Sound Changes in Hong Kong Cantonese: A Multi-perspective Study

WONG, Ying Wai

A Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of
Doctor of Philosophy
in
Chinese Studies

The Chinese University of Hong Kong
August 2011

ACKNOWLEDGEMENT

I am grateful to William S-Y. Wang, my dissertation supervisor. The dissertation topic in fact started from a question raised in oral defense session of my master's degree. His solid foundation in a wide variety of related fields, like phonology / phonetics, acoustics, physiology, neurology, and computational modeling, always gave me support and courage to build linkages between experimental observations from different perspectives.

I thank Thomas Lee of the Language Acquisition Laboratory, my previous supervisor during my master's degree, for his continuous input to my whole work. Thanks to Yi Xu, who from time to time offered me insightful discussion on phonetics research.

During over four years of my PhD study, colleagues (James, Peng Gang, Hongying, Susan, Francis, as well as many other colleagues from the Language Engineering Laboratory) offered me useful criticism and comments at different stages of my current work. The weekly Wednesday seminar always urged me to take a multi-perspective approach to the current study.

Thanks to my classmates in the Centre for East-Asian Studies, who shared many memorable classes, seminar, and discussion with me. Though coming from different countries, with different languages and colors, they offered invaluable moral support towards completion of the dissertation. I am indebted to professors in the same centre, who offered many classes centering around Chinese studies.

There were also numerous anonymous subjects who have participated in my speech production / perception tasks and questionnaire survey. Their voluntary help yielded the various observations obtained in this dissertation. Credits have to be given to Peng Gang for his kind effort in helping subject recruitment for the questionnaire survey.

Though always from a distance, my parents and my two brothers always showed their unfailing concern and love, which were essential in my facing different

stages of work towards producing this dissertation. I am also indebted to my dearest wife, who tried hard to maintain a warm and harmonious environment for my dissertation writing.

Last but not least, I thank God for always leading my way. Without Him, this work would never complete.

ABSTRACT

Speech sound assimilation is a phenomenon widely attested across languages where speech units occurring in proximity in continuous speech become more like each other. From previous literature, there are opposite asymmetries in its direction of operation for segmental feature (e.g. consonant voicing, nasality, place of articulation, etc.) and tonal feature (i.e. fundamental frequency, F_0) assimilations: the former one is dominantly leftward / regressive while the latter one mainly rightward / progressive. This dissertation presents a series of six experiments conducted to investigate the effect of speaking rate on directionality of speech sound assimilation, with Cantonese as the target language. Segmental feature assimilation was studied in the first four experiments: Experiment 1 elicited native speakers' production of target syllables containing two consecutive consonants embedded in a carrier sentence, under different speaking rates. We noticed assimilative contextual effects from onset consonants acting regressively on F_2 (second formant) transition of the previous coda consonant, and that as speaking rate increases, this influence becomes more prominent. Interestingly, this regressive direction of influence agrees with the cross-linguistically dominant (leftward) direction of stop place assimilation; To study the perceptual consequences of the observed rate-induced formant variations, in Experiment 2, we presented to another pool of subjects a subset of samples from Experiment 1 for a consonant identification task. Identification error analysis showed that those acoustic variations indeed caused a corresponding confusion pattern at the listener side, rather than being perceived merely as allophonic variations; Experiment 3 and 4 presented our effort to draw additional empirical support from real-world language uses apart from behavioral data obtained under tightly controlled laboratory settings. Experiment 3 was a frequency analysis into a contemporary spoken corpus on consonant variations. Agreeing with our previous acoustic experiments, it was found that coda consonants were often influenced in an assimilative manner by its following onset consonant rather than the other way round; Experiment 4 compared

two Cantonese spoken corpora constructed in different times for revealing the diachronic trajectory of consonants. Several well-attested changes surfaced in the comparison. Besides, frequency results were matched against results from simulation modeling to provide additional pieces of supporting evidence to our conclusion of regressive segmental feature assimilation; We tried to extend the methodology in Experiment 1 and 2 to study tonal feature assimilation in Experiment 5 and 6. Data from Experiment 5 revealed, under high speaking rate, a rightward shift of certain tonal features, in particular, F_0 peak, relative to segmental boundaries. The corresponding perception Experiment 6 made use of samples from Experiment 5 to elicit subjects' perceptual responses when presented those rightward shifted F_0 contours. Perception results indicated that listeners were able to recover the intended tones even given the distorted F_0 contours, probably with the help of other co-varying acoustic cues present; On a slightly different topic, Experiment 7 presents an attempt to quantitatively document ongoing merger of two tone pairs in Hong Kong Cantonese, namely, T_2-T_5 and T_3-T_6. Results showed statistically significant tone confusion rates across subject ages. Among the two tone pairs studied, the first one was more severely confused while the second one was found to merge at a slightly faster pace. To sum up, since speech rate fluctuation is inevitable in everyday oral communication, if the speech rate-induced acoustic variations obtained in production experiments are found to cause perceptual confusion at the listener side, as exemplified in Experiment 2, in the same direction as reported in the literature on historical sound assimilation processes, speech rate is highly probably one of the candidates driving diachronic speech sound assimilation processes, as found in Experiment 3 and 4. The whole dissertation attempted to re-construct the whole speech sound evolution process with experimental results from multiple perspectives.

摘要

在連續語流中相鄰語音單位變得彼此相似，稱為語音同化現象，在不同語言中都能被觀察到。根據文獻報告，音段特徵（例如聲母清濁、鼻音性、發聲部位等）及聲調特徵（即基頻）在產生語音同化時有不同的方向性：前者主要為逆向（影響力向左），後者主要為順向（影響力向右）。本論文報告一系列六個以粵語為目標語的多角度實驗，探討語速與語音同化作用方向性的關係。首四個實驗研究音段特徵：實驗一要求受驗者以不同的語速發話，以提取試驗句子中兩個相鄰聲母的聲學參數。從對第二共振峰的分析中，我們發現一逆向的同化作用，並此同化作用於高語速時更為明顯。這逆用作用的實驗結果正與學者們報告的跨語言的結果相符。於實驗二中，我們以從實驗一中收集到部分的聲音樣本作為材料，向另一批受驗者播放，要求他們記錄所聽到的輔音。錯誤分析顯示於實驗一中找到因語速導致的聲學變異會令聽者產生相應的輔音感知，即逆向同化。我們在實驗三及四中嘗試從口語語音庫中找尋更多確實証據，以肯定前述從實驗室得到的結果可在日常生活中應用。實驗三對一近代口語語料庫作輔音頻率分析，確認節尾輔音通常會逆向受後接節首輔音的同化影響，亦即跟前述語音學實驗結果相符。實驗四比較兩個於不同年代完成的香港粵語口語語音庫，以觀察輔音的歷時演變，可觀察到很多學者們報告的音變（例如 *n*-去鼻音化）的踪影。另外，輔以電腦模型的結果，我們再次得出輔音主要進行逆向同化的結論。我們採用實驗一及二的方法進行實驗五及六，探討聲調特徵的同化作用。實驗五表明在高語速環境下基頻峰值會有右向的延遲，亦即可視為順向的特徵同化。部分實驗五的語音樣本被作為實驗六的材料，以觀察聽者對這些延遲出現的基頻峰值的感知。實驗結果告訴我們，聽者大概是從基頻以外的聲學特徵準確辨出本來的聲調。環繞另一課題，實驗七以香港粵語兩個廣為談論的聲調合併（即第二五聲，及第三六聲）為背景，作一大規模跨年齡問卷調查。結果顯示不同年齡的參加者對上述四個聲調的混淆程度不一，並年齡

經統計分析為有效的影響因素。其中，第二五聲的合併較明顯，但第三六聲合併速度較快。總括來說，由於語速在日常口語交談中不斷變化，我們証明了在高語速下的發話會導致聽者的混亂（實驗二），而這混亂的方向性與一眾歷時語音演變的結果相符，語速因而極可能是推動語音同化作用的其中一重要原因。本論文正是嘗試從多角度重現整個語音演變的過程。

CONTENTS

CHAPTER 1

INTRODUCTION

Language is seen as a fascinating part of human culture partly because it is full of variations. Properties like word-order, sound inventories, tense and aspect systems all show great diversity cross-linguistically. Much effort has been invested to explore similarities and differences among languages all over the world. If those studies are like taking static pictures of various existing current languages, our view of it is incomplete if their motions are not taken into consideration[1]. The dynamic aspect of language all lies in their development along history. Taking English as an example, the great vowel shift in its sound inventory, morphologization of the adverbial suffix *-ly*, as well as grammaticalization of the lexical verb *will* to a grammatical marker (marking future), are all representative cases illustrating language change at different levels of language structure.

This dissertation, in particular, deals with language change at the level of speech sounds, namely, the directionality in speech sound assimilation. Simply put, assimilation is a process in which neighboring speech sound elements become more similar to each other in terms of different articulatory (e.g. lip rounding, vowel height)

---

[1] The famous statement "Nothing in biology makes sense except in the light of evolution" appeared as title of an essay by Theodosius Dobzhansky (Dobzhansky, 1973), emphasizing the importance of studying evolution, the historical development of biological species, in understanding the interrelatedness of various aspects of biology. We try to adopt this idea here to highlight the significance of studying language change in understanding synchronic language patterns. In fact, several decades ago, Wang (1974) already stated the importance of historical linguistics in studying speech sounds.

or phonatory features (e.g. F_0 (fundamental frequency[2]) level, F_0 slope). Intuitively, adjacent neighbors are the most natural candidate for speech sounds to assimilate to (i.e. *contact assimilation*). Interestingly, a variety of studies, more than that, report cases where speech sounds assimilate to other sound units two or more segments away from them (i.e. *distant assimilation*), e.g. vowel harmony in German[3], consonant harmony in Turkish. This dissertation, however, is limited to discussion of the former one. More will be reviewed in later sections.

Depending on formality or constraints of communication situations, language speakers vary their speech along a continuum of *hyper-* and *hypo-speech* (Lindblom, 1990). In formal or careful speech, speech elements are produced with enough contrast for listeners' accurate perception, e.g. different vowels are articulated with distinctive formant patterns; F_0 contours contain enough F_0 contrasts among the tonal inventory. Chance of mis-comprehension is thus rare. To convey the same meaning in a casual conversation scenario, or under circumstances where there is a tight time constraint, strategies like vowel reduction, the aforementioned speech sound assimilation or even syllable fusion are often observed. These all render the comprehension task more difficult for the listeners. In extreme situations where the listeners fail to perfectly perceive the speakers' intended message, sound change may occur as a result (Ohala, 1981).

---

[2] For a signal with recurrent patterns, such as speech waveform, the smallest repeating unit with period T is obtained. Fundamental frequency (abbreviated as F_0), expressed in Hertz (Hz) or cycle per second (cps), is then obtained by taking reciprocal of it (i.e. $F_0 = 1/T$). It is the acoustic correlate of our impressionistic pitch, and thus a signal with higher F_0 sounds with a higher pitch, and vice versa. While F_0 carries contrasts at various levels in a linguistic system (e.g. *syntactic accent*, *morphological accent*, and *lexical accent* (Wang, 1972)), we only focus on the last one (i.e. *lexical accent*, or referred to as *etymological tone* in Chao (1933)), which is the F_0 variation associated with individual morphemes to signal lexical contrasts.

[3] In German, plural forms of many nouns are derived from their singular forms by appending the suffix *-e* plus vowel fronting. E.g. *Gast* 'guest' > *Gäste* 'guests', *Fluss* 'river' > *Flüsse* 'rivers'. In both cases, the vowels in the singular form are fronted ([a] becomes [e] and [u] becomes [y] respectively) to become more similar to the front vowel [i] in the historical plural suffix *-iz*. This particular anticipatory vowel assimilation in German is also named umlaut by specialist in Germanic.

From the above introduction it becomes apparent that speaking rate is one of the factors behind sound change. This dissertation presents a multi-perspective research with a pursuit along this direction to investigate the effect of speaking rate on directionality of speech sound assimilation. This dissertation consists of 7 Chapters and is organized as follows: Chapter 2 presents different studies related to the current research on topics like speech sound assimilation, temporal organization of syllable; Chapter 3 and 4 report results on segmental sound assimilation from various experimental perspectives including phonetic (Chapter 3), corpus study and modeling (Chapter 4); Chapter 5 tries to extend the speech-rate phonetic experiments to tone assimilation phenomenon, though with partial success. Chapter 6 details a questionnaire survey on the undergoing mergers of two tone-pairs, namely, T_2-T_5 and T_3-T_6, in Hong Kong Cantonese. Finally, Chapter 7 concludes the whole dissertation and outlines direction for further explorations.

CHAPTER 2

LITERATURE REVIEW

While phonological analyses assume concrete units of phonemes, speech sounds are full of variation when examined acoustically. Produced in isolation, sound segments usually show enough contrast to make it distinguishable from other units in the phonological inventory. However, when speech sounds are produced continuously, sound segments appear with much variation depending on their surrounding phonetic environment[4]. Assimilation is a typical case of such context-dependent variation, being "the process by which two sounds that occur close together in speech become more alike." (Trask, 1998, p53). Similarly, according to Laver (2000, p176), "... given the rate that segments follow each other in the stream of speech, that one segment may influence the articulatory characteristics of segments yet to be spoken, or be influenced by those that precede it." More details such as typology of speech sound assimilation, as well as cross-linguistic examples will be presented in Section 2.1.

At a first glance, speech sounds influence each other both progressively (rightwards) and regressively (leftwards), so is the directionality of speech sound assimilation. This intuitive conclusion, however, is disconfirmed by previous studies, where skewed frequency of occurrence with respect to the directionality of speech sound assimilation is reported. We suspect that this directional asymmetry is related

---

[4] Apart from the surrounding phonetic environment, there are yet more factors contributing to the tremendous amount of variations in speech signals reaching our ears: inter-speaker (e.g. F_0 range, intensity), intra-speaker (e.g. time of the day (Garrett & Healey, 1987), emotion (Chuenwattanapranithi *et al.*, 2006)) and properties of transmission media (e.g. background noise level, frequency response characteristics of the transmission medium), just to name a few.

to a number of onset-coda asymmetries revealed in previous literature, which will be surveyed in Section 2.2; Section 2.3 reviews some consequences of varying speaking rate on temporal organization of syllable. All the materials presented will then together lead to our current study.

2.1 Direction of assimilation

One of the major classifications of speech sound assimilation is based on its direction of operation. Consider two sounds S_a and S_b appearing successively in an utterance, *progressive* assimilation[5] is said to occur if S_b becomes more like S_a. For instance, English plural marker *-s* is pronounced as [s] or [z] depending on voicing feature of the final consonant of the base (e.g. [-s] for *books* and *myths* vs. [-z] for *dogs* and *dolls*); For regular English verbs, the past tense marker *-ed* is realized phonetically following a similar logic as [t] or [d] (e.g. [-t] for *laughed* and *walked* vs. [-d] for *cleaned* and *saved*). In both cases, voicing feature of the suffices *-s* and *-ed* changes to become more like their preceding sounds.

Conversely, for cases where S_a instead becomes more like S_b, *regressive* assimilation is said to occur. In colloquial speech of English, *have to* becomes /hæftə/ and *in case* becomes /iŋkeys/ (Hill, 1955). /v/ in *have* is devoiced (to [f]) and /n/ in *in* is velarized (to [ŋ]) to become more similar to the immediately following consonant; English prefix *in-* (meaning negation) appears as *im-* before bilabial consonants in words like *impossible* and *immature*. The alveolar nasal stop is changed to agree in labialization with its following consonant; Latin words *eandem* from *eamdem*, *condere* from *comdere*, and *impar* from *inpar* (Sadler, 1973), all have coda consonant of their first syllable changed to agree in POA (place of articulation) feature with the following consonant (i.e. onset consonant of the second syllable); Another widely cited case is a set of words developed from Latin to Italian: *octo* > *otto* 'eight', *noctem* > *notte* 'night', and *septem* > *sette* 'seven'. In each disyllabic word, stop consonant coda of the first syllable changes to become identical to its following consonant; In

French, there is a well-documented case of diachronic vowel nasalization. Historically, vowels in many words (e.g. *pain* 'bread', *bon* 'good') were nasalized before syllable-final [n] or [m] (i.e. regressive assimilation of nasality), and then the nasal consonants got lost later on. In all the above examples, changes occur on sound segments in an assimilatory manner due to influences from their *following* sounds.

In the segmental assimilation processes listed above, articulatory gestures (e.g. lip-rounding, tongue and velum position) for sound segments appearing in succession to become more alike. The situation is similar in the supra-segmental domain such that phonatory configurations (realized acoustically as F_0 trajectories) of adjacent syllables become more alike. In a classical study on tonal patterns of Tangxi, a Wu dialect, Kennedy (1953) reported a tonal pattern of disyllabic words in which "the original tone of the first syllable is spread over the two combined syllables" (Kennedy, 1953, p369), e.g. *kà*[6] [51][7] 'artificial' + *sāe* [33] 'mountain' to become *kà-sàe* [53-31] 'rockery', and *dhú* [24] 'large' + sāe [33] 'mountain' to become *dhū-sāe* [22-44] or *dhū-sáe* [22-24] 'large mountain'. In both examples here, the original tone of the first syllable spans the total duration of the disyllabic word, as seen by observing initial pitch value of the first and ending pitch value of the second syllable of each disyllabic word; Historically, Gwari, a language spoken in Nigeria, has developed contour tones through assimilative process (Hyman & Schuh, 1974, p88): /òkpá/[8] > [òkpă] 'length', and /súkNù/ > [súkû] 'bone'. In these examples, the low (of /òk/ in /òkpá/) / high (of

---

[5] It must be noted that the terms *progressive* and *regressive* assimilation are found to carry exactly the opposite meaning in some literature. Nevertheless, we adhere to the conventional usage of them in this dissertation (i.e. change in S_a as regressive and change in S_b as progressive).

[6] Diacritics appearing on top of vowels visualize tone of the corresponding syllable, e.g. ā (level tone), á (rising tone), and ă (falling-rising tone).

[7] The numbers enclosed in square brackets here are pitch values which, according to Kennedy's description, denote different pitches (from 1 to 5 in increasing order of pitch), while the precise intervals between these points are immaterial. [51], for instance, represents a sharp fall in pitch. This scheme roughly resembles that of the tone-letter system first proposed by Chao (1930).

[8] In the literature, a different notation (from East-Asian languages) is used to mark tones for African languages: á (high level tone), à (low level tone), ă (rising tone), and â (falling tone).

/sú/ in /súkNù/) tonal feature spreads to the following syllable to yield a rising ([pă] in [òkpă]) / falling ([kû] in [súkû]) tone.

Apart from these, there are some more subtle yet acoustically detectable contextual tonal variations among neighboring tones in continuous speech, as has been demonstrated for languages like Thai (Gandour *et al.*, 1994; Potisuk *et al.*, 1997), Vietnamese (Han & Kim, 1974), Mandarin (Shen, 1990; Xu, 1997) and Cantonese (Liu, 2001; Chang, 2003; Wong, 2006b). According to these studies, these contextual tonal variations exist in both forward (rightward) and backward (leftward) directions. Despite some discrepancies on relative magnitude and nature of such contextual variations (e.g. assimilatory vs. dissimilatory) among these studies [9], tonal assimilation exists progressively. To put it another way, in continuous speech, F_0 contour of a (non-utterance-initial) syllable is affected by tone of the preceding syllable in a way that its initial portion is *raised* by a preceding *high* tone, whereas *lowered* by a preceding *low* tone.

Before proceeding further, it must be noted that speech sound assimilation exists at different levels of language sound system. Some of them developed diachronically (e.g. Italian words developed from Latin, English *in-* prefix, and Tangxi tone-sandhi[10]) and have been fossilized into the corresponding phonological systems. In contrast, others are optional (e.g. colloquial English speech examples from Hill (1955), and contextual tonal variations), selectively applied depending on the communication scenario. A test, though not perfectly accurate, may be applied to separate the two: In a scenario where language users are allowed to produce utterance slowly, formally and carefully, those diachronically developed assimilation rules (e.g. English *in-* prefix) still apply while those optional ones (e.g. contextual tonal variations) diminish to a considerable extent.

---

[9] A survey of such discrepancies and some possible explanations are given in Wong (2007).

