speech errors. Phenomena relevant for determining whether an error had occurred are: (a) synchronic alternation between initial [ŋ] and null initial, between initial [k^w, k^{wh}] and [k, k^h] respectively before [ɔː], and the more prevalent substitution of final [t] for final /k/ than vice versa; (b) long/short vowel alternations related to literary versus colloquial styles of reading for certain lexical items (e.g. [tsek^5 h**ɐ**k^5~h**aː**k^5] 'immediate'); (c) lexically conditioned segmental alternations (e.g. [**k^h**ɵy^{23}~**h**ɵy^{23} waː^{22}] 's/he said'; [tsiːu^{55} **t^h**ɐu^{21}~**h**ɐu^{21} tsou^{35}] 'morning'); and (d) deletion of place of articulation of final /t/ and final /k/, retaining just a certain amount of glottalization or a glottal stop, even when produced in isolated citation form.

Production tokens counted were the ones that either met the following criteria, or could be interpreted as syllable fusion. The criteria are:

(i) The onset consonant was produced in the form given on the test tapes.
(ii) The onset consonant was not produced in the form given on the test tapes, but it could be ascribed to synchronic alternations in that position (cf. (a) above).
(iii) The vowel was produced in the form given on the test tapes.
(iv) The vowel was not produced in the form given on the test tapes, but it could be ascribed to stylistically governed or lexically conditioned segmental alternations (cf. (b) and (c) above).
(v) The coda consonant was one of the final consonants [p, t, k, m, n, ŋ] or [ʔ] (cf. (d) above).
(vi) The lexical tone was produced in the form given on the test tapes.
(vii) The lexical tone was not produced in the form given on the test tapes, but it could be interpreted as having tonal coarticulation, or tonal target undershoot due to, for example, intonation phrase-final effects, etc.

Both non-fused forms and fused forms were counted towards the total number of elicited fusion targets for each subject. Fusion targets elicited were subject-dependent.

3.6. Selecting two degrees of fusion

Two degrees of fusion were noted. First, there is the deletion of at least one segment contiguous to the syllable boundary between two syllables (or bisyllabic fusion). Second, there is a "merging" of vowel qualities of adjacent syllables to having a single intermediate quality. That is, I arbitrarily chose these two degrees of fusion in the fusion continuum for the current analysis.

4. Results

Each data point in Figures 2 and 3 represents the proportion of fusion forms for each speaker — i.e. the number of fusions (Figure 2) or the number of coalesced vowels (Figure 3), divided by the number of fusion targets that were not produced with speech errors. Figure 2 shows that syllable fusion is highly predictable by speech rate: the faster the speech rate, the more the occurrence of fused forms [r = .854, p =.000, n = 64]. This result may not surprise many, since oral gestures/target shooting is time-linked. Of interest may be how properties of the participating syllables could be merged or changed. Figure 3 shows that syllable fusion with participating vowels coalesced is well-predicted by speech rate: the faster the tempo, the more the fusion forms with participating vowels coalesced [r = .759, p = .000, n = 64].

In the fast rate condition in the two figures the average number of syllables per second across all 32 subjects is 7.40. The total number of fusion targets elicited for all 32 subjects is 897 (or 28 targets per subject in average). Among the elicited fusion targets, 546 (or 61%) exhibit bisyllabic fusion; 159 (or 18%) of the elicited fusion targets exhibit vowel coalescence. In the normal rate condition in the two figures the average number of syllables per second across all 32 subjects is 5.11. The total number of fusion targets elicited for all 32 subjects is 902 (or 28 targets per subject in average). Among the elicited fusion targets, 203 (or 23%) exhibit bisyllabic fusion; 31 (or 3%) of the elicited fusion targets exhibit vowel coalescence.