[10] The alternation of phonetic shape of adjacent tones when tones come into contact with each other in connected speech, not limited to assimilation. It is also named *neutral intonation* in Chao (1933). Interested readers are referred to Chen (2000) for an extensive list of such processes in Chinese dialects as well as their analyses within several phonological frameworks, and to Wang (1967) for phonological features of tone, with an insightful application of it to formulate tone circle in Amoy Hokkien.

Cross-linguistically, the distribution of two types of assimilation based on directionality is observed to be quite skewed. More than half a century ago, an early report by Kent (1936) already concluded that "An examination of many examples of assimilation and dissimilation of consonants shows that the natural direction of the influence is regressive; I have attributed this to the fact that the thought of the speaker is ahead of his utterance, which tries to overtake the thought, but only at the expense of confusion in the order or the nature of the sounds uttered." (Kent, 1936, p258). Later on, numerous other scholars made a similar observation that regressive assimilation is much more abundant for segmental feature assimilation (Hill, 1955; Ohala, 1990; Steriade, 2001; Jun, 2004, among others).

The aforementioned directional asymmetry can also be observed supra-segmentally. However, interestingly, the dominant pattern is in the opposite direction, mostly rightward (progressive). For instance, Hyman & Schuh's (1974) study of tone rules for West African languages reported that tone-spreading occurs dominantly from left to right (c.f. anticipatory spreading do exist in languages like Auchi and Mitla Zapotec, as mentioned in Maddieson (1978)), where spreading was defined as an assimilatory process. In another study (Zhang, 2005) on Chinese tone-sandhi systems, *left-dominant* sandhi systems (i.e. base tone of the initial syllable in a sandhi domain is preserved while other tones undergo sandhi, also named as *first-syllable dominant* type in Yue-Hashimoto (1987)) tend to involve spreading. In other words, in such tone-sandhi systems, the initial tone within a sandhi domain affects tones rightwards through spreading, which is assimilatory in nature[11].

Up till now, a number of examples of speech sound assimilation processes have been presented. Furthermore, the cross-linguistic directional asymmetry observed for such assimilation processes has been introduced (i.e. dominantly *regressive* for segmental features vs. dominantly *progressive* for tonal features). The next Section reviews some related studies which, we speculate, can assist in giving

---

[11] Here, we do not neglect right-dominant sandhi system, which has abundant presence in southern Wu dialects. Yet, it must be noted that, right-dominant sandhi systems usually involve tonal processes not assimilatory in nature, such as default insertion.

explanation to such directional asymmetry, namely, temporal organization of syllables.

2.2 Temporal organization of syllable

Cross-linguistically, frequency of occurrence of consonants at syllable-initial and syllable-final positions is quite asymmetric: Among languages of the world, CV syllable (i.e. a syllable with a consonant only at syllable-initial position) is the most common type found, and is also the only syllable type present in all languages (Jakobson, 1966, p267). Some languages even do not allow consonants at the syllable-final position. To illustrate these observations, two languages, Cantonese and Japanese are mentioned here. Among the 19 consonants in Cantonese (Zee, 1991), only 6 of them (i.e. [p], [t], [k], [m], [n], [ŋ]) can appear at the syllable-final position. According to a study of spoken corpora of Hong Kong Cantonese (Leung *et al.*, 2004), 87.71% of the syllables used in everyday speech begin with an onset consonant, while only 32.23% (34.9% according to Fok (1979)) of them end with a coda; A typical syllable in Japanese is of CV structure (e.g. *hana* 'flower', *kagi* 'key'), except the alveolar nasal [n], which can be attached syllable-finally to form CVN syllables (e.g. *san* 'three').

This differential status between syllable-initial and -final positions is also reflected in a cross-linguistic survey of 104 languages (Greenberg, 1978). At syllable-initial position, 90 languages among the surveyed 104 languages (86.54%) allow consonant clusters, while the number of languages is reduced to 62 (59.62%) for syllable-final position. Judging from syllable-initial position's higher capacity to accommodate consonant sequences, that syllable-initial position is more suitable than syllable-final position is somewhat confirmed.

According to Krakow's (1999) review of physiological literature on speech production, syllable-initial and syllable-final consonants behave quite differently in terms of their physiological settings. For example, alveolar stop consonants /t/ and /d/, are associated with greater area of tongue-palate contact (measured by a palatometer) at syllable-initial position (e.g. *timid*, *deaf*) than at syllable-final position (e.g. *limit*, *fed*), according to palatometer data; Velic lowering, which allows air to escape through the nose, making nasal sounds possible, is linked to different events depending on position of the nasal inside a syllable: For syllable-initial /m/, offset of

velum lowering is closely linked in time with lip raising offset, while for syllable-final /m/, it is closely linked in time with lip raising onset instead. These observations exemplify that, from physiological measurements, syllable-initial and syllable-final consonants differ both in terms of *intra-articulator organization* and *inter-articulator timing*.

Fujimura *et al.*'s (1978) perception experiment confirmed that CV transition is perceptually more salient than VC transition. In their experiment, VCV sequences are synthesized such that conflicting information about the place feature of the intervocalic consonant is present in the stimuli. An utterance synthesized from /ab-/ and /-ga/, for instance, leads to consonant perception of /b/ if the VC transition (from /a/ to /b/) is more relied on, but perception of /g/ if instead the CV transition (from /g/ to /a/) is perceptually dominant. When presented these synthetic stimuli, subjects consistently made their consonant identification according to the CV transition (with the above example, results are dominantly /g/). In order to eliminate the physical asymmetry from speech production factor (e.g. position-dependent different duration and slope of formant transitions), stimuli were played back in both forward (natural) and backward (reverse) directions. Even with this experimental control, subjects still identified the consonants according to the CV transitions (in the reverse playback condition, these transitions were produced in natural speech as VC transitions).

After reviewing phenomena of speech segments across positions in a syllable, we proceed to tone, the supra-segmental component of speech. When tones appear in succession in continuous speech, their influences on the preceding and the following tones are not symmetric. The progressive effect is assimilatory, as mentioned before, while the regressive one is dissimilatory in nature. Also, more importantly, the magnitude of variation is much larger in the progressive direction than that in the regressive direction (Xu, 1997; Wong, 2006b, among others).

Perceptually, tones also show position-dependent effect. With a split-syllable design, Wong's (2006a, 2007) perception experiments confirmed that the second half of a syllable contributes much more than the first half in perception of the associated

tone. To rule out the factor of intensity differences across portions of a syllable[12], as a follow-up, Wong (2007) flattened intensity profile of the target syllables and presented them to subjects for tone identification. The same bias was consistently obtained, indicating an inherent perceptual bias towards the second half of a syllable in tone identification.

So far, we have reviewed a number of studies related to syllable structure in human speech. These studies are from different perspectives, production, perception, physiology, etc., yet they all demonstrate the complicated nature of a syllable: for segments, onset is more prominent than coda, for tone, the second half is more salient than the first half[13] (Wong, 2007). Next we step forward to the speech rate-dependent behavior of a syllable.

2.3 Rate effect on syllable organization

Kent *et al.*'s (1987) report summarized maximum performance for a variety of measures related to speech production, e.g. maximum phonation volume, maximum vocal intensity, fundamental frequency range, maximum repetition rate. One of the conclusions made was that, the demand for ordinary speech communication all falls within the limit of their measures of maximum performance, with one exception, namely, the maximum repetition rate, expressed as either syllables/second or phones/second[14]. In other words, ordinary speech communication may already be

---

[12] The general pattern is that the vocalic part, which often resides in latter portion of a syllable, is more sonorous. This uneven intensity profile pattern can be argued as the sole factor behind the perceptual bias observed in Wong (2006a). As a result, some further follow-ups like that in Wong (2007) became necessary.

[13] It is suggested that the first half is more salient than the second half in the falling tone case. Since Wong (2007) used Cantonese as the target language, which only contains level and rising tones, such opposite pattern about falling tone is yet to be verified.

[14] Kent *et al.*'s (1987) measure of maximum repetition rate mainly deal with performance limits of different articulatory organs like the jaws, the tongue, etc., in rapid syllable production. Along the same direction, Xu & Sun's (2002) study evaluated performance ceiling of human phonatory organs in

conducted at a maximum rate allowed by our physiological characteristics. At speaking rates faster than that, no matter due to whatever reasons, speakers either have to employ compensatory strategies to cope with the physiological limit[15] or otherwise the speech production task cannot be accomplished due to failure in accurate temporal sequencing and coordination of different involved speech production organs.

Tuller & Kelso's (1990) study revealed differential stability of consonants in prevocalic and postvocalic positions. In their speech production experiment, participants were requested to produce repetitively the utterances /ip/ (/ip-ip-ip-.../) and /pi/ (/pi-pi-pi-.../). Initially, utterances were produced at a slow speaking rate, and during a production trial (without breath), the experimenter signaled the subjects to speech up approximately 4 to 6 times. Acoustic speech signals, movement of glottis and lips, and intraoral pressure were simultaneously acquired during the speech production tasks.

The major result was that upon speaking rate increases above a certain speaker-dependent threshold, phase relationship between glottal and lip movement changes such that the utterance /ip/ becomes to /pi/, while the utterance /pi/ remains the same, as /pi/. This physiologically detected phase transition[16] (from /ip/ to /pi/) was reconfirmed by a follow-up perception experiment. The observed shift from VC to CV sequence but absence of shift from CV to VC sequence in this study indicates that CV sequence is more stable. Formulated another way, at higher speaking rate, the consonant at coda position (C in VC) is possibly the earlier component (compared to consonant at the onset position) to change (due to failure of inter-articulator

---

producing changing pitches. They arrived at a similar conclusion: the maximum speed of pitch change is often approached in speech.

[15] Also treated as part of the operating characteristics (OC) in speech mechanism (Wang, 1972).

[16] Phase transition is not unique to speech production. For bimanual periodic movements, such as finger wiggling, it is well observed that above a certain repetition rate, the bimanual coordination pattern always shifts abruptly from an asymmetric out-of-phase mode to a symmetric in-phase mode. Interested readers are referred to Kelso (1984).

coordination or speakers' intentional adjustment) to meet the corresponding higher physiological demand.

From Gay's (1978) acoustic study, it is found that under high speaking rate, different components, namely prevocalic consonant, vowel, and postvocalic consonant, of a syllable contribute differently to utterance duration reduction. More precisely, vowels always show the largest proportion of durational reduction. While for consonants, the postvocalic consonants are reduced proportionally more in duration than the prevocalic ones. Taking Hong Kong Cantonese as the target language, Zee's (2002) study made a similar observation. When the syllable /sa55/, for example, is produced under high speaking rate, whereas the consonant /s/ is reduced by 45.29%, the vowel /a/ is reduced by a larger ratio of 55.69%. These two acoustic studies are common in reporting the differential contribution among speech segments to the durational reduction of the whole utterance during high speaking rate.

To summarize, this Chapter reviewed previous literature on speech sound assimilation, both for segmental and supra-segmental components. It has been shown that segmental assimilation is dominant regressively, while tonal assimilation progressively. Next, reports on asymmetries in temporal organization of syllable from different perspectives were introduced. Following that, results on temporal organization of a syllable under high speaking rate were reviewed. From these reports, we suspect that increased speaking rate leads to temporally asymmetric adjustment of syllable organization, which is in turn related to the directional bias in speech sound assimilation. Our current study is to explore the possibility to establish a link between speaking rate and direction of of feature assimilation. While previous studies gave a general picture of differential adjustment of different parts of a syllable under increased speaking rate, we contribute by exploring more detailed dynamics of speakers' strategies when the speaking rate is raised beyond their ordinary one. In particular, acoustic measurement will be made to detect such rate-dependent speech signal variation, which will be presented to subjects for perception tasks. Besides, comparative corpus studies will be carried out, supplemented by modeling effort, to draw relevant empirical evidence from real-world languages. Results thus obtained, we expect, will lead to a convincing account of the observed directional asymmetries in speech sound assimilation, with a single factor *speaking rate*.

CHAPTER 3

SEGMENTAL CHANGE – PHONETIC EXPERIMENTS

This chapter presents two phonetic experiments on the effect of *speaking rate* on segmental features. First a production experiment (Experiment 1) is conducted to elicit behavior of formant trajectories under different speaking rates. Following that, a perception experiment (Experiment 2) is carried out to verify the perceptual consequences of the obtained rate-dependent acoustic variations.

Many features like nasality, voicing, labialization, etc., can take part in segmental feature assimilation, as exemplified in Section 2.1. In this study, we limit our scope of investigation here to first deal with *stop place feature*, namely, distinction between bilabial, alveolar, velar stops ([p], [t] and [k]). In particular, F_2 (second formant) movement becomes subject of our study since it has been demonstrated by Liberman *et al.* (1954, 1957) that direction and degree of such F_2 transitions can serve as cues for stop consonant distinction[17]. Experiment 1 examines stop place distinction manifested in terms of F_2 movement.

With these production data, Experiment 2 is conducted to imitate an ordinary speech communication situation under different speaking rates: Participants in this

---

[17] As shown by Lisker (1986), a list of 16 acoustic feature differences can be listed for the voicing distinction between the pair *rapid-rabid*. Similarly, we do not claim F_2 transition as the only feature of stop place. In fact, apart from formant transitions, Wang's (1959) study revealed that, in English, the release associated with syllable-final stops serves as a salient perceptual cue to stop identification. However, it must be noted that in Cantonese, our subject language, syllable-final stops are unreleased (Zee, 1991). Consequently, F_2 transition plays an even more dominant role in coda stop identification.

experiment will be required to identify utterances from Experiment 1. Test results will be matched against the target sentences presented to subjects to relate speaking rate and listeners' identification performance. Cantonese is chosen as our target language. Results obtained in our study supposedly reflect physiological and perceptual constraints of language users in general (known as operating characteristics (OC) in Wang (1972)) and is expected to be applicable to other languages.

3.1 Speech Production - Experiment 1[18]

Aim of the current experiment is to elicit second formant (F_2) transition trajectories of different intervocalic stop consonants. With different speaking rate conditions, this experiment will (1) demonstrate rate-dependent bias in F_2 movement, and (2) provide data for the subsequent perception Experiment 2.

Material

In this experiment, there are two consecutive target syllables S_1 and S_2, embedded sentence-medially in the carrier test sentence is *ngo5*[19] *soeng2 maai5* $\underline{S_1}$ $\underline{S_2}$ *nei1 zek3 zau2* 'I want to buy the wine named $\underline{S_1}$ $\underline{S_2}$', where S_1 is one of the three syllables *laap6* 'to stand' [lap], *laat6* 'spicy' [lat], *laak6* 'to tighten' [lak], and S_2 one of the three syllables *baa1* 'bus' [pa], *daa1* 'a dozen' [ta], *gaa1* 'plus' [ka]. With this design, formant transitions due to different stop place specification (as either bilabial [p], alveolar [t] or velar [k]) at both syllable-initial and syllable-final positions can be compared, and their mutual influence, if any, can be examined. The same long vowel [a] is used for both target syllables to minimize vowel-dependent formant variations (Peterson & Barney, 1952). Possible bias from the lexicon is minimized through the use of bigram S_1-S_2, where all the nine combinations of it carry no meaning in

---

[18] Part of the results reported here appeared previously in Wong (2009a).

[19] Unless otherwise specified, *Jyutping* (LSHK, 1997), romanization system adopted by LSHK (Linguistic Society of Hong Kong) is used to transcribed Cantonese. Syllables in *Jyutping* are expressed by concatenating segmental and tonal labels, for instance, *baa1* 'father' refers to a syllable *baa* ([pa]), associated with T_1 (high-level tone). Note that their system labels the three traditional entering tones, from high to low pitch, as T_1, T_3 and T_6 respectively.

Cantonese. Both the lateral consonant [l] from S_1 and the nasal murmur associated with post-target *neil* provide characteristic acoustic landmarks for accurate and objective segmentation. There are altogether (3 coda stops) × (3 onset stops) = 9 test sentence combinations. Note that all the syllables for S_1 and S_2 are of an identical tone (T_6, mid-low level tone for S_1 and T_1, high-level tone for S_2).

Recording procedure and data extraction

Five native Cantonese speakers (5M) were recruited to take part in the speech production task. Recording sessions were conducted in a quiet room, with a SONY ECM-MS907 electret condenser microphone. During the recording sessions, each of the test sentences was produced at various speaking rates for 20 repetitions (10 for data analysis, remaining 10 as spare). We had four levels for the *speaking rate* factor in this experiment. To control the speaking rate, a series of *pace-keeping clips* were played back to subjects during the course of recording. These clips consisted of a chain of "ding" sounds linked up by silence intervals of fixed durations (from 500ms to 2000ms, in the step of 500ms). Subjects were instructed to make their best effort to articulate the presented test sentences within those silence intervals. Each subject thus had to produce (3 coda stops) × (3 onset stops) × (4 speaking rates) × (20 repetitions) = 720 utterances to complete the task. Recordings were done with PRAAT (Boersma & Weenink, 2005), at a sampling rate of 22.1kHz. Utterances thus obtained were fed into the same software for segmentation, which was based primarily on formant patterns (such as the aforementioned nasal murmurs), with the aid, if necessary, from intensity profile fluctuations also. Then a PRAAT script had been coded to obtain formant values from the segmented utterance portions. Twenty samples at 5% temporal steps were extracted per portion.

Results and discussion

Figure 3.1 plots durational measures, across different speaking rates, of three segments relevant to the current experiment, namely (1) F_2 for S_1, (2) the silence portion between S_1 and S_2, and (3) F_2 for S_2. For each segment, duration surfaces as a strictly increasing series (e.g. 119ms, 141ms, 163ms, 177ms for the first segment from fast to slow speech) as speaking rate decreases (from 500ms to 2000ms

conditions), indicating that our pace-keeping clips served their role well in regulating subjects' speaking rate.

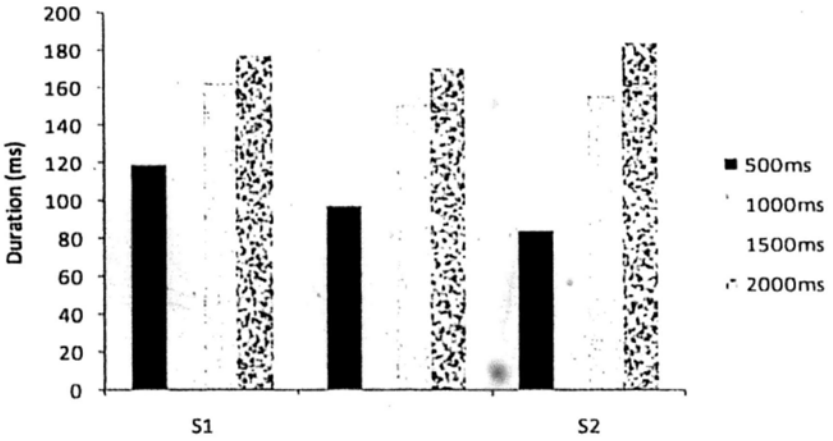
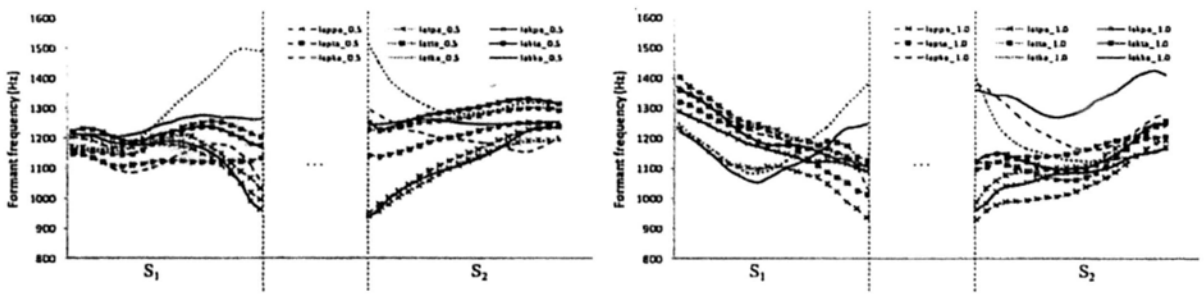


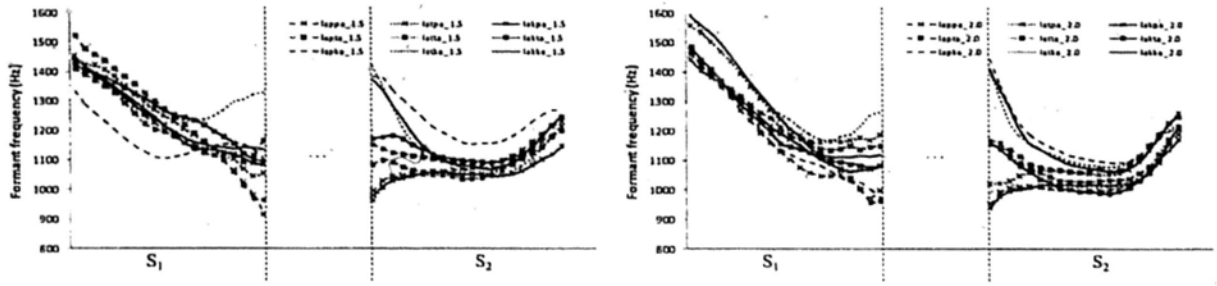
Figure 3.1: Durational measures for 3 segments (F_2 for S_1; silence portion between S_1 and S_2; F_2 for S_2) obtained after segmentation.

Following this, F_2 trajectories for the target syllable bigram S_1-S_2 across different intervocalic stop combinations and speaking rate conditions (500ms, 1000ms, 1500ms, and 2000ms) are plotted in Figure 3.2[20]. Each curve results from averaging 10 repetitions. The general pattern of these F_2 trajectories is to start at a narrow range of formant frequencies, diverge when approaching offset of S_1 (VC formant transition due to coda consonants), appear to come from a wide range of frequencies (CV formant transition due to onset consonants) at S_2 onset and finally converge again.



(a) 500ms (b) 1000ms

[20] The plots are based on data of Subject 1. Unless otherwise stated, all following relevant description and discussions are applicable to data from all the subjects.



(c) 1500ms (d) 2000ms

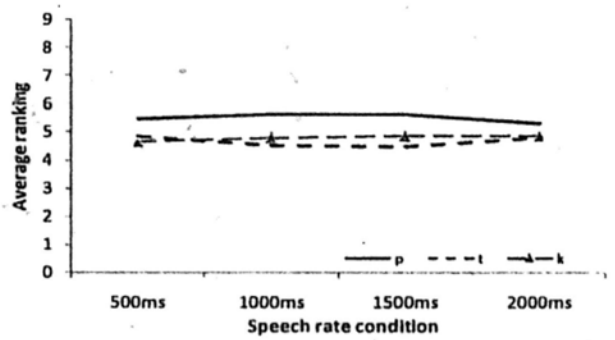
Figure 3.2: Consonantal perturbation of F_2 trajectories for different S_1-S_2 combinations across speaking rates. Data points are plotted time-normalized.

We may gain some insights into differences between onset and coda consonants by comparing the VC and CV transitions. Here we take the 1500ms condition (in Figure 3.2c) as an example for the following discussion, since the corresponding speaking rate approximates common conversation scenario the most. Depending on coda consonants, F_2 frequency terminates at S_1 offset from 918.34Hz (*lap-pa*) to 1329.51Hz (*lat-ka*), spanning a range of 411.17Hz. At S_2 onset, the corresponding range of initial F_2 frequency is 454.84Hz, from 964.65Hz (*lat-pa*) to 1419.49Hz (*lat-ka*); F_2 contours in S_1 overlap up to approximately 75% of the total duration before they diverge, while the non-overlapping proportion of F_2 contours in S_2 can be seen to be larger, going about 50% into vowel of S_2; Terminal F_2 frequency at S_1 offset, sorted in ascending order, is *lap-pa* < *lap-ta* < *lat-pa* < *lak-pa* < *lak-ta* < *lap-ka* < *lak-ka* < *lat-ta* < *lat-ka*. Roughly speaking, contours with bilabial stop /p/ as coda (*lap-pa*, *lap-ta* and *lap-ka*) end lower while the remaining ones with /t/ or /k/ as coda end higher; Compared to the VC transition, clustering pattern of CV transition is much clearer. The series, in ascending order of initial F_2 frequency at S_2 onset, *lat-pa* < *lak-pa* < *lap-pa* < *lat-ta* < *lap-ta* < *lak-ta* < *lak-ka* < *lap-ka* < *lat-ka*, clearly shows three groups of contours depending on the onset consonant associated with S_2. Specifically, F_2 contours with /p/ as onset begin at the lowest value, followed by those with /t/, and finally those with /k/ as onset, agreeing with results by Öhman (1966) on the voiced series.

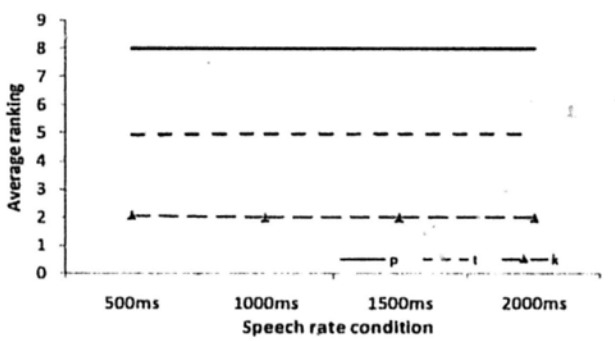
Essentially, CV transitions show a clearer clustering pattern than VC transitions, and this is true across all speaking rates. Upon closer examination, however, relative ranking among different F_2 trajectories are observed to vary across

speaking rates. To facilitate comparison, we numerate the rankings with numbers from 1 to 9, corresponding to items with the highest value to the lowest value. VC transitions for the 500ms condition, for instance, are in the descending order of *lat-ka* > *lak-ka* > *lat-ta* > *lak-ta* > *lap-ta* > *lap-ka* > *lat-pa* > *lap-pa* > *lak-pa*. In such case *lat-ka* is assigned a ranking score of 1 while *lak-pa* 9.

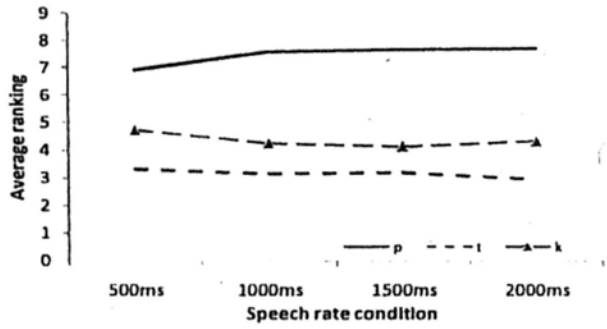
With the aforementioned ranking scheme, we proceed to study the effect of speaking rate on F_2 movement by comparing average ranking scores with different factors fixed, as shown in Figure 3.3. First, Figure 3.3b plots the average ranking of F_2 trajectories at CV transition (i.e. onset of S_2) given different S_2 onset consonants. The plot shows that average ranking of F_2 contours is almost constant under different speaking rates, that CV transition with [p] as S_2 onset (i.e. *lap-pa*, *lat-pa*, *lak-pa*) starts at the lowest value (average ranking of 8), while that with [k] (i.e. *lap-ka*, *lat-ka*, *lak-ka*) the highest (average ranking of 2). This correctly reflects the clear and non-overlapping clustering pattern at the CV transition portion, as previously shown in Figure 3.2; Figure 3.3a plots the average ranking of CV transitions given different S_1 coda consonants. The rankings centre around 5, with crossings among individual conditions. Taking Figure 3.3a-b together, we may conclude that speaking rate has no obvious effect on CV transition under our current experimental settings.



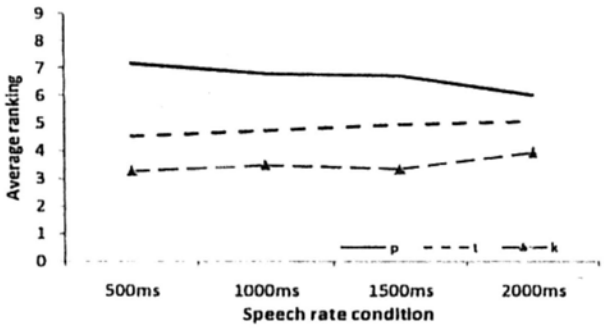
(a) Avg. CV ranking | S_1 coda



(b) Avg. CV ranking | S_2 onset



(c) Avg. VC ranking | S_1 coda



(d) Avg. VC ranking | S_2 onset

Figure 3.3: Average rankings of CV (3a-b) and VC (3c-d) transitions with different factor of interest.

Next, Figure 3.3c-d show respectively average rankings of VC transitions with respectively different S_1 coda (Figure 3.3c) and S_2 onset consonants (Figure 3.3d). Figure 3.3d is particularly interesting: As speaking rate increases (from 2000ms condition to 500ms conditions), *F_2-lowering* is observed for tokens with [p] as S_2 onset (i.e. *lap-pa*, *lat-pa*, *lak-pa*), *F_2-raising* instead for tokens with [k] as S_2 onset (i.e. *lap-ka*, *lat-ka*, *lak-ka*). Recall from Figure 3.3b that [p] onset is associated with the lowest F_2 while [k] onset the highest F_2, Figure 3.3d seems to suggest that as speaking rate increases, VC transition is *regressively* influenced more and more prominently by the following consonant (i.e. S_2 onset in this case). Conversely, from Figure 3.3a-b, *progressive* influence is not observed for a preceding consonant on the following CV transition. Simply put, an asymmetry is observed here that for two consecutive consonants in speech, as speaking rate increases, it is the *second consonant* which exerts larger and larger effect on the first consonant. In other words, the first consonant is more readily reduced in such situation.

3.2 Speech Perception - Experiment 2

Ladefoged & Broadbent (1957) pioneered to study the contextual effect of a carrier sentence on speech perception. Depending on formant frequency settings of the carrier sentence, a fixed target word yielded different perception. Later on, scholars extended the paradigm to supra-segmental domain of human speech. Numerous speech perception studies (Lin & Wang, 1985; Wong, 1999; Wong & Diehl, 2003; among others) reported that context is heavily relied on in tone perception. Parallel to

perception of segments, the target syllable was perceived to be associated with tonally different as F_0 level of the carrier sentence alone was raised / lowered.

The influence of context on the target probably reflects our perception system's effort to restore the linguistic message intended by speakers. Miller *et al.* (1986) reported such an effort in speech rate normalization. It is widely attested that as speech slows down, *voice onset time* (VOT), the primary acoustic cue for voicing distinction, increases. In their experiments, the perceptual boundary in VOT for voicing distinction was found to be larger when subjects were presented with slow speech, compensating for the rate-induced production bias. Equally, listeners have to possess the strategy to remove voice pitch variations of individual speakers. In several speaker normalization studies (Leather, 1983; Moore & Jongman, 1997), Mandarin subjects adjusted their identification responses of the target words according to the carrier sentences with F_0 manipulated to reflect different speaker characteristics.

In extreme conversation scenario where a target linguistic unit is masked for some reasons, there is still the possibility for one to recover the masked target, as demonstrated by the pioneering work on phonemic restoration by Warren (1970). In their listening experiment, part of the speech sample was replaced with a segment of cough, and presented to subjects for identification of the position of the sound replacement. Surprisingly, majority of the subjects did not report any missing sound and that they could not identify accurately position of the cough. In the supra-segmental domain, we conducted tone perception experiments (Wong, 2008), successfully eliciting subjects' ability for 'tonemic restoration'.

The aforementioned studies all point to the same capability of the human perception system to make use of various cues to remove acoustic variations due to context, recovering the linguistic units intended by speakers. This observation highly relevant to our study: There are certainly numerous acoustic features co-varying with the formant lowering / raising as found in Experiment 1[21]. If human perception system

---

[21] Although our emphasis in Experiment 1 is on F_2 variations, there must be various other acoustic cues co-varying with F_2 across speaking rates. Lisker (1986) set a good example demonstrating this acoustic co-variance in human speech. While the pair *rabid-rapid* differ from each other minimally in the

can accurately recover the stop consonants from those co-variances, it is just another instance of speech rate normalization, without much contribution to sound change. Following that, a corresponding perception experiment is necessary to verify this conjecture.

Methodology

To preserve all the possible co-varying acoustic cues in the speech signal, we make use of the speech tokens obtained from Experiment 1. From each subject in Experiment 1, one token was drawn randomly from each condition. Altogether, we got 180 tokens (5 speakers × 9 coda-onset combinations × 4 speech rates) for the current perception test. We had considered making use of more repetitions to better elicit listeners' performance when presented a particular stimulus, but finally abandoned the idea as doing so would lead to the undesirable subject fatigue. To minimize possible influence due to intensity variation across different subjects / tokens, all the sound samples were intensity-normalized to 80dB.

Nine native speakers of Cantonese (6M3F) participated in the experiment. All of them did not take part in the production experiment (Experiment 1) leading to the current set of stimuli, and thus had no bias from listening to their own voice. The 180 tokens, grouped in a randomized order into 20 blocks of 9 tokens, interleaved with silence interval of 1.5s (i.e. ISI = 1.5s), were presented to subjects for character identification task. The experiment was carried out in a quiet venue, with speaker volume tuned to a comfortable level. The participants were requested to identify the two target characters (S_1 and S_2) in the stimulus sentence "*ngo5 soeng2 maai5* S_1 S_2 *nei1 zek3 zau2*" 'I want to buy the wine named S_1 S_2' by marking the corresponding Chinese characters on an answer sheet. Six characters (three for coda combinations (i.e. *laap6*, *laat6*, *laak6*), the remaining three for onset combinations (i.e. *baa1*, *daa1*, *gaa1*)) were given as choices for each question item.

---

phonological feature of voicing, the author argued that as many as 16 acoustic features can contribute to their contrast, like F_0 *contour* and *duration of closure*.

Results and discussion

Figure 3.4 shows the resulting error rates of the target coda and onset identification. For coda identification, the error rate drops from 42.72% to 34.07% when speaking rate decreases from 500ms to 1500ms condition, and rises again for the even slower 2000ms condition. For onset confusion, the error rate fluctuates between 4.44% and 9.38%. Comparing the two target consonants, error rate is much higher for coda than onset, which is a compatible consequence from the clearer clustering pattern of onsets (compared to codas) as observed in Experiment 1.

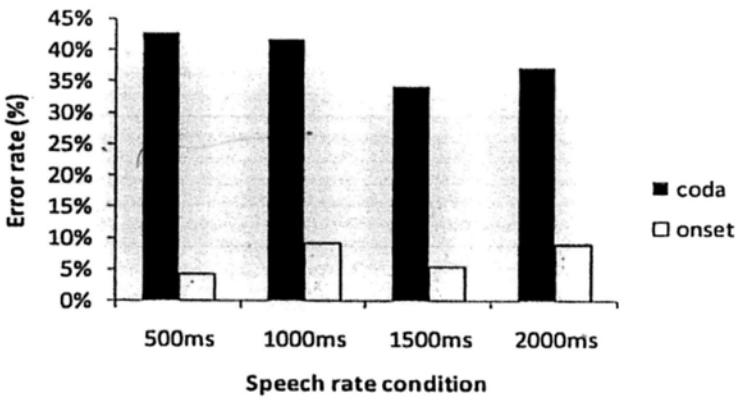
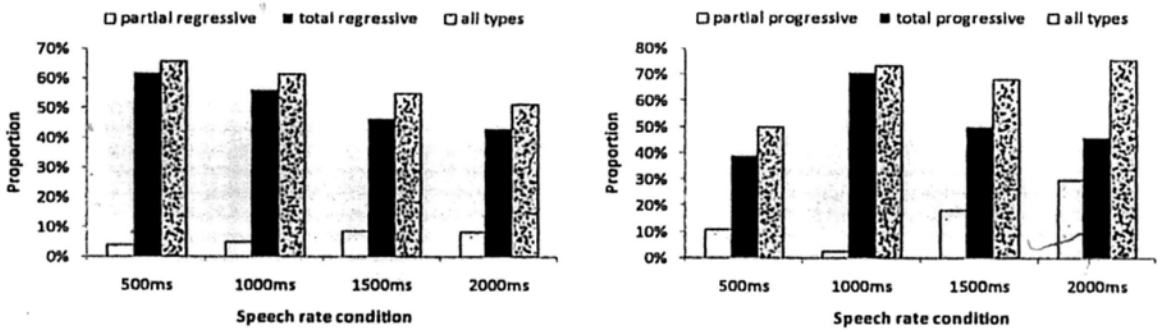


Figure 3.4 Error rate of consonant identification.

A grouping of identification errors according to individual consonants reveals that consonants at onset and coda positions are considerably different. For consonants at the coda position, error rate follows the order of the target coda consonant being *p* (22.96%) < *t* (43.89%) < *k* (50.00%). For onset identification, 98.26% of the identification errors go to the case when the target onset consonants were *p*- (9.63%) and *t*- (11.30%), and there are merely only two errors (0.37%) obtained if the target onset is *k*-. The outstanding performance of *k*- onset can be attributed to its relatively distinct F_2 onset, as exemplified in Figure 3.2.

Next, we conduct a further analysis by breaking down the identification errors into different types, namely, *total assimilation* and *partial assimilation*. Figure 3.5a shows the effect of speech rate on proportion of the two types of regressive assimilation. Total regressive assimilation here refers to the case when the coda consonant is erroneously perceived to be *identical* in POA to the following onset consonant, e.g. *laat* is perceived as *laap* when followed by *p*- as in the token *laat-paa*;

Partial regressive assimilation refers to the case when the coda consonant is perceived to be *more* like the following onset consonant in POA. There are only two cases for this category: (1) *-p* is perceived as *-t* when followed by *k-* (in the token *laap-kaa*), and (2) *-k* is perceived as *-t* when followed by *p-* onset (in the token *laak-paa*). Both cases involve the shifting of the perceived POA towards that of the following onset consonant.



(a) regressive assimilation (b) progressive assimilation

Figure 3.5 shows the rate of total / partial regressive assimilation

The proportion of total assimilation in Figure 3.5a is shown to increase from 43.05% to 61.85% when speaking rate goes up from 2000ms to 500ms condition. This shows the conclusion from Experiment 1 that rapid speech enhances regressive assimilation is also valid in the perceptual domain. The instance of coda confusion classified as partial assimilation is much smaller in number than total assimilation. The figure decreases from 13 (8.61%) to 7 (4.05%) as speech rate increases. This apparent contradiction to the results of total assimilation, however, is readily explainable by considering also figures summing all the assimilation types: The aforementioned increasing trend (as speech rate goes up) still persists (from 51.66% to 65.90%). Some of the coda consonants are perceived more like the following onset (as reflected as partial assimilation), but as speech rate increases, regressive assimilative effect becomes more prominent such that some of the partial assimilative cases become the total assimilation instead, leaving a smaller number in the partial assimilation class.

Since only three choices are available to subjects for consonant identification, there is a 50% probability that a perceptual confusion is classified as total assimilation. The expected rate is thus approximately 50% when the answer was obtained by pure

random guessing. Taking this figure into consideration, regressive influence is in fact quite insignificant when the speaking rate is slow (conditions 2000ms and 1500ms), where the rate is 51.66% and 55.07%, merely slightly over the by-chance level.

Similarly, Figure 3.5b shows the picture for progressive assimilation. Interpretation of *total* and *partial* progressive assimilation here is the same, depending on whether the erroneous perceived onset is totally or partially identical to the following onset consonant. *laap-kaa* perceived as *laap-paa* and *laap-kaa* perceived as *laap-taa* are respectively examples of these two types of assimilation. Compared to regressive assimilation in Figure 3.5a, there is no clear pattern of speech rate-dependence. Besides, it must be noted that the number of onset identification errors is much smaller than that at the coda position, with a ratio of about 5.5 : 1, such that the percentages are less representative.

In the perception experiment, subjects, though presented with the reported formant trajectory variation, as well as other co-varying acoustic cues, cannot fully recover those stop consonants (especially codas). In another communication round, if the inaccurately perceived consonant is adopted by the current listeners, a minute sound change will have occurred. With iterations of such minute changes, a phonological change (regressive assimilation) may finally emerge.

Empirically, previous studies on secret languages hinted some indirect evidence of such rule in Hong Kong Cantonese. Chao (1931) reported an extensive survey of secret languages (i.e. *Fanqie-yu*) in various Chinese dialects, detailing the basic construction rules as well as additional language-specific idiosyncrasies. Chao suggested an important reference value in studying those secret language construction rules to studying the phonology of the host language. E.g. *Fanqie-yu* also follows phonological processes in their host language in *Beiping fanqie-yu*: When the *fanqie-yu* target are a sequence of two syllables of tone-3, the first tone-3 changes to tone-2, following exactly the tone-3 sandhi rule in Mandarin.