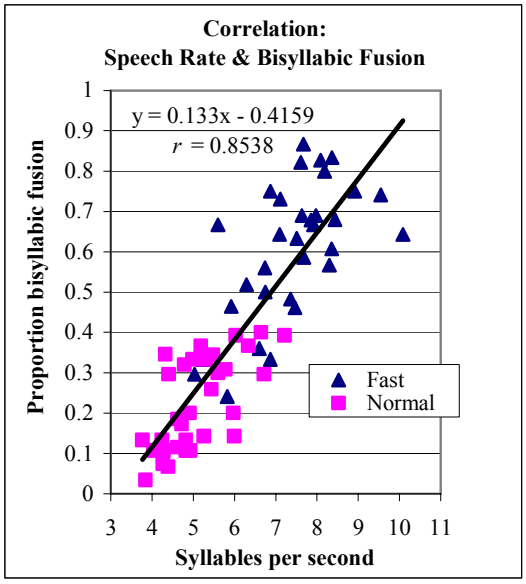

Figure 2

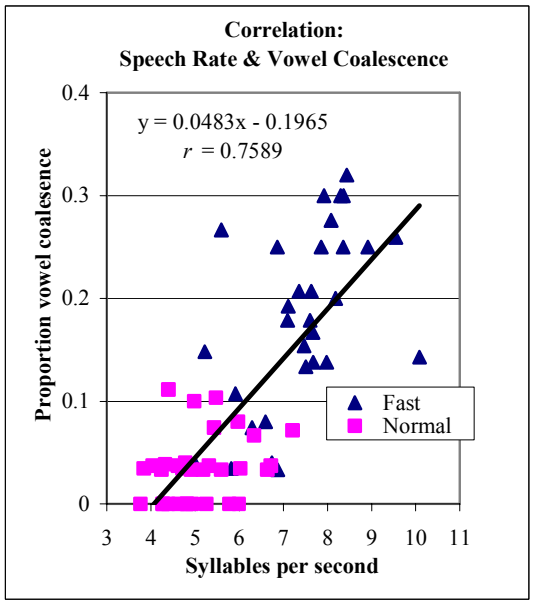

Figure 3

5. Implications for synthesizing prosody

Results of this study show that speech rate is a good predictor for syllable fusion, both in terms of the amount and the degree of fusion. One implication of the results for the application level of developing a Cantonese speech synthesis system is that the capacity to model syllable fusion forms needs to be instilled in the system to ensure naturalness and good intelligibility, if the synthesizer is to model different speech rates in Cantonese connected speech. While the citation monosyllabic flavor is one major characteristic of Cantonese as described in the past literature, one can be certain that simply stretching or shrinking the duration of a string of citation (or non-fused) monosyllables would not be a good idea to model different speech rates when naturalness is desired. Rather, segment(s) contiguous to the syllable boundary would need to be deleted, and participating vowels would need to be changed in terms of duration and/or quality.

Results of this study also shed light on the issue of unit selection for synthesizing Cantonese connected speech using a concatenative approach. The present study shows that in a fused form "within-syllable diphthongs" (see an example in (1b)) or even "triphthongs" (see an example in (1d)) can emerge. Some of these forms could be absent in the standard sound inventory of Cantonese (for example, see [11]). In the "extreme" form of fusion, the merging of qualities of the adjacent vowels can create new C[onsonant]V[owel] or VC sequences. (For example, in /t^{h}e$ŋ^{55}$ jɐt^{2}/ 'tomorrow' → [t^{h}ɛt^{55+2}], while the CV sequence [t^{h}ɛ] is present in the standard Cantonese phonotactics, the VC sequences [ɛt] is not.) These above fusion forms would pose problem to Cantonese speech synthesis systems that are developed based on the standard inventory and phonotactics of Cantonese [3, 5, 8]. Possible considerations for improving the synthesizers include: how to augment the basic inventory of concatenative units with fusion forms; what degree(s) of fusion need(s) to be modeled; whether it is desirable to have a mixed inventory of diphones and triphones; whether some other sizes of units are necessary, etc.

6. References

[1] Bauer, R.S.; Benedict, P.K., 1997. *Modern Cantonese Phonology*. Berlin and New York: Mouton de Gruyter.

[2] Cheung, K.-H., 1986. The Phonology of Present-day Cantonese. Unpublished Ph.D. dissertation. University College London.

[3] Chu, M.; Ching, P.C., 1997. A Cantonese synthesizer based on TD-PSOLA method. In *Proceedings of the 1997 International Symposium on Multimedia Information Processing*. Academia Sinica, Taipei, Taiwan, Dec. 1997.

[4] Hashimoto, O.K., 1972. *Phonology of Cantonese*. Cambridge, England: Cambridge University Press.

[5] Law, K.M.; Lee, T., 2000. Using cross-syllable units for Cantonese speech synthesis. In P*roceedings of the 2000 International Conference on Spoken Language Processing*. Beijing, China, Oct. 2000.

[6] Li, P.Y.-C., 1986. Contraction in Cantonese: A first probe. Manuscript. University of California, San Diego.

[7] Wong, W.Y.P., 1996. Tempo, processing rate and clarity drive in Hong Kong Cantonese connected speech. M.A. thesis. The Hong Kong Polytechnic University.

[8] Wong, W.Y.P.; Brew, C.; Chan, S.D.; Beckman, M.E., 2002. Using the Segmentation Corpus to define an inventory of concatenative units for Cantonese speech synthesis. In *Proceedings of the First SIGHAN Workshop on Chinese Language Processing in conjunction with COLING 2002*, B.K. T'sou; O.O.Y. Kwong; T.B.Y. Lai (eds.), 119-123.

[9] Wong, W.Y.P.; Chan, M.K.-M; Beckman, M.E., in press. An autosegmental-metrical analysis and prosodic annotation conventions for Cantonese. In *Prosodic Models and Transcription:* Towards *Prosodic Typology*, S.-A. Jun (ed.). Oxford University Press.

[10] Yuan, J. et al., 1983. *Hanyu Fangyan Gaiyao*. Wenzi Gaige Chubanshe.

[11] Syllabary of the Jyutping Romanization Scheme, 1993. <http://cpct92.cityu.edu.hk/lshk/>.