In Cantonese's case, Tse (1979) discussed *Mang-gong-wa* (MKW) (a type of *fanqie-yu*) in Cantonese. Interestingly, some of the MKW words undergo a regressive assimilation before its phonetic realization. The word *pin* 'side', for instance, is translated to *lim-pi*, instead of *lin-pi* according to the original rule. The alveolar coda *-n* is regressively assimilated to the bilabial coda *-m* before realization. Several parallel examples were given, all demonstrating the same phenomenon.

Combining these two studies, it can be induced that regressive assimilation observed in MKW may not be an idiosyncratic case, rather it reflects the speakers' internalized phonological rules (i.e. regressive assimilation) in its host language. In other words, regressive assimilation is possibly a fertile rule in Cantonese.

3.3 Summary of findings

We reported two phonetic experiments in this Chapter. The production experiment (Experiment 1) investigated the effect of speaking rate on acoustic properties of stop coda and onset consonants. It was found that as speaking rate increases, stop codas become more and more regressively assimilated to the following onset consonant, as reflected in the inflection of F_2.

Following that, our perception experiment (Experiment 2), making use of the natural tokens from Experiment 1, demonstrated significant regressive assimilative effect when speech rate goes up. Listeners could not recover the original coda consonants even though there are many acoustic cues co-varying with the rate-induced phonetic variations in F_2.

Combining results from the two experiments, we have demonstrated under a laboratory setting an instance of sound change: the rate-induced phonetic variations from production propagate to the listeners, leading to eventually a categorical perceptual error. If the wrongly perceived coda consonant is used when the listener plays the role of speaker in another communication round, a small sound change will occur.

In the following chapter, we continue to find empirical evidence for such regressive assimilation change, by a series of analyses on spoken corpora of Hong Kong Cantonese.

CHAPTER 4

SEGMENTAL CHANGE - CORPUS STUDIES

As reviewed in Chapter 2, it is widely reported in the literature that, cross-linguistically, segmental feature (e.g. voicing, place of articulation) assimilation among successive consonants show a strong bias in its directionality, namely, a stronger trend to observe regressive (e.g. labialization into *-m* when *-n* appears before a labial onset consonant in Japanese) than progressive (e.g. voicing of English plural marker *-s* depends on voicing feature of the final consonant of the base) assimilations (Kent, 1936; Hill, 1955; Ohala, 1990, among others).

The reasons underlying such asymmetry have been explored in previous literature from various perspectives like physiology (Krakow, 1999), production (Tuller & Kelso, 1990), and perception (Fujimura *et al.*, 1978). In essence, these reports investigated internal organization of syllable and all concluded with a common theme that onset and coda consonants are handled differently along the speech production-perception pathway.

Experiment 1 and 2, under a laboratory setting, demonstrated the assimilative influence of neighboring consonants in continuous speech and that it is much more likely that coda consonants change in response to the following onset consonants than the opposite case (i.e. onset consonants being progressively assimilated by the preceding codas). In the experiments, speech rate was manipulated resulting in different degrees of place assimilations. As speech rate fluctuation is omnipresent in spontaneous speech, we find investigating spontaneous speech data a good testing ground for verifying empirically the experimental results.

We report in this Chapter a series of corpus analyses on a spoken corpus of Hong Kong Cantonese (HKC) to study variation pattern in consonant assimilation. A

corpus study covers a much larger pool of data under a much wider range of phonetic contexts than those elicited under tight experimental control. If, after inclusion of so many factors which are present in the spoken corpus, the aforementioned directional asymmetry still holds, we can then be more confident in believing that the directional asymmetry suggested by Experiment 1 and 2 indeed underlies the human speech production mechanism. Experiment 3 and 4 detailed below give such empirical evidence in spontaneous speech in Hong Kong Cantonese.

4.1 Synchronic study - Experiment 3[22]

Material

The corpus we used in this study was Hong Kong Cantonese Adult Language Corpus (HKCAC) (Leung & Law, 2001)[23]. It contains phonetically transcribed data of more than 8 hours of phone-in programs and forums on radio, from 69 (not counting program hosts) native Hong Kong Cantonese speakers. These radio programs spanned more than one year, from November 1998 to February 2000. Totally, 141,149 tokens were recorded in the database, much larger than an earlier similar attempt (Fok 1974, 1979).

The corpus was transcribed by two female university graduates from the Chinese and Bilingual Studies Department at The Hong Kong Polytechnic University (Law, personal communication). They did the phonetic transcription while listening to the tape recordings. Ten percent of the transcription was later on randomly selected by the project investigators for cross-checking. In case of disagreement, the investigator and the transcriber listened together to the audio playback and resolved the difference.

The resulting corpus contains phonetic and orthographic transcriptions of the recordings. The transcription data are stored in lines, where each line roughly corresponds to an utterance. Each line is further divided into syllables, for which IPA

---

[22] Results of this experiment has been partially reported in Wong (2010).

[23] Phonetically transcribed data are available online at http://shs.hku.hk/corpus/index.htm.

symbols are provided denoting the segmental and tonal composition, with the corresponding Chinese characters, whenever possible, listed directly below. Besides, other information such as the radio program title, gender of the speaker, and the speaker's role (i.e. program host vs. phone-in audience) are also provided.

Since the corpus text was originally accessible in Microsoft Excel format, we chose Excel as the tool for analyzing them. Specifically, Visual BASIC for Applications (VBA) programs were coded for data collection and analyses.

Coda confusion

We focus on coda variations in HKC spontaneous speech in Experiment 3-I. Here we limit our search to the six stop codas[24] in HKC, namely *-p*, *-t*, *-k*, *-m*, *-n*, *-ŋ*. We extract from the database all the characters having one or more variants with codas within the above set. The character 發 'to expand', canonically pronounced as /fat$^{\underline{33}}$/, is found to have instances in the corpus to be produced without a coda (i.e. [fa^{33}]). Conversely, the character 做 'to do' has a lexical pronunciation of /tsou^{22}/ (no coda). However, it has tokens in the corpus transcribed with pronunciation ending with *-k* ([tsok$^{\underline{22}}$]) and *-ŋ* ([tsoŋ^{22}]). Under our current selection criteria, these two characters are included in our analysis.

There are altogether 1,257 characters having tokens with coda consonants. Among them, 220 characters have coda variants, ranging from two (e.g. 各 'each' appearing as [kɔk$^{\underline{33}}$] and [kɔt$^{\underline{33}}$]) to six (e.g. 我 'I' having tokens with all the six codas) variants. Table 4.1 lists the type and token frequency of characters with two coda variants. Each cell shows the type and token frequency (enclosed in brackets) of characters having coda variants with the codas indicated by the row and column indices.

---

[24] These six consonants are lexical codas, while the glottal stop coda [-ʔ], though with negligible token frequency, were also recorded in the HKCAC corpus.

	-p	-t	-k	-m	-n	-ng
-p						
-t	3 (149)					
-k	7 (189)	45 (10141)				
-m	1 (23)	2 (3038)	1 (404)			
-n		8 (1990)	1 (528)	14 (742)		
-ng		1 (8)	8 (1064)	1 (245)	83 (7752)	

Table 4.1: Type and token frequency of characters having two coda variants.

The cell, for instance, located at the intersection of the row marked "-t" and the column "-k" indicates that there are 45 characters (with a total of 10,141 occurrences in the corpus) having two coda variants, (i.e. *-t* and *-k*). The coda variations in Table 4.1 can be classified into three types. The first is variation with respect to place of articulation (POA), e.g. variant pair (*-k*, *-p*) and (*-m*, *-ŋ*). The character, for instance, 能 'can do' has tokens transcribed as [lam^{21}] and [laŋ^{21}] in the corpus. The second type is variation across the nasality boundary. There are three such variant pairs, namely (*-m*, *-p*), (*-n*, *-t*), and (*-ŋ*, *-k*), with the minimal contrast that the first one is nasal, while the second oral. The last type includes all the remaining variations, like (*-m*, *-k*), which awaits further investigation for the underlying patterns.

The primary observation is that are coda variations are more frequently seen in terms of POA than nasality, as observed in the far much larger figures in the corresponding cells in Table 4.1 (total frequency of 153(19,218) vs. 17(3,077)). Next, among such variations in POA, oral stops (*-p*, *-t*, *-k*) are more prone to confusion than nasal stops (*-m*, *-n*, *-ŋ*)[25]. This result gives empirical support to the perceptual

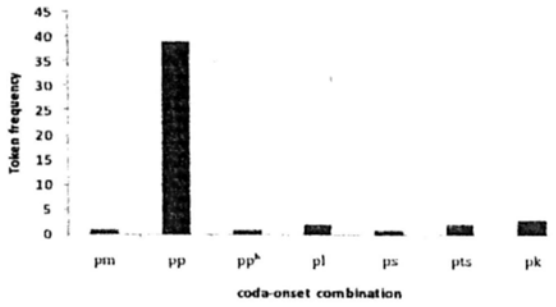
[25] Total type(token) frequency of the confusion among oral stops (*-p*, *-t*, *-k*) and nasal stops (*-m*, *-n*, *-ŋ*) are 55(10,479) and 98(8,739) respectively. Apparently, the frequency data show contradictory results, with oral stops / nasal stops being the more frequent ones if token frequency / type frequency is compared. However, the picture becomes clearer if we also take the total frequency of these stops into consideration. From the corpus report (Leung *et al.*, 2004, p503), nasal stop tokens (32,692) are approximately 2.5 times that of oral stop tokens (12,793). In other words, nearly 81.9% of the oral stop tokens in the corpus are not uniquely pronounced. With the same line of argument, oral stop coda

experiment results from Chen & Wang (1975). This bias is explainable in terms of their physical properties, summarized as "... the only perceptual cue for (oral) stops is the formant transition. Nasals; on the other hand, have spectral properties similar to those of vowels, and can be identified by means of their formant distribution (like vowels) as well as by virtue of the formant transitions (like plosives)." (Chen & Wang, 1975, p270).

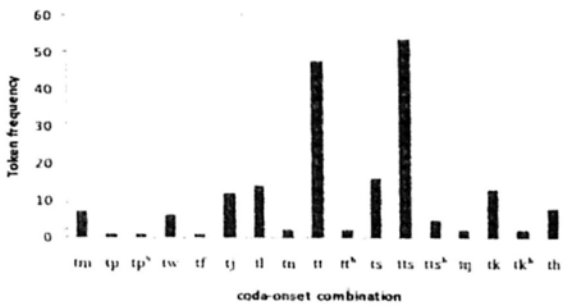
After obtaining this basic pattern of coda variation, we proceed to study the underlying factors behind it. From Experiment 1 and 2, it was found that onset consonants have considerable regressive influence on their preceding codas. We thus follow this result to investigate from the spoken corpus the effect of a following consonant on coda variation we obtained so far.

To conduct the investigation, we extract all the character tokens with stop-coda deviants from the lexical pronunciation, with the following segment. An example here may illustrate the idea, the character 特, with lexical pronunciation of /tak$^{\underline{22}}$/, is observed to have deviant transcriptions of [tap$^{\underline{22}}$] (in 特别 'special') and [tat$^{\underline{22}}$] (in 特登 'deliberate'), differing in terms of the coda consonant (*-p* and *-t*). These two tokens are thus included in the current investigation. With such selection criteria, a total of 1,368 tokens are collected for data analysis. Figures 4.1a-f plot the token frequency (*y*-axis) for *-p*, *-t*, *-k*, *-m*, *-n*, *-ŋ* respectively, with the coda-onset consonant sequence combination shown on the *x*-axis. For example, there are 39 tokens of coda consonant *-p* followed by another *p-* (as onset of the following syllable). They are labeled as *-pp-* in Figure 4.1a. (the aforementioned token 特别[tap$^{\underline{22}}$-pit$^{\underline{22}}$] is also counted here) and 9 occurrences of an un-aspirated voiceless velar stop coda followed its aspirated counterpart (*-kk^{h}-* in Figure 4.1c).

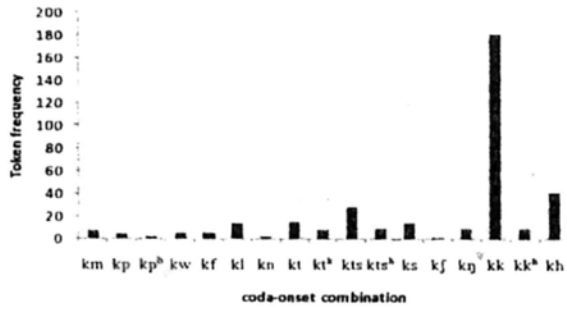
(frequency of 12(7,503)) is again observed to be more prone to confusion than nasal stop (frequency of 13(7,084)) in token pool with 3-coda variants.



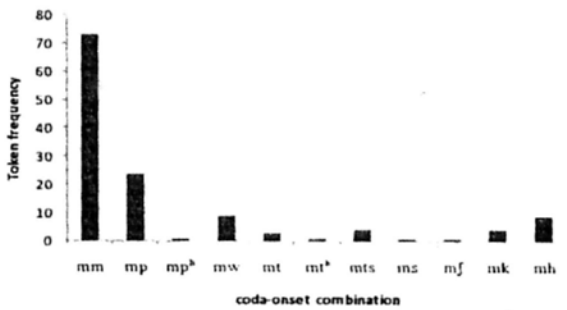
(a) onset consonants following *-p*



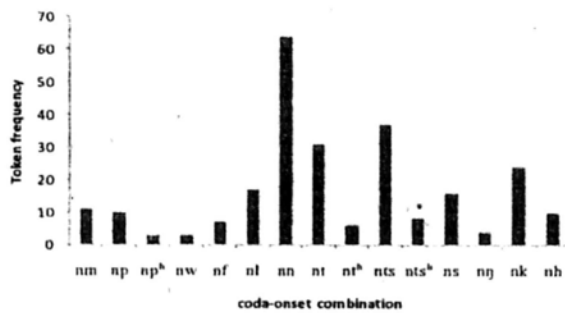
(b) onset consonants following *-t*



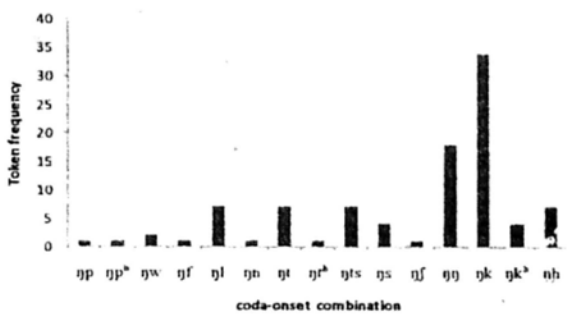
(c) onset consonants following *-k*



(d) onset consonants following *-m*



(e) onset consonants following *-n*



(f) onset consonants following *-ŋ*

Figure 4.1: Token frequency of different coda-onset consonant sequence combinations: a-c (oral stops *-p*, *-t*, *-k*); d-f (nasal codas *-m*, *-n*, *-ŋ*).

The first emerging pattern here is the regressive effect of POA. The most frequent coda-onset combinations from those six figures are respectively *-pp-*, *-tts-*, *-kk-*, *-mm-*, *-nn-*, *-ŋk-*. The coda consonants from these deviant tokens all agree in terms of POA with its following onset consonant. Four of them (*-pp-*, *-kk-*, *-mm-*, and *-nn-*) represent total assimilation in which the coda becomes identical to the following onset consonant. This regressive effect of POA even works across the nasality boundary, where codas are regressively influenced by a following /k-/ onset to

become /-ŋ/ coda. Another observation is related to the effect of aspiration: We observe that un-aspirated oral stop onsets are less powerful in assimilating its preceding coda than their aspirated counterparts. Comparisons between *-pp-* vs. *-pp^h-*, *-tts-* vs. *- tts^h-*, *-ŋk-* vs. *-ŋk^h-* well illustrate this point[26].

In summary, the first part of Experiment 3 reveals that regressive assimilative effect plays an important role in coda variations in spoken HKC. Now a similar test is carried out for onset variations in spoken HKC.

Onset confusion

Experiment 3-II focuses on phonetic variations of onset consonants in spontaneous Hong Kong Cantonese. For comparison purposes, we limit our target to characters starting with one of the six onsets used in Experiment 3-I (i.e. *p-*, *t-*, *k-*, *m-*, *n-*, *ŋ-*). All the characters, whatever their lexical onset is, are included in this analysis if they have records in the corpus with one of these six onsets. For instance, while the character 學 'to learn' is lexically specified as having a glottal fricative onset (/hɔk^{22}/), it is included in our data pool due to the fact that records of its being produced as [ŋɔk^{22}] are captured in the database. Eventually, 846 characters fall into our data pool, among which 107 of them have more than one onset variants.

Table 4.2 shows the type and token frequency of the characters (totally 72 of them) with two onset variants. A similar notation as Table 4.1 is used here. We may find a similar 3-way classification of onset variation here: (1) variants differ in terms of POA (e.g. (*t-*, *k-*), (*m-*, *n-*)); (2) variants differ minimally in terms of nasality (e.g. (*m-*, *p-*); and (3) others (e.g. (*k-*, *m-*)).

---

[26] To facilitate fair comparison, the effect of token frequency must be considered here. From Leung *et al.* (2004), aspirated onset consonants are consistently less frequent than their unaspirated counterparts (e.g. [k^h] (3,044) vs. [k] (19,319) and [ts^h] (3,274) vs. [ts] (11,564)) such that larger number of instances of "coda-unaspirated consonant" sequences, like *-pp-* and *-tts-*, is expected. However, even after accounting for this distributional bias, the assimilative effect for the unaspirated onsets is still seen to be obviously larger than the aspirated ones.

	p-	t-	k-	m-	n-	ng-
p-						
t-	1 (79)					
k-	3 (899)	7 (1425)				
m-	10 (1585)		3 (1844)			
n-	1 (97)	9 (3931)	2 (335)	3 (3438)		
ng-		1 (79)	17 (3490)	10 (2380)	5 (1086)	

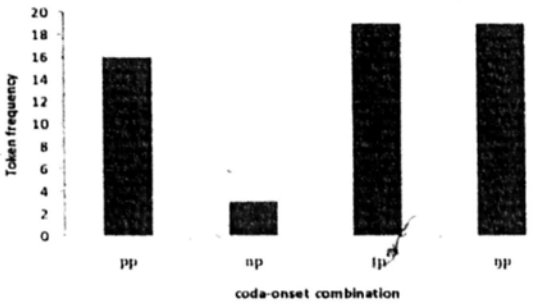
Table 4.2: Type and token frequency of characters having two onset variants.

Upon closer examination, the onset variation pattern is qualitatively different from that of codas. First, there is no longer a strong bias towards variations in terms of POA as in coda variation. In fact, the figures are approximately the same (POA: 29(9,307) vs. nasality 36 (9,006)). Second, with regard to POA onset variation, nasal onsets are more commonly attested. The number of tokens involved in POA onset variations for oral and nasal stop onsets are 2,403 and 6,904 respectively. Dividing these numbers by the token frequency of 33,699 (with *p*-, *t*-, *k*- onsets) and 11,470 (with *m*-, *n*-, *ŋ*- onsets) (Leung *et al.*, 2004, p504), only 7.13% of the oral stop tokens, but 60.19% of the nasal stop tokens are involved in POA variation.

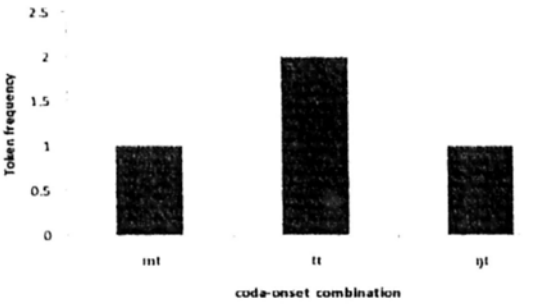
We then proceed to investigate the effect of neighboring consonants' (preceding segment in our case) influence in such onset variation. To serve that purpose, from our data pool, all the character tokens with onset consonant deviant from the lexical pronunciation are extracted, together with its preceding segment. The character 頭 'head' is lexically pronounced as /t^{h}au^{21}/, but has instances in the corpus transcribed as [kau^{21}] and [tau^{21}]. These two tokens [kau^{21}] and [tau^{21}] are included in the current analysis. The resulting token frequency (*y*-axis) are plotted in Figure 4.2a-f for *p*-, *t*-, *k*-, *m*-, *n*-, *ŋ*- respectively, with the coda-onset consonant sequence combination indicated on the *x*-axis. Note the significantly larger token frequency for nasal onsets.

We first take a look at the oral stops (Figure 4.2a-c). Generally speaking, a similar obvious trend as in coda variation is not present in the figures: The leading combinations for *p*- onsets (Figure 4.2a) are *-ŋp-* and *-tp-*, both are from different POA than the bilabial stop *p*-; *-tt-* is leading in Figure 4.2b, but the difference is only 2 to 1 occurrence; All the candidates in Figure 4.2c agree in POA with the velar stop *k*-.

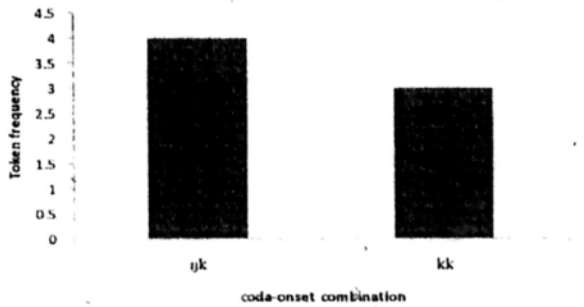
But the total token frequency is only 7. Essentially, for the oral stop onsets, due to sparsity of data, any conclusive result can unlikely be drawn.



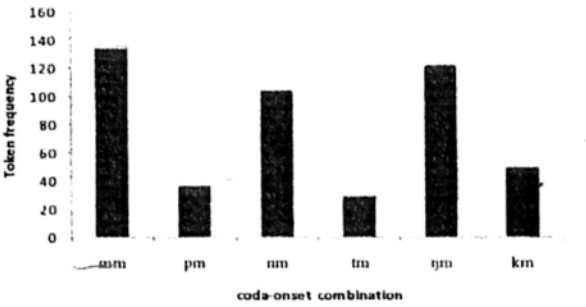
(a) coda consonants preceding *p*-



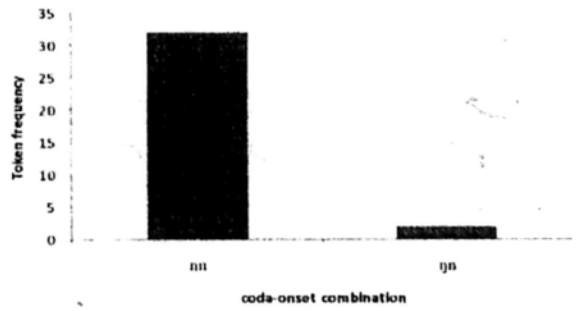
(b) coda consonants preceding *t*-



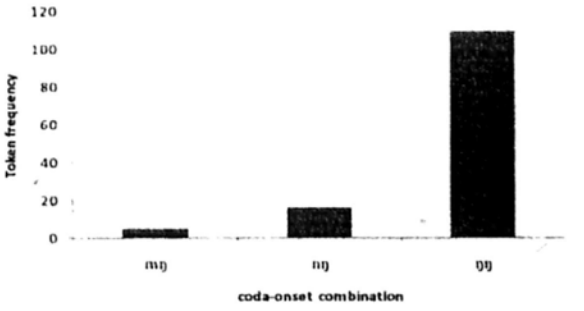
(c) coda consonants preceding *k*-



(d) coda consonants preceding *m*-



(e) coda consonants preceding *n*-



(f) coda consonants preceding *ŋ*-

Figure 4.2: Token frequency of different coda-onset consonant sequence combinations: a-c (oral stops *p*-, *t*-, *k*-); d-f (nasal onsets *m*-, *n*-, *ŋ*-).

Next, we move on to study the case for deviant characters with nasal onsets. Effect of the preceding coda consonant becomes more evident as we find the most frequent coda-onset combinations in the figures (*-mm-* in Figure 4.2d, *-nn-* in Figure 4.2e, *-ŋŋ-* in Figure 4.2f) all agree in their POA specifications.

To sum up, Experiment 3-II shows that onset consonants are not dominantly influenced by its preceding coda consonants in spontaneous speech. Some progressive assimilative influence can be observed for nasal onsets though. However, as the trend is much less obvious than that for coda variations, there may be other major factors (not considered in this study) conditioning such changes.

Discussion

Comparing distribution pattern of consonant variations as well as magnitude of neighboring effect obtained from Experiment 3, the differences between onset and coda consonants is clearly shown. Essentially, coda consonant is more malleable than onset consonant under the influence of its neighboring segment. Since synchronic variation can be seen as a snapshot of the longer-term diachronic changes, our findings on synchronic variations in HKC is compatible with previous scholars' observation on diachronic segmental feature assimilation.

Since no audio data is available, the oral communication stage at which the variation occurs may be questioned. In other words, for a particular variation extracted from the corpus, whether there was accurate pronunciation but inaccurate transcription, or inaccurate pronunciation followed by accurate transcription, or even both, are responsible, is queried. Nevertheless, there has been a stage along the communication chain in which confusion (i.e. an instance of sound change) occurs. From our previous experiments, we are inclined to believe that such variation comes from production factors. If, however, the pattern is later on found to have roots in perception, a whole new horizon of perception research will be opened.

Zee (1985) reported a study of nasal coda consonants in modern Chinese dialects. The study covered 19 Chinese dialects, affiliated with all the major Chinese dialectal families. In some dialects, part of the inventory (esp. bilabial nasal *-m*) has merged with other coda categories (e.g. *-m* > *-n*, e.g. Beijing, Nanchang, Suzhou). In some other cases, the codas disappeared, leaving only as a trace nasality in its preceding vocalic element (e.g. Jinan and Xi'an). In the same study, Guangzhou is among the best in preserving its nasal codas, apparently indicating a language-internal stability in the syllable-finals. Yet, our synchronic corpus study in HKC (which is a close variant to Guangzhou dialect) still shows phonetic variations in onset/coda consonants. Through iterations of communication, such accumulative phonetic

variations will eventually spread through sociolinguistic processes to the whole linguistic community, eventually leading to categorical change of sounds (Ohala, 1981).

4.2 Diachronic study - Experiment 4

This Section reports a comparative study between two Hong Kong Cantonese (HKC) speech corpora. The main aim of this comparison is to obtain quantitative diachronic drifts, if any, in frequency distribution of various speech sound units in HKC. In addition, we will give some explanation to the observed pattern through modeling efforts. Such modeling may eventually contribute to projection of the future development in of the sound system in HKC.

We make use of two spoken corpora of HKC, namely Fok (1979) and Leung *et al.* (2004) (i.e. the one used in the synchronic corpus study Experiment 3 reported above). Fok (1979) covers a total of one hour of speech data, consisting of 15 minutes of radio program recording and 45 minutes of conversation (6 participants). Totally, over 13,000 tokens are present in the corpus. Leung *et al.* (2004), being a later attempt, is considerably larger than Fok (1979). It contains more than 8 hours of speech data recorded from radio broadcasting with 69 native speakers (excluding program hosts). These radio programs spanned more than one year, from November 1998 to February 2000. Altogether, 141,149 tokens are included in the database. The two corpora, separated in their production by 25 years, should be able to show some middle-term sound changes.

General results

Figure 4.3 plots the frequency of occurrence of the six coda consonants in HKC, sorted in POA and grouped according to nasality of the consonants. Nasal codas obviously outnumber oral ones. The result is compatible with the observation in Chen & Wang (1975) that oral obstruent codas generally experience faster attrition than nasal ones. A general trend can be observed with respect to place of articulation that bilabial codas ([-m] and [-p]) are the least frequent, followed by alveolar ([-n] and [-t]), and velar ([-ŋ] and [-k]) the most frequent.

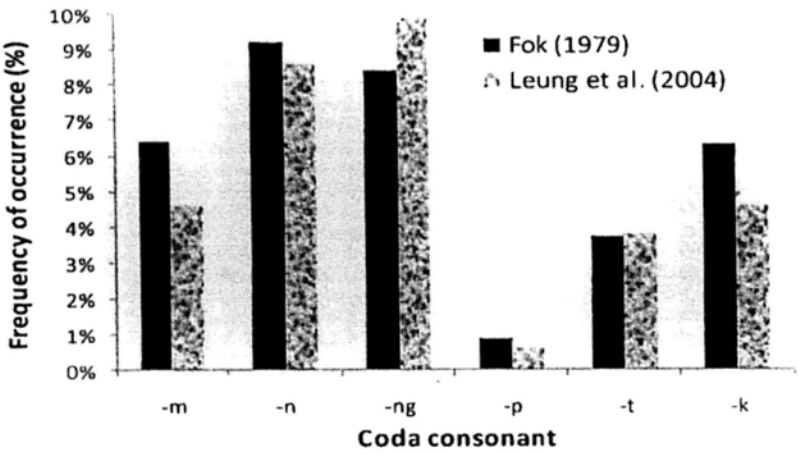


Figure 4.3: Frequency of occurrence (in percentage) of different coda consonants in the two corpora.

We proceed to Figure 4.4, in which frequency of occurrence of onset consonants are shown[27]. Obviously, *un-aspirated* consonants (*b-* [p], *d-* [t], *g-* [k]) are significantly more frequent than their *aspirated* counterparts (*p-* [p^h-], *t-* [t^h-], *k-* [k^h-]). Several obvious diachronic changes can be observed in the figure. First is the great decrease in number of *n-* tokens and great increase in number of *l-* tokens. This reflects the widely reported *n-/l-* merger (Zee, 1999). Second, virtually there are only very small amount of tokens for initials *gw-* [k^w-] and *kw-* [k^{wh}-] (0.41% and 0.06% respectively) left in Leung *et al.* (2004). This is due to de-labialization of those rounded initials to become *g-* [k-] and *k-* [k^h-] correspondingly (e.g. *gwok6* 'country' becomes *gok6*; and *kwong4* 'crazy' becomes *kong4*). Lastly, though not too obvious, the slight decrease in *ng-* initial may come from the well-documented *ng-* >> zero-initial process in Hong Kong Cantonese (e.g. *ngo5* 'I' pronounced as *o5* and *ngai4* 'dangerous' pronounced as *ai4*).

---

[27] The graph does not show exhaustively all the consonants in the two papers, since consonants in the two corpus studies were not transcribed with the same phonological system. For instance, [tʃ] appears in Leung *et al.* (2004) only while the syllabic [ŋ] appears in Fok (1979) only.

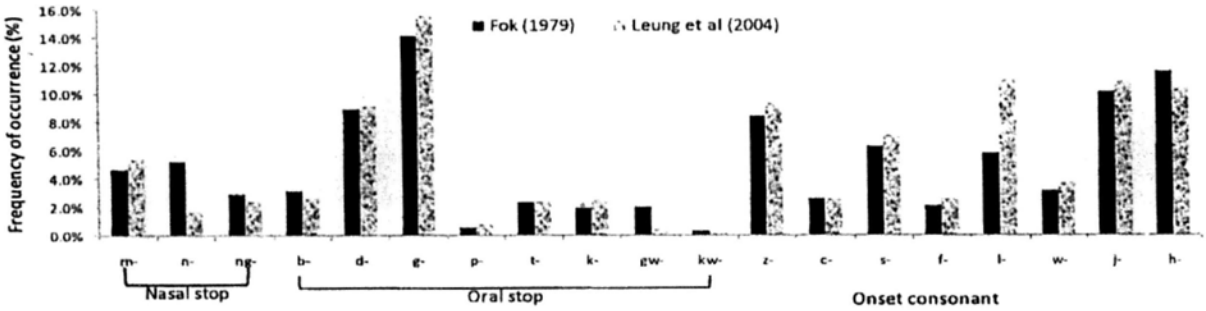


Figure 4.4: Frequency of occurrence (in percentage) of different coda consonants in the two corpora.

Modeling diachronic consonantal changes

The above comparison demonstrates only one of the many ways we can observe diachronic changes from comparative studies of corpora. We now explore another way to probe changes which gives the picture of a larger time-frame.

From Production Experiment 1, it was found that whenever speech rate goes up, coda consonants get assimilated into the immediately following onset consonant. For examples, *lap-ka* becomes *lak-ka*, *lak-pa* becomes more like *lap-pa*. Given the omnipresence of speech rate fluctuation, there is a constant force for the coda consonants to become more like the following onsets. We now try to model the process and thus investigate the long term behavior of different consonants.

Let the number of syllables (uttered within a finite time-frame) with coda consonants *-p*, *-t*, *-k* at time t be $C_p(t)$, $C_t(t)$, $C_k(t)$ respectively. Further, let P_p, P_t, P_k be the probability of a consonant following the current syllable being [p], [t], and [k] respectively, which is primarily determined by the relative frequency of syllables with initials *p-*, *t-* and *k-*. From Production Experiment 2, we learn that some codas were wrongly perceived more easily, and thus undergoes regressive assimilation more easily, e.g. the coda consonant *-k* is more readily assimilated to the following onset consonant than *-p*, according to results from Experiment 2. We use D_{ij} to denote such factor of "likeliness of regressive assimilation" of a coda consonant i to get assimilated to the following onset j. D_{kt} (the likeliness of *-k* to get assimilated to *-t*) is probably larger than D_{pt} (the likeliness of *-p* to get assimilated to *-t*), considering the perception results in Experiment 2.

Consider the case of syllables closed with *-p*. At any time *t*, there is an *inflow* of syllables (through regressive assimilation) from syllables originally ending with *-t* and *-k*. There contribution can be expressed as

$$C_t(t)P_pD_{tp} + C_k(t)P_pD_{kp} \quad \cdots\cdots\cdots (4.1)$$

At the same time, there is also an *outflow* of syllables (through regressive assimilation) into syllables ending with *-t* and *-k*, expressed as

$$C_p(t)P_tD_{pt} + C_p(t)P_kD_{pk} \quad \cdots\cdots\cdots (4.2)$$

Lastly, as reported in Chen & Wang (1975), cross-linguistically, obstruent codas have the trend to disappear along the history, leaving only open syllables, in the form of CV syllables. We arbitrary name it as *decay factor*, denoted as DF_p. The corresponding *outflow* of syllables is thus

$$C_p(t) * DF_p \quad \cdots\cdots\cdots (4.3)$$

Summing up these three terms, a differential equation representing the rate of change of number of syllables with *-p* coda is thus

$$\frac{dC_p(t)}{dt} = C_t(t)P_pD_{tp} + C_k(t)P_pD_{kp} - C_p(t)P_tD_{pt} - C_p(t)P_kD_{pk} - C_p(t) * DF_p \quad \cdots\cdots\cdots (4.4)$$

Similarly, the corresponding differential equations for $C_t(t)$ and $C_k(t)$ are

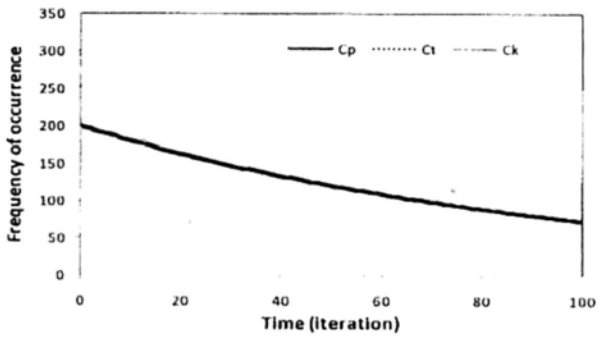
$$\frac{dC_t(t)}{dt} = C_p(t)P_tD_{pt} + C_k(t)P_tD_{kt} - C_t(t)P_pD_{tp} - C_t(t)P_kD_{tk} - C_t(t) * DF_t \quad \cdots\cdots\cdots (4.5)$$

$$\frac{dC_k(t)}{dt} = C_p(t)P_k D_{pk} + C_t(t)P_k D_{tk} - C_k(t)P_p D_{kp} - C_k(t)P_t D_{kt} - C_k(t) * DF_k$$

..... (4.6)

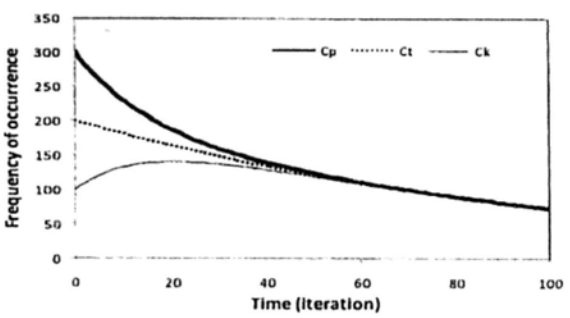
Equations (4.4)-(4.6) depict the interactive dynamics of syllables with different codas. Since there is no obvious analytical solution, we resort to numerical simulation to investigate their behavior. To be specific, equations have been set up in Microsoft Excel to capture diachronic evolution of the values $C_p(t)$, $C_t(t)$, $C_k(t)$.

We now study behavior of the system by applying different parameter values. The first candidate is initial number of syllables with different codas, i.e. $C_p(0)$, $C_t(0)$, $C_k(0)$. Figure 4.5 shows the evolutionary trajectories of those syllables given different initial population, with other parameters kept constant[28]. For each figure, the simulation is run for 100 iterations.

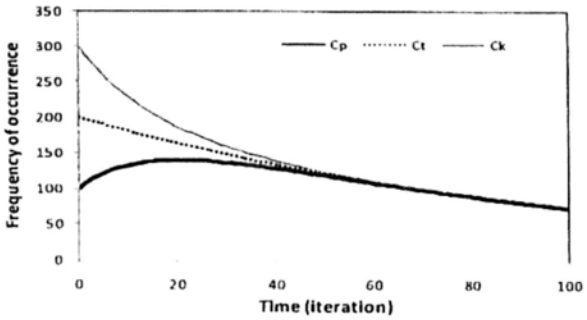


(a)

[28] Unless otherwise stated, default values of the parameters are $(C_p(0), C_t(0), C_k(0)) = (300, 200, 100)$, $(D_{tp}, D_{kp}, D_{pt}, D_{kt}, D_{pk}, D_{tk}) = (0.1, 0.1, 0.1, 0.1, 0.1, 0.1)$, $(P_p, P_t, P_k) = (0.2, 0.2, 0.2)$, $(DF_p, DF_t, DF_k) = (0.01, 0.01, 0.01)$.



(b)

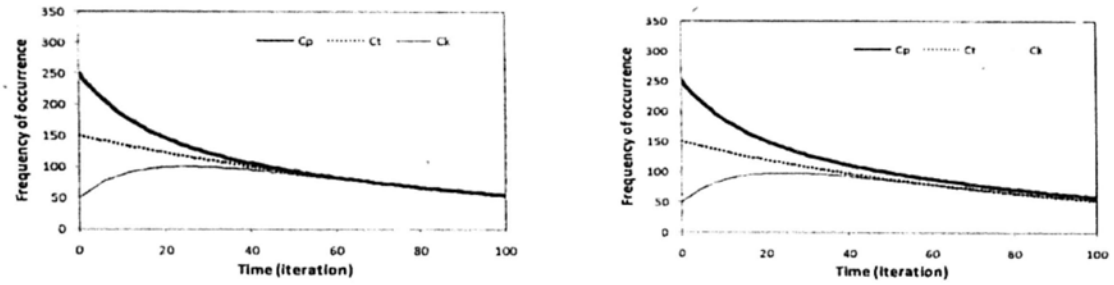


(c)

Figure 4.5: Effect of initial number of syllables for (C_p, C_t, C_k). (a) (C_p, C_t, C_k) = (200, 200, 200); (b) (C_p, C_t, C_k) = (300, 200, 100); (c) (C_p, C_t, C_k) = (100, 200, 300).

All syllables start with the same initial condition in Figure 4.5a and they develop with the same trajectory throughout the whole duration of simulation. A general trend of decrease in number can be observed, due to the decay factor *DF*. Initial population C_p is set to triple that of C_k in Figure 4.5b, as can be seen at $t = 0$. C_p then decays the most rapidly right from onset of simulation, while C_k starts with a slight rise, followed by a gradual fall afterwards. Eventually, all the three syllable sets converge again. Parameters for C_p and C_k are exchanged, and a reversed result is obtained as expected in Figure 4.5c. From these figures, we learn that initial number of syllables does have impact at first, yet the effect diminishes quite rapidly as time goes by.

Next, we investigate the effect of "likeliness of regressive assimilation" by altering the variables (D_{tp}, D_{kp}, D_{pt}, D_{kt}, D_{pk}, D_{tk}). With (C_p, C_t, C_k) set to (250, 150, 50), all the *likeliness* variables are set to 0.1 in Figure 4.6a. All the three syllable families, starting from different numbers, eventually converge from about half of the iterations. The likeliness of coda consonants to change to *-p* is set to slightly higher than that of other codas (i.e. $D_{tp} = D_{kp} = 0.11$, while others set to 0.1, same as before). The obvious effect is that the number of *-p* syllables decays more slowly than before, because of the greater inflow (the first two terms in Equation 4) of syllables in each round of simulation. The *-p* syllables converge again with the other two syllable families, but only when it approaches the end of simulation. These set of results indicate that the *likeliness* variables have a more prolonged effect on diachronic sound change, but it is still not the major factor behind syllable-coda evolution.

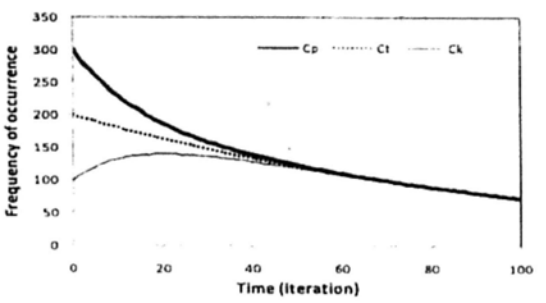


(a) (b)

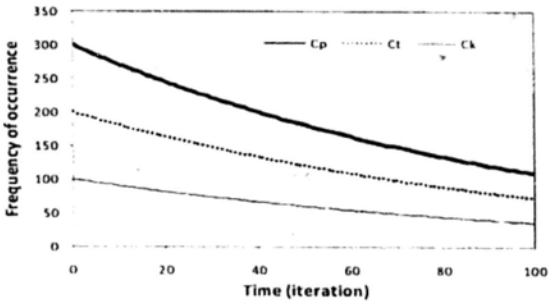
Figure 4.6: Effect of relative ease of regressive assimilation (D_{tp}, D_{kp}, D_{pt}, D_{kt}, D_{pk}, D_{tk}) (a) all equal to 0.1; (b) $D_{tp} = 0.12$, $D_{kp} = 0.11$, others remain the same.

Finally, we study the effect of frequency distribution of different onset consonants by tuning the parameters P_p, P_t and P_k. They are set identically to 0.2 in Figure 4.7a, resulting in a typical convergent pattern as described in previous paragraphs. In Figure 4.7b, the parameters are set to (P_p, P_t, P_k) = (0.3, 0.2, 0.1), emulating a linguistic environment in which the ratio of syllables with *p*-, *t*-, *k*- initials of 3:2:1. Consequently, all the three syllable families decay much more slowly than before, and their population is in the ratio of 3:2:1[29] at the end of simulation. Next, we reverse the parameter values to (P_p, P_t, P_k) = (0.1, 0.2, 0.3) such that the frequency distribution of different onset consonants does not match that of initial number of syllables grouped according to coda consonant (i.e. initial C_p, C_t, C_k). In this case, syllables with *-p* coda is the most abundant initially (compared to syllables with *-t* and *-k* coda), but the number of syllables with *p-* onset is the smallest among others (i.e. those starting with *t-* and *k-*). Interestingly, the initial population does not affect the end-of-simulation results. Syllables with different codas again follow the ratio of C_p : C_t : C_k = 1:2:3 as specified by the ratio P_p : P_t : P_k. Simply put, the relative frequency of different onsets dominates initial population.

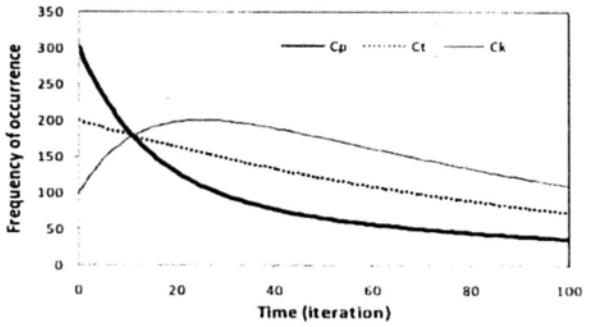
[29] To be precise, the final C_p, C_t, C_k are 109.81, 73.21, and 36.60.



(a)



(b)



(c)

Figure 4.7: Effect of frequency distribution of syllables with different onsets (P_p, P_t, P_k). (a) all equal to 0.2 ; (b) (P_p, P_t, P_k) = (0.3, 0.2, 0.1); (c) (P_p, P_t, P_k) = (0.1, 0.2, 0.3).

Empirical-modeling result match

From the above language modeling, the obvious observation is that in a long-enough timeframe, relative frequency of different codas is driven by onset token frequency. We now try to verify if it is true in the spoken corpus data.

First, we extract all the stops (oral and nasal) at the syllable-initial and syllable-final positions, and plot the frequency distribution, sorted according to POA, in Figure 4.8. We may find a general trend of bilabial < alveolar < velar for both onset and coda consonants.

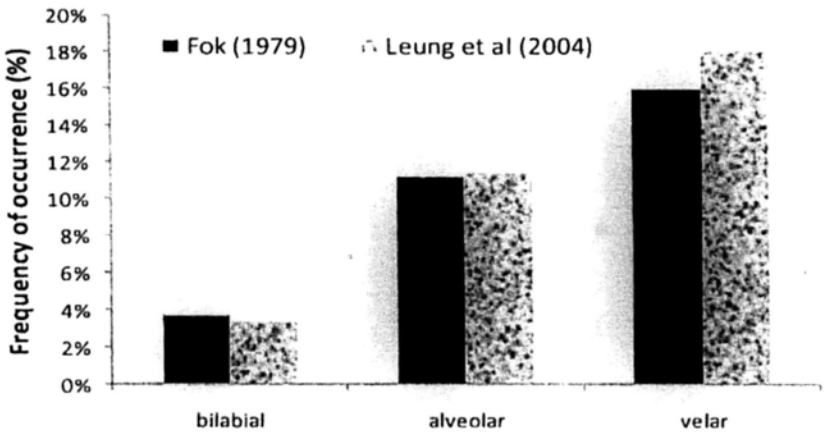


Figure 4.8: Frequency of occurrence (in percentage) of different onset consonants in the two corpora.

Discussion

The match between empirical data and modeling can be argued to result from assimilative effect between neighboring onset and coda consonants, in either direction. However, combining results from both our production experiment (Experiment 1) and synchronic corpus analysis (Experiment 3), it seems regressive assimilation is more likely the case.

Diachronic changes can be effectively revealed by directly comparing frequency measures of speech corpora constructed at different times in history like what we described above. Comparing speech corpora (as opposed to written corpora) is supposed to be the most effective way as sound changes always originate from minute drifts in spoken form of speech units, only propagating to the written form much later on. This approach generally requires corpora farther apart in history for processes that take a longer time-frame. What we could show at the moment was just an observation *compatible* with the hypothesized diachronic process. To rigorously verify whether the regressive assimilation process proposed above did occur in history of Chinese development, a comparable database of speech sound frequency distribution of classical Chinese is necessary. Historical texts from different historical stages are possible sources for compiling such databases. Yet, it must be noted that these texts record Chinese of old ages in written form (rather than spoken form), which reflect language usage with subtle differences from the spoken form.

Apparently, the three families of syllables compete for survival in the language in the simulation. However, a closer in to Equation 4.4-4.6 reveals that their populations change in an *interactive* way. Take Equation 4.4 (also applicable to Equations 4.5 and 4.6) as an example, the first two terms are *inflow*, while the remaining three are *outflow*. It tells that the evolutionary trajectory not only depends on the number of syllables with *-p* coda (i.e. $C_p(t)$), but also its competitors, those with *-t* and *-k* codas (i.e. $C_t(t)$ and $C_k(t)$). A sudden increase of C_p not only benefits survival of *-p* syllables, but also increases the *inflow* of syllables into *-t* and *-k* families (the first term in Equations 4.5 and 4.6). In other words, a change in one component may trigger a chain reaction within the whole system, just like the Great Vowel Shift (GVS) in England at about the end of the Middle English period (Trask, 1998, p85), where the up-shifting of low-mid vowels lead to further up-shifting of high-mid vowels to preserve enough vowel space for vowel discrimination [30]. Simply put, full understanding of the dynamics of an evolving language is approachable only if we take a system perspective.

Apart from the regressive assimilation effect under investigation, there are probably numerous other factors, for instance, ease of articulation, and perceptual distinctiveness, affecting the distribution of onset and coda consonants. The current corpus study concerns only Hong Kong Cantonese, leaving the possibility that the identified empirical-modeling match being merely a co-incidence. It thus certainly awaits parallel studies in other languages to verify cross-linguistic applicability of the current modeling[31].

---

[30] The phonological chain reaction that starts with a speech sound moving close to a second one, causing that second one to move further away, and do the same thing to a third one, is regarded as a *push chain* (Trask, 1998, p86-87). Another proposal to the GVS is a *drag chain*, where the whole reaction starts with a speech sound moving away from its initial position, leaving a hole in the acoustic space, dragging other ones to fill its position, creating yet another hole to drag other ones into it.

[31] Interested readers may consult Krámský (1959) for a snapshot of a cross-linguistic attempt to survey various groups of languages, drawing data from literary text.

4.3 Summary of findings

We conducted a series of corpus analyses to study the pattern of consonantal assimilations in Hong Kong Cantonese spontaneous speech. Both onset and coda consonants were found to be influenced by their neighboring segments, yet with different behavior: The effect of regressive segmental feature assimilation was found to be much more prominent than that in the progressive direction. This provides empirical support to our experimental results as well as the cross-linguistically observed directional asymmetry in segmental feature assimilation.

Next, we carried out a diachronic comparison of two spoken corpora of Hong Kong Cantonese to provide a piece of partial evidence of the historical trajectory of regressive assimilation. In the same study, we also learned that to thoroughly understand the evolutionary dynamics of a linguistic system, one must take a system perspective to take as many factors as possible into consideration.

CHAPTER 5

TONAL CHANGE – PHONETIC EXPERIMENTS

Characteristic patterns are observed for segmental features across speaking rates in Experiment 1. Speech samples thus obtained were then fed into Experiment 2, eliciting subjects' perceptual responses. The rate-dependent second-formant trajectory shifts observable in Experiment 1 induce corresponding perceptual shifts in Experiment 2. Specifically, segmental features spread *leftwards* (*regressively*). This matches the cross-linguistic dominant directions of segmental feature assimilation.

This chapter explores speech feature assimilation in the tonal domain by adopting a similar methodology as Experiment 1 and 2. Two experiments are devised to test speaking rate's effect on tone acoustics (Experiment 5) and tone perception (Experiment 6). To study tonal feature assimilation, F_0 movement will be considered as the subject of study as it has been well demonstrated as the primary acoustic correlate of linguistic tone (Abramson, 1972; Fok, 1974; Lin, 1988, among others).

Production Experiment 5 follows a similar design as Experiment 1 to investigate tonal distinctions. Participants were requested to produce the presented target sentences under different time constraints. Production data will be analyzed acoustically to establish possible correspondences between speaking rate and the resulting acoustic signals. Following that, perception experiment (Experiment 6) was conducted, with stimuli consisting of tokens recorded in Experiment 5, from different speakers, at different speaking rates, and different tonal feature combinations.

5.1 Tone Production - Experiment 5

Under different speaking rate conditions, we try to elicit rate-dependent variations of tonal features from the current experiment. In particular, F_0 (fundamental frequency) measure is taken to be the subject of investigation.

Material

Target syllable *lau* is used for the current task. It is associated with either one of the six lexical tones in Cantonese, surrounded by two tonal context syllables *maa* (tonally specified as either high as *maa1* 'mother' or low as *maa4* 'sesame'), which are in turn embedded sentence-medially in the carrier test sentence *ngo5 heoi3* $\underline{S_c}$ $\underline{S_t}$ $\underline{S_c}$ *maai5 sung3* 'I go to $\underline{S_c}$ $\underline{S_t}$ $\underline{S_c}$ to buy some food', where S_t is the target while S_c are the tonal context syllables. With such design, full course of F_0 movement contour can be extracted because both the two tonal context syllables and the target start with voiced onset; Both the lateral consonant [l] from the target S_t and the nasal murmur associated with the tonal context S_c provide characteristic acoustic landmarks, contributing to accurate and objective segmentation. Finally, all combinations of the trigram target do not form a word in Cantonese, minimizing lexical bias in production.

Recording procedure and data extraction

Five native Cantonese subjects (4M1F) participated in the production experiment. 10 repetitions were obtained for each condition. Pace-keeping clips similar to that used in Experiment 1 were used again to elicit rate-dependent variations, but at only three levels of speaking rates (500ms, 1000ms, 1500ms) for this experiment. Consequently, (6 target tones) × (2 tonal contexts) × (3 speaking rates) × (10 repetitions) = 360 utterances were produced by each speaker. Utterances recorded were fed into PRAAT for pitch extraction, then the pitch values were rectified by a trimming algorithm[32] to

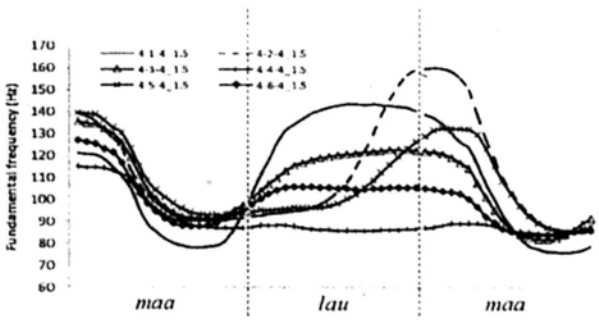
---

[32] The trimming algorithm follows the one used in Xu (1999) for removing spikes in raw F_0 contours extracted. Essentially, individual pitch values are compared with their neighbours. In case a sudden jump of value, defined by an adjustable threshold, is detected, the particular pitch point is replaced by an interpolated values calculated from its two neighbours. Pseudo codes are listed in the Appendix.

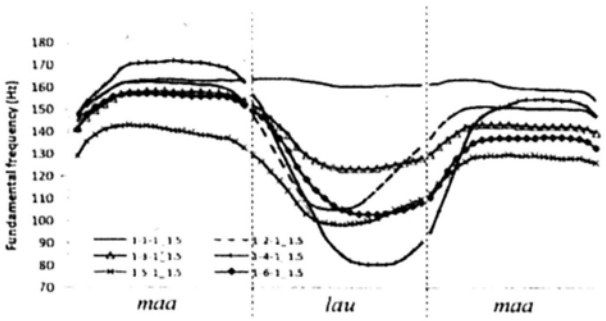
remove sharp spikes for data analysis. Apart from the above specificities, all other recording procedure and experimental settings followed that of Experiment 1.

Results and discussion

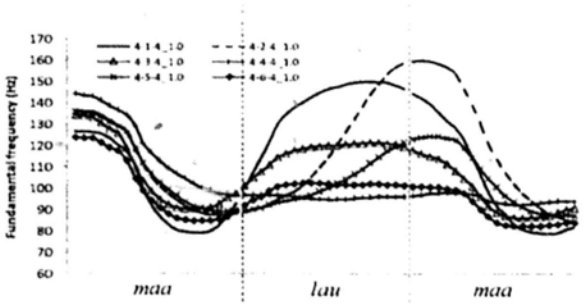
Time-normalized F_0 trajectories for the tri-gram S_c-S_t-S_c from Subject 1 are plotted in Figure 5.1, with durational measures in Figure 5.2. Durational measures show that individual syllables of the tri-gram are shortened under rapid speech. Each trace in Figure 5.1 came from averaging 10 samples. Generally speaking, F_0 starts at onset of the first context syllable S_c (the fist *maa*) at around 135Hz, moves to either 80Hz (for *low* tonal context *maa4*) or 150Hz (for *high* tonal context *maa1*) before it reaches onset of S_t (*lau*). During course of target syllable S_t, F_0 moves according to its associated pitch targets. After leaving S_t, F_0 starts approaching tonal target of the second context syllable. Simply put, F_0 movement here can be understood as consecutive approximating actions toward tonal targets associated with the corresponding syllable, according to the Target Approximation (TA) model (Xu & Wang, 2001).



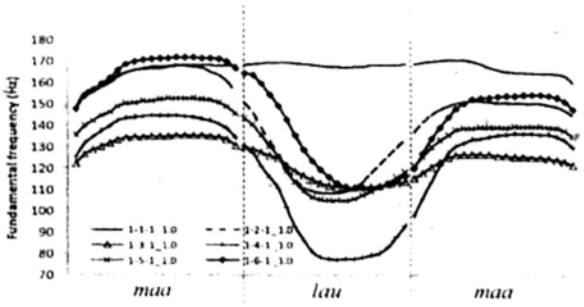
(a) LOW-Target-LOW, 1500ms



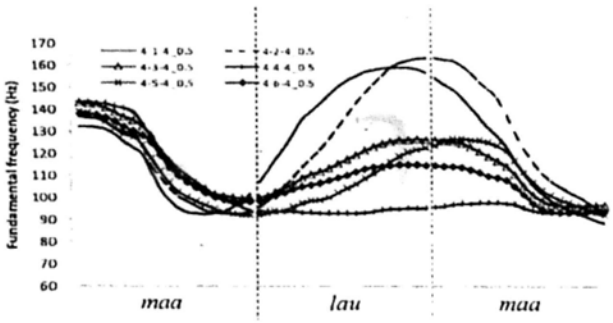
(d) HIGH-Target-HIGH, 1500ms



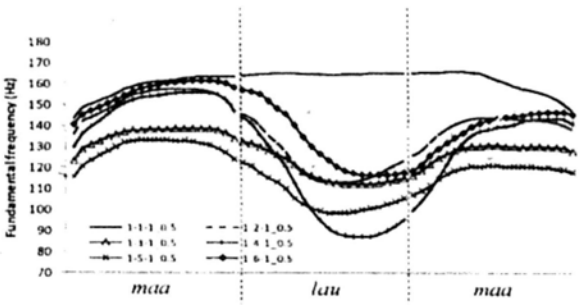
(b) LOW-Target-LOW, 1000ms



(e) HIGH-Target-HIGH, 1000ms



(c) LOW-Target-LOW, 500ms



(f) HIGH-Target-HIGH, 500ms

Figure 5.1: Time-normalized plot of F_0 trajectories across speaking rates, from *slow* (a, d), to *medium* (b, e), to *fast* (c, f). The left column (a-c) shows results of the target syllable surrounded by low tonal context, while the right column (d-f) high tonal context.

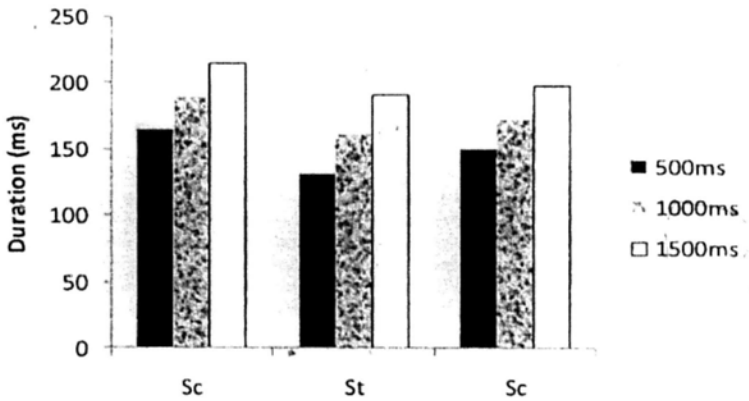
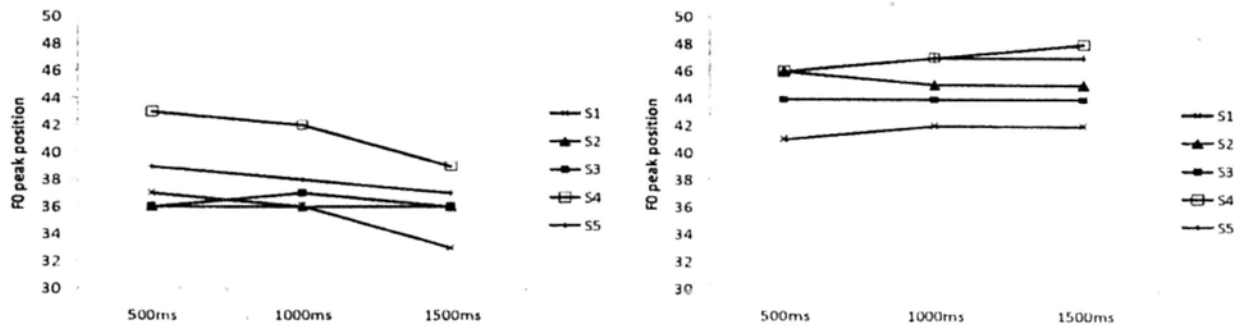


Figure 5.2: Durational measures of the two context (S_c) and one target (S_t) syllables for different speaking rate conditions.

Next we proceed to examine, if there is any, F_0 variation related to speaking rate. Comparing Figure 5.1a-c (*maa4-lau-maa4*), we find a gradual trend for F_0 contour for the condition *maa4-lau1-maa4* (the solid line) to shift rightwards as speaking rate increases, whereas the condition *maa4-lau2-maa4* (the dashed line) behaves more rate-independently. As a result, the two F_0 contours become more similar to each other under higher speaking rates. To obtain evidence quantitatively, *F_0 peak* is taken as the measure.

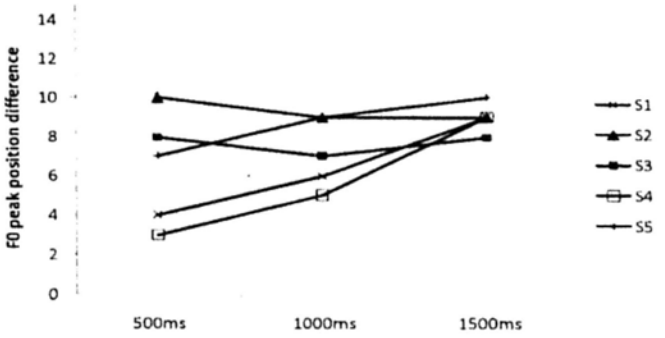
Figure 5.3 shows positional analyses of F_0 peaks for different subjects across speaking rates. The numbers on *y*-axis of Figure 5.3a and 5.3b denote the exact F_0

samples containing the *peak* F_0 value. Recall that 20 samples were obtained from each of the three syllables under investigation in Figure 5.1, numbers within the range (31-40) lie in S_t while those within the range (41-50) lie in the post-target context syllable. A general increasing trend of F_0 peak location is observable in Figure 5.3a for more rapid speaking rates. In other words, the higher the speaking rate, the later F_0 peak occurs. For the case of *maa4-lau2-maa4*, no such obvious trend is found in Figure 5.3b, matching our impressionistic judgment. Figure 5.3c shows the position difference among the two conditions, which is obtained by subtracting the position figures in Figure 5.3a from their counterparts in Figure 5.3b. A lower value indicates a smaller F_0 positional difference. A general trend[33] for such F_0 peak difference to diminish as speaking rate increases can be seen in Figure 5.3c.



(a) (b)

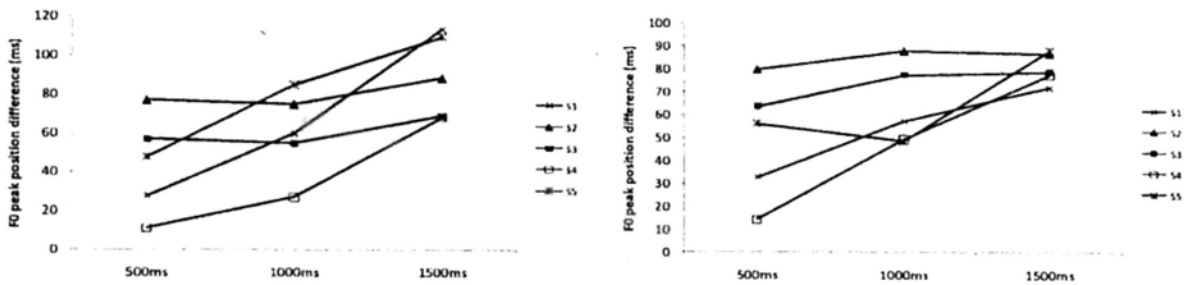
[33] Cross-subject variations are evident, as for S2 and S3 in this case. More discussions on the topic will be given in Chapter 7 of the dissertation. However, these two subjects agree with the prevalent trend in another position analysis (Figure 5.4b).



(c)

Figure 5.3: Positional analyses of F_0 peak location for different speaking rate conditions: (a) F_0 peak position for *maa4-lau1-maa4*; (b) F_0 peak position for *maa4-lau2-maa4*; (c) F_0 peak position difference.

Rather than the relative position, Figure 5.4 takes timing information into consideration to give the time difference (in millisecond) of occurrence of F_0 peak between the two conditions. Again, the timing difference decreases when speaking rate increases. As another case to demonstrate the rate-dependent F_0 spreading, the case between mid level (T_3) and rising tones (T_5) has been studied following the same procedure and the results are plotted in Figure 5.4b. The analogous phenomenon, where F_0 peak of the *maa4-lau5-maa4* shifts rightwards, is found under rapid speech. Such phenomenon of a rightward shift of F_0 peak, also named *F_0 peak delay*, has also been reported for Mandarin (Xu, 2001).



(a) (b)

Figure 5.4: F_0 peak difference between (a) high target tones *maa4-lau1-maa4* and *maa4-lau2-maa4*; (b) mid target tones *maa4-lau3-maa4* and *maa4-lau5-maa4*.

With reference to the condition *maa4-lau2-maa4* (the dashed line), which behaves more consistently across speaking rates, the contrast between

maa4-lau1-maa4 and *maa4-lau2-maa4* diminishes as speech rate increases, due to *lau1*'s closer and closer resemblance in F_0 movement to *lau2*. To restate the same phenomenon with more language-general terms, a low-high-low (LHL) tonal sequence becomes more like a low-rise-low (LRL) tonal sequence when speaking rate increases. Since the R in LRL is characteristic of a rise from low to high. Compared to the H in LHL, F_0 contours in R rises at a later time. Thus, a high speaking rate induces a delay in the F_0 rise. Consequently, the delay can be understood as a rightward (i.e. progressive) spreading of tonal features (L in this case) to the high tone from its immediately preceding low tone. The delay may be represented as the phonological process: LL-HH-LL >> LL-LH-LL[34].

This spreading phenomenon may be due to physiological limit of the larynx: As speaking rate increases, the larynx cannot catch up with the syllable rate in adjusting F_0 to accommodate consecutive high and low tonal targets. The observation that slightly larger F_0 range (from low (*maa4*) to high (*lau1*)) is recorded for slower speech (Figure 5.1a-c) gives support to such argument.

To sum up, Experiment 5 elicited rate-dependent variations in F_0 movement. Specifically, a higher speaking rate causes the delay of some F_0 trajectories, and thus the corresponding low tone feature. Formulated phonologically, *rightward* tone spreading can be observed.

5.2 Tone Perception - Experiment 6

Now it has been shown that as speech rate increases, there is a trend for tonal features to spread rightwards (progressive), a natural question follows: Do the observed F_0 peak delay lead to corresponding shifts in perception?

To resolve the query, with the utterances obtained from Experiment 5, we will move further to study whether such speaking rate-dependent change can cause a corresponding shift in perception of the target tone. If so, there is an instance of sound change demonstrated.

---

[34] There have been quite extensive debates on the validity to represent a rising tone as two consecutive tone levels. Rather than go into such details, we make use of the notation here simply for the sake of building the bridge between phonological processes and our acoustic observations.

Methodology

Ten Native Cantonese speakers were recruited to take part in Experiment 6. These subjects all had not participated in Experiment 5 to produce the utterances so as to avoid any priming effect. The listening task was conducted in a quiet room. After a short practice session to familiarize subjects with format of the test, the real test began. They were to complete a forced-choice tone identification task to circle the characters heard on the given answer sheet. Breaks were provided between sessions upon request to relieve subjects' fatigue.

To simulate a natural communication scenario as much as possible, natural tokens obtained in Experiment 5 were used. Three out of the ten repetitions (repetitions for averaging out token idiosyncrasies) for each condition were determined randomly and chosen as the material. To simplify the experiment, only conditions with LHL (*maa4-lau1-maa4*) and LRL (*maa4-lau2-maa4*) were used. As a result, (5 speakers) × (2 target tones) × (3 speaking rates) × (3 chosen tokens) = 90 tokens were presented in a random order to each participant. Before presentation to subjects, intensity of all the sound samples was normalized to 80dB to minimize possible influence due to cross-subject / cross-token intensity fluctuations.

These sets of 90 stimuli, grouped in a randomized order into 10 blocks of 9 tokens, interleaved with silence interval of one second (i.e. ISI = 1s), were presented to subjects in a quiet venue, with speaker volume tuned to a comfortable level. The participants were requested to identify the target characters (S_t) in the stimulus sentence "*ngo5 heoi3 maa4* S_t *maa4 maai5 sung3*" 'I go to *maa4* S_t *maa4* to buy some food' by marking the corresponding Chinese characters. Two characters were available to be circled as the answer, namely, *lau1* and *lau2*.

Results and discussion

Figure 5.5 shows the tone perception error rate across different speaking rate conditions. Compared to perception of coda in Experiment 2, tone confusion is much lower, starting from 2.67%, and reaching 16.67% as speaking rate increases. The F_0 peak delay under high speaking rate, as obtained in Experiment 5, is correlated with a higher tonal confusion rate.

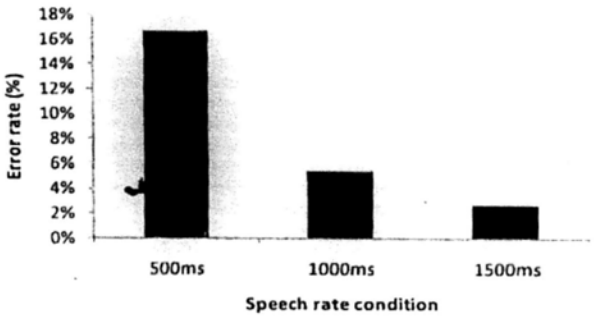


Figure 5.5: Tone perception error rate across speech rates.

Next we have a break-down of the tonal errors into different source speakers of the stimuli in Figure 5.6. It can be observed that errors from the stimuli produced by S1 and S4 clearly outnumber those from other speakers. Recall from Figure 5.3c and 5.4a that the subjects S1 and S4 showed the most obvious F_0 peak delay under high speaking rate, now we can be even more confident that F_0 peak delay leads to the observed tonal confusion.

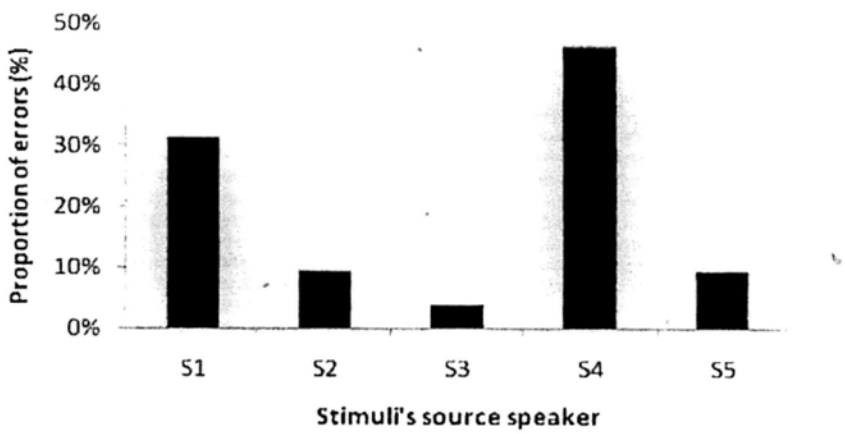


Figure 5.6: Proportion of errors from stimuli by different speakers.

Regarding the direction of tonal confusion, we do not obtain the expected result. Since acoustic data from Experiment 5 shows a close resemblance of LHL sequence to LRL one under high speaking rate, we thus expect a corresponding mis-identification of LHL stimuli as LRL under high speaking rate. In fact, we found that, especially for stimuli from speakers S1 and S4, some tokens produced as LRL were perceived as LHL.

We have not yet been able to find a definite answer to these results. We suspect that when listeners were presented with F_0-peak delayed stimuli, they made use of

other co-varying cues to resolve the tonal identity. Such co-varying acoustic variations other than F_0 directed them to choose the final answer.

Experiment 6 could not give full-blown evidence of progressive assimilation in rapid speech at the perception stage. The widely attested progressive tonal assimilation probably comes from other factors, which await further investigation. What we can be sure from these results are that (1) high speaking rate is one of the catalysts behind tonal confusion, which is a source of sound change; (2) F_0 peak delay observed in rapid speech is directly correlated with such tonal confusion.

5.3 Summary of findings

In the chapter, we presented our attempt to extend the methodology used in segmental feature assimilation to the tonal domain, though only with partial success.

In Experiment 5, speakers produced utterances at various requested speaking rates. F_0 peaks were taken as a measure to show the acoustic rightward shift of F_0 contours at high speaking rates. Translated phonologically, rightward spreading of L tone feature was observed.

A subset of natural utterances obtained in Experiment 5 were presented to participants in Experiment 6 to elicit their perceptual responses. Unexpectedly, the acoustically more similar (in terms of F_0 peak) stimuli obtained under high speaking rates did not lead them to mis-identify the tones (intended T_1 identified as T_2). We thus suspect other co-varying cues, say, F_0 peak level, F_0 range, may play a role in identification of the tones, from the distorted F_0 signals resulting from high speaking rate.

CHAPTER 6

TONAL CHANGE - QUESTIONNAIRE SURVEY

This chapter reports a large-scale questionnaire survey (Experiment 7) to quantitatively document trajectories of possibly ongoing tone-mergers (T_2/T_5 and T_3/T_6) in Hong Kong Cantonese.

6.1 Related studies

Table 6.1 summarizes the tone letters assigned to the six long tones in Cantonese by different scholars. Since the tone values were assigned based on different criteria and samples were collected at different times in history, minute deviations are inevitable. Nevertheless, they correspond quite well to Figure 6.1, an F_0 plot reproduced from Figure 2 in Wong (2007). Each contour in Figure 6.1 is obtained by averaging 20 tokens from four native Cantonese speakers. In the figure, our first target tone pair, T_2/T_5, starts with nearly the same onset, diverges to rise to different F_0 values towards its offset. T_3 and T_6, our second targets, are both assuming a level contour with slight declination, with a mere pitch difference of 10Hz.

	T_1	T_2	T_3	T_4	T_5	T_6
Chao (1947)	53	35	33	21	23	22
Hashimoto (1972)	53	35	44	21	24	33
Vance (1976)	55	35	33	11	13	22
Zee (1991)	55	24	33	21	23	22
Bauer & Benedict (1997)	55	25	33	21	23	22

Table 6.1: Tone-letter labels by different scholars.

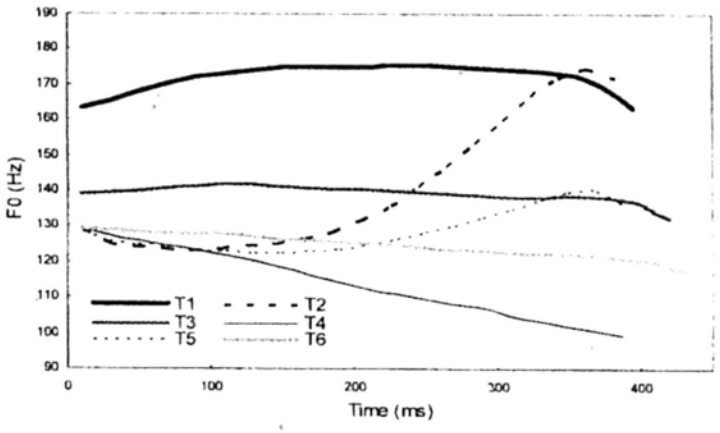


Figure 6.1: F_0 contour of the six long tones in Hong Kong Cantonese (from Wong, 2007).

Generally speaking, except T_1, which is located relatively distant from other tones, the lower tonal space of Hong Kong Cantonese is rather crowded. Obviously, judging from the F_0 values, the primary acoustic correlate of linguistic tone, many tone-pairs are expected to cause confusion. Yet, T_2/T_5 and T_3/T_6 are the most widely reported among them, as reviewed below.

So (1996) gave a brief overview of several changes in rising and level tones in Hong Kong Cantonese. In the discussion on tonal contours of the two rising tones (T_2 and T_5), the author mentioned that a previous study (So & Varley, 1991), involving 101 Hong Kong Cantonese speakers, reported that the participants often confused the two rising tones. They attributed this to the similar onset fundamental frequency of the two tones.

Mok & Wong (2010) conducted perception tasks to investigate the status of perception of merging tones in Hong Kong Cantonese. There were two groups of participants, one *merger* group consisting of subjects doing worse than average in a screening test, and one *control* group who discriminated all tones well. They had to go through an AX discrimination task, responding either 'same' or 'different' after listening to each sequence of two testing syllables. Results showed that the control group generally did better, as expected, than the merger group in terms of both accuracy rate and reaction time. However, same as the merger group, they also find the pair T_2-T_5 the most difficult to distinguish.

Bauer *et al.* (2003) carried out a production study to investigate tone production in Cantonese. Two subjects of interest were found. One of them merged T_2 into T_5, while the other T_5 into T_2.

In Peng & Wang's (2005) pursuit of an accurate tone recognizer, using SVM techniques, 71.5% accuracy for tones in Cantonese continuous speech was achieved. What is relevant to the current discussion is their recognition results in Tables 6-8, where T_2 is confused with T_5, and T_3 with T_6. Of course, it cannot be assumed that machine recognition and human perception are essentially the same. Yet, when human subjects were presented T_3 and T_6 tokens that were confused by the tone recognizer, they also made considerable erroneous responses, bearing some degree of resemblance to machine recognition. Since the tone recognizer mainly considered F_0 contours of the speech samples, the mis-recognition results implied that T_2-T_5 and T_3-T_6 are associated with similar F_0 contours, the most prominent acoustic cue to tone perception (Abramson, 1972; Lin, 1988, among others).

Peng's (2006) paper reported a large-scale survey of corpus analysis on tones in Mandarin and Cantonese. By means of automatic speech segmentation and pitch extraction, a large pool of tokens were collected and analyzed. In particular, for the tonal aspect, two pitch dimensions, *height* and *slope* were used to represent the token distribution (Figures 7-8). The Cantonese tones showed heavy overlap between T_2 and T_5, as well as between T_3 and T_6 in the long tone figure (Figure 8), indicating a possible confusion of the two tone-pairs in every speech communication.

Though from different perspectives, these reports and results all carry the same theme that the ongoing status of tone-mergers of T_2/T_5 and T_3/T_6 is unarguable. Given this, we wanted to further our understanding of its developmental dynamics, such as the speed of its development, its penetration into the speech community, age-effect, and etc. To achieve so, we applied the *age-grading*[35] method on Hong Kong Cantonese tones to reveal its developmental trajectory of merging.

---

[35] For some applications on studying modern Chinese dialects, like Tianjin tone-sandhi and tonal usage in producing transliterated English loanwords, readers are referred to a report (Wong, 2009b) on the first Conference in Evolutionary Linguistics.

Briefly speaking, the fundamental principle is to survey synchronic variations across individuals from different age groups and make use of such variations to project diachronic changes not directly observable otherwise (Labov, 1966). Shen's (1997) monograph played an important role in publicizing the age-grading method. In the monograph, Shen elicited responses from subjects from a wide age range to obtain trajectories of vowel mergers in Shanghai and Wenzhou dialects. The major finding was that the younger the subjects were, the more confusion between words with the two vowels under study resulted.

According to Shen (1997), the age-grading methodology could be traced back to nearly a century ago in studies by Gauchat (1905) and Hermann (1929). The method works based on the assumption that linguistic behavior of subjects is fossilized at their language acquisition stage of life, say, 15 years old. Consequently, responses from subjects from different age groups reveal linguistic forms that once appeared in different periods in history. In other words, Shen's studies implied that the Shanghai and Wenzhou vowel pairs concerned were more distinctive back in history than the time he carried out the survey.

6.2 Methodology

Following the age-grading paradigm, we elicited tonal responses of subjects of different ages in the format of questionnaire survey. Presented with the questionnaire, participants were requested to indicate the character with unlike pronunciation. There was no time limit for completing the questionnaire. Options were also given for the case where the subject found all the pronunciation to be the same, and the case where the subject was unsure about the answer.

We selected the target tokens from 粵語審音配詞字庫[36]. To form each question item, three characters were selected from the database. To avoid guessing in situations where subjects were unfamiliar with the characters, we only chose common characters for the survey. To minimize possible interference from orthography, all the

[36] The corpus can be accessed at http://arts.cuhk.edu.hk/Lexis/lexi-can/. It was produced by the Research Centre for Humanities Computing, The Chinese University of Hong Kong.

characters in the same question item were orthographically dissimilar such that pairs like (管, 館) *gun2* and (艦, 纜) *laam6* were excluded. Also, only characters with a unique pronunciation across all contexts were used to avoid confusion (e.g. 死, pronounced as either *sei2* and *si2*, was avoided). In the survey, we did not include characters with entering tones.

Based on the selection criteria above, there were altogether 36 items in the questionnaire, categorized into three groups according to the pronunciation contrast: 16 items on T_2/T_5 distinction, 16 items on T_3/T_6 distinction, and the remaining 4 items as distracters. For the first two groups, one among the three characters differed from the remaining two in terms of tonal specification. The number of characters in each tonal category in the first two groups was balanced. For instance, there were 8 items with two T_2 and one T_5 characters, and another 8 items with two T_5 and one T_2 characters within the group of T_2/T_5 distinction; The distracter items served two functions: They reduced the monotonicity in completing the questionnaire by presenting pronunciation contrasts other than the target tonal distinctions[37]. Besides, since those items were expected to be highly distinguishable by a normal Cantonese speaker, they could help detect questionnaire completed based on pure guessing. All the 36 items were listed in the questionnaire in a randomized order. Besides the questionnaire items, subjects were requested to enter information about their exposure to Cantonese, like education level, and year of residence in Hong Kong. Full details about the questionnaire like the list of items and the exact layout can be found in the Appendix.

There were altogether 137 (67M, 70F) subjects contributing to the current data analysis. Majority of them completed the questionnaire in a face-to-face survey, while the others submitted the completed questionnaire through electronic mail. The subjects were aged 13-33, with the exact age frequency distribution shown in Figure 6.2.

---

[37] The distracter triplets are #4 (來, 朱, 豬) (*loi4*, *zyu1*, *zyu1*), #6 (紙, 指, 只) (*zi2*, *zi2*, *zi2*), #11 (真, 珍, 陳) (*zan1*, *zan1*, *can4*) and #32 (他, 流, 留) (*taa1*, *lau4*, *lau4*). Items #4, #11 and #32 all have the unlike

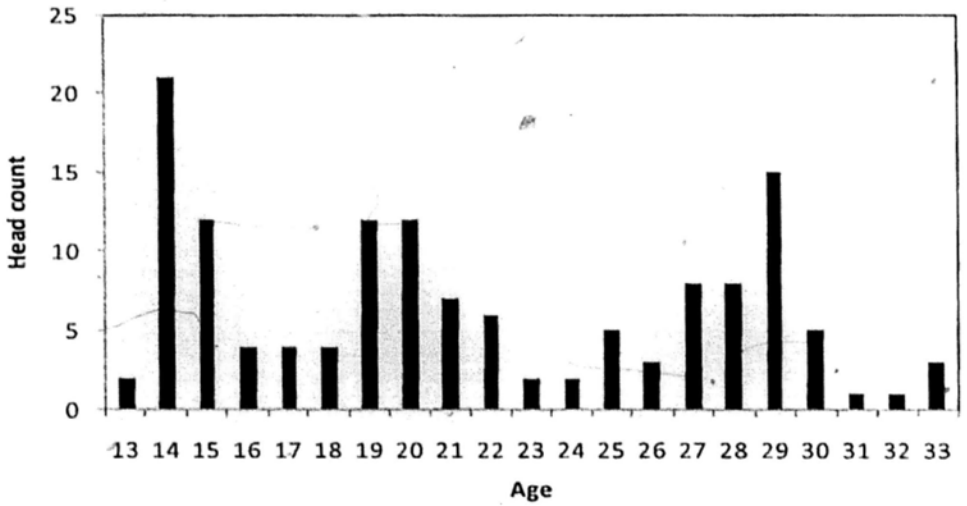


Figure 6.2: Age distribution of participants of the questionnaire survey.

6.3 Results and discussion

Subjects were requested to try their best to have every question attempted. Cases were found for some subjects to leave a few questions unnoticed, but they were requested to fill in the missing items right after they returned the questionnaire. For the distracter items, only 4.74% (26 tokens)[38] were found to be wrong, indicating that all the participants did not complete the questionnaire by pure guessing.

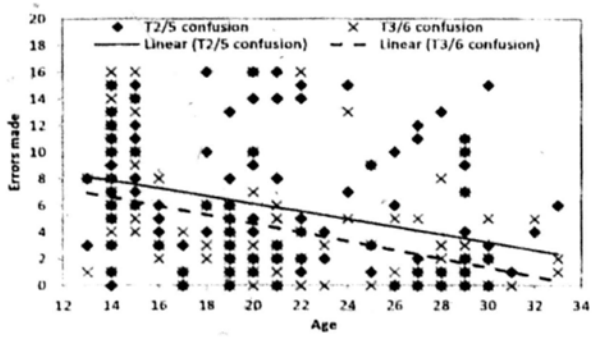
We now try to explore age effect of tonal confusion[39] in Hong Kong Cantonese speakers. Figures 6.3 plot the confusion rate of different tones obtained from subjects of different ages. *x*-axis denotes the subject age, while *y*-axis the number of errors made in discriminating the tones. The three figures differ in terms of age grouping. Figure 6.3a shows the individual results; Figure 6.3b shows the average confusion

---

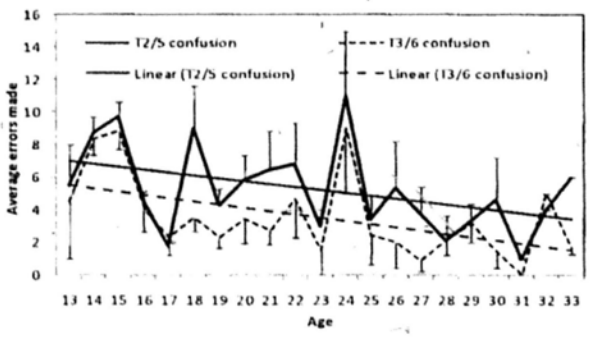
character differing from the others in both segmental composition and tone. Characters in item #6 all sound the same, and subjects were expected to choose 'the same' for this item.

[38] Majority of them (23 out of 26) happened in item #6 (紙, 指, 只), where all three items are pronounced as *zi2*. It is conjectured that subjects already got used to assume one item must be different, and could not switch to choose the response "all the same".

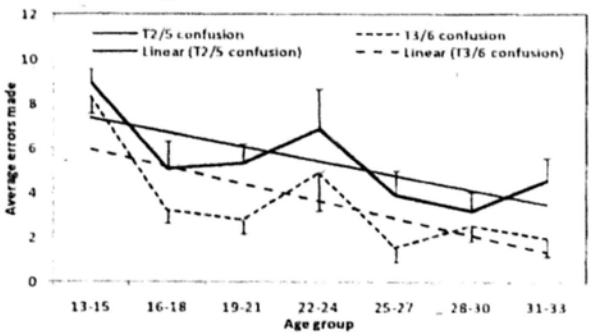
rate for each age; Figure 6.3c shows the average by grouping subjects from a range of 3 years. Linear regression results are also shown in the figures. Regardless of the age-grouping used, there is a general decreasing trend in confusion rate as the age of participants grows higher. Based on the basic assumption of age-grading that linguistic forms of subjects of different ages represent different historical linguistic forms, a tone merger can be observed: T_2/T_5 and T_3/T_6 were more discriminable back in history than now.



(a) individual subjects



(b) average by age



(c) average by age group (3 years)

Figure 6.3: Age distribution of participants of the questionnaire survey.

Comparing the two tone pairs, T_2/T_5 is consistently more severely confused than T_3/T_6. Judging from the steeper slope of the linear-regression lines for T_3/T_6 in

[39] Confusion here merely means that the tonal distinction present in the source dictionary corpus is absent from the subject's response. There is no prescriptive judgement of the subjects' response since

Figure 6.3, the age-dependent differential confusion rate mentioned above is slightly more obvious for T_3/T_6[40]. In other words, the pace of tone merger is faster. Regarding the sex factor, male subjects (6.13 for T_2/T_5, 5.57 for T_3/T_6) did worse than female ones (4.36 for T_2/T_5, 3.81 for T_3/T_6). However, the individual variability, measured in terms of standard deviation, is larger for male than female subjects.

Statistical significance of the results obtained above is tested by conducting a 3-way repeated-measures ANOVA. *Sex* (male vs. female), *age-group* (3-years in a group), and *tone-pair* (T_2/T_5 vs. T_3/T_6) are used as the factors (IV) and *confusion rate* as the dependent variable (DV). Main effect is obtained for *age-group* ($F = 14.545$, $p < 0.01$) and *tone-pair* ($F = 7.913$, $p < 0.01$) but not *sex* ($F = 2.830$, $p = 0.094$).

All the above results were obtained by treating all the tokens equal. We proceed to investigate confusability among different stimulus tokens, as plotted in Figure 6.4. x-axis shows the probability that a particular token (in y-axis) is confused by the subjects, sorted in descending order. Sum of the probability figures for T_3/T_6 is lower than T_2/T_5, as expected. Interestingly, there is quite a large variation in the confusion probability among the tokens, ranging from 0.10 to more than 0.65. This gives solid evidence to lexical diffusion theory (Wang, 1969), where a sound change (i.e. the likely tone merger from tone confusion obtained in our case) is hypothesized to originate from a small group of items in the lexicon, and gradually spreads to other phonologically similar items.

---

there is no eternal right or wrong in the context of diachronic sound changes.

[40] The linear regression results are, for Figure 6.2a, T_2/T_5: $y = -0.290x + 11.932$: $y = -0.330x + 11.245$; for Figure 6.2b, T_2/T_5: $y = -0.179x + 7.208$, T_3/T_6: $y = -0.202x + 5.723$; for Figure 6.2c, T_2/T_5: $y = -0.646x + 8.015$, T_3/T_6: $y = -0.772x + 6.712$.

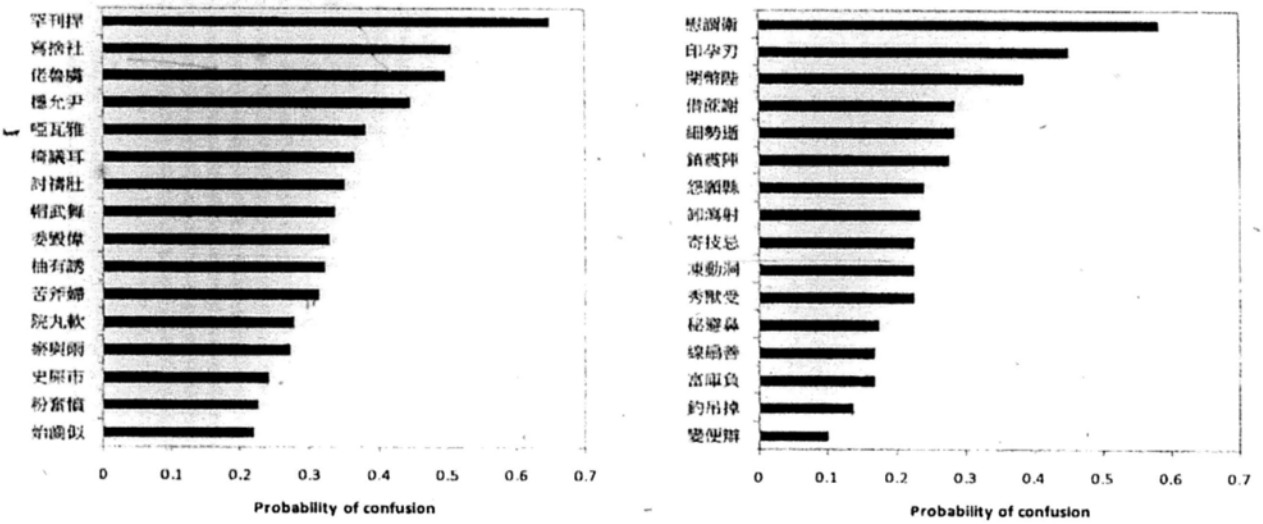


Figure 6.4: Tone confusion rate plotted against different tokens for T_2/T_5 (left) and T_3/T_6 (right).

No two lexical items in a particular language share exactly the same role in day-to-day communication scenarios, be it oral or written. Such difference may cause the observed differential confusion rate among characters obtained from the questionnaire survey. Two such closely related factors, *lexical frequency* and *segmental sonority*, will be studied now, as detailed below.

Lexical frequency effect

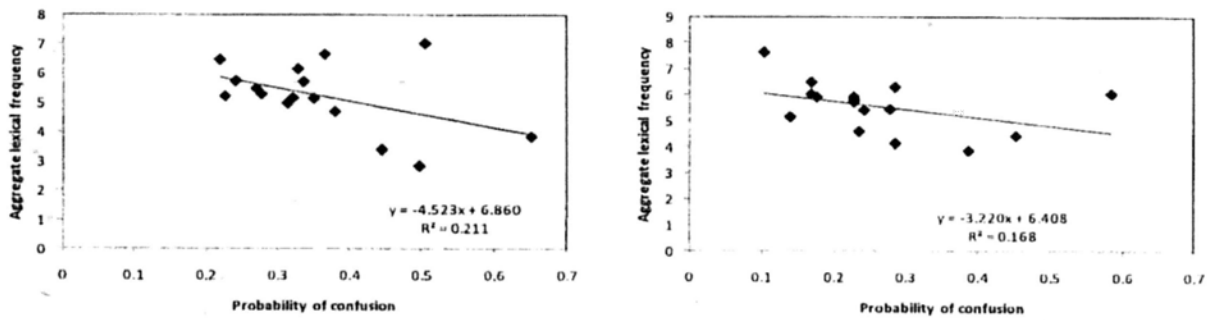
Depending on likeliness of occurrence, pronunciation of a lexical item can be stabilized (i.e. not to change) or drifted (i.e. to change) quite differently compared to others in a language. Here, we explore the relationship between *lexical frequency* and the confusion rate by making use of an online Chinese character frequency database named "Hong Kong, Mainland China & Taiwan: Chinese Character Frequency - A Trans-Regional, Diachronic Survey"[41], prepared by the Chinese University of Hong Kong. It contains more than three million characters obtained from written materials like magazine, newspaper published in Hong Kong, mainland China, and Taiwan.

[41]The database is also titled in Chinese as "香港、大陸、台灣 - 跨地區、跨年代: 現代漢語常用字頻率統計" and is publicly accessible at http://arts.cuhk.edu.hk/Lexis/chifreq/.

The written materials are divided into two groups according to their publication date, 60's (1960-1969), and 8/90's (1980-1993).

For each of characters used in our questionnaire survey, we query its frequency from the character frequency database. In particular, we limit our search by specifying the region as Hong Kong and period as 8/90's, which is supposedly the closest to that experienced by the participants of the questionnaire survey.

Each item in the questionnaire consists of three characters. To illustrate the frequency effect, we obtain its total frequency by the formula $freq = \log(f_1) + \log(f_2) + \log(f_3)$, where f_1, f_2 and f_3 denote the frequencies of the component characters. Confusion probability is plotted against lexical frequency in Figure 6.5. The linear regression results included show that more frequent characters generally have a lower tone confusion rate, both for $T_2/_5$ and T_3/T_6.

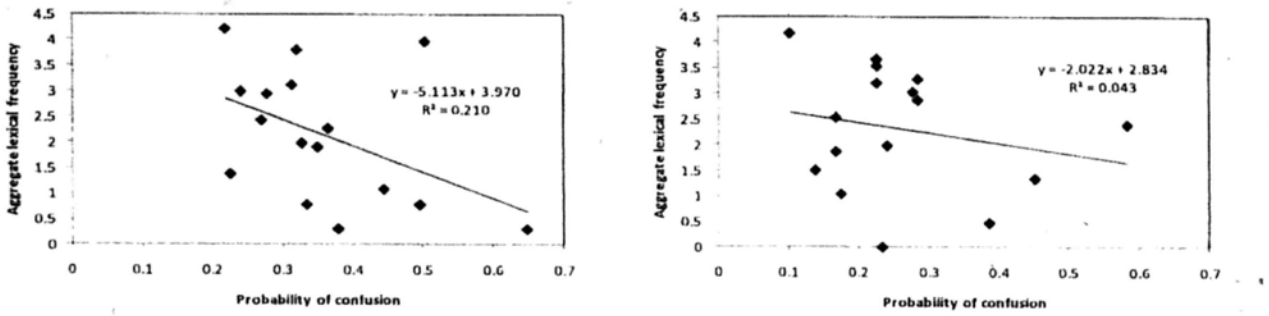


(a) T_2/T_5 (b) T_3/T_6

Figure 6.5: Lexical frequency (written) effect on tone confusion rate.

Results in Figure 6.5 come from a frequency database sampling only written material, the frequency data obtained may not fully represent the frequency of individual characters as experienced by the language users in their day-to-day oral communication. For instance, the characters 屎 'faeces' and 瘀 'shameful' are frequently used in colloquial oral communication, but they have extremely low frequency in the database we used. A parallel frequency analysis with spoken corpus is necessary for helping us verify our results.

We made this attempt with the HKCAC spoken corpus (Leung & Law, 2001) we used in Experiment 3, which contains phonetically transcribed data of hours of radio programs. Similar data extraction and analysis procedures as for Figure 6.5 were involved to obtain results plotted in Figure 6.6.



(a) T_2/T_5 (b) T_3/T_6

Figure 6.6: Lexical frequency (spoken) effect on tone confusion rate.

Comparing corresponding plots of Figure 6.5 and 6.6, we may find much similarity in the two sets of results. Confusion rate varies inversely with lexical frequency, for both tone pairs. Besides, linear regression results also agree in terms of their relative magnitude (that T_2/T_5 is associated with a steeper slope). This indicates that the observed frequency-confusability correlation is quite consistent even after accounting for the oral-written language differences.

Segmental sonority effect

Next, we explore the relationship between *phonological sonority* and tonal confusion rate. As Katamba (1989, p159) stated, "The phonological sonority hierarchy has the phonetic correlates of openness and propensity for voicing. The more sonorous a sound is, the more audible it is likely to be", sonority roughly corresponds to the auditory visibility of a linguistic item to listeners.

Many sonority hierarchies have been proposed by various scholars[42], which do have differences among them. To facilitate investigation, taking the common factors

---

[42] Here are some example sonority hierarchies proposed by different scholars: *voiceless obstruents* < *voiced obstruents* < *nasals* < *liquids* < *glides* < *vowels* in Katamba (1989); *complex plosives* < *voiceless plosives* < *voiced plosives* < *voiceless fricatives* < *voiced fricatives* < *nasals* < *laterals* < *flaps* < *high vowels / glides* < *mid vowels* < *low vowels* in Burquest and Payne (1993, p101); *voiceless stops / voiceless fricatives* < *voiced stops* < *voiced fricatives* < *voiced nasals / voiced laterals* < *voiced*

from these proposals, we establish a simplified sonority scoring scheme for each character, as discussed below.

First, each character is decomposed into two constituents, namely, *onset* and *rhyme*. Following that a sonority score is given for each component according to the scheme as shown in Figure 6.7. Generally speaking, voiced onset (e.g. /j/, /l/, /m/) is more sonorous than its voiceless counterpart (e.g. /f/, /t^h/, /s/), and a low vowel (e.g. /a/) is more sonorous than a high one (e.g. /i/, /y/). The sonority score of a particular character is simply the sum of these two scores. The character 市 'market' /si/, for instance, has a voiceless sibilant onset (score = 1) and a high main vowel (score = 1), summing to a sonority score of 2. As another example, the character 衛 'to guard' /wai/ is much more sonorous, having a sonority score of 5, from its voiced onset /w/ (score = 2) and low main-vowel /a/ (score = 3).

type	score
[-vce]	1
[+vce]	2

(a) consonant voicing

type	score
high	1
mid	2
low	3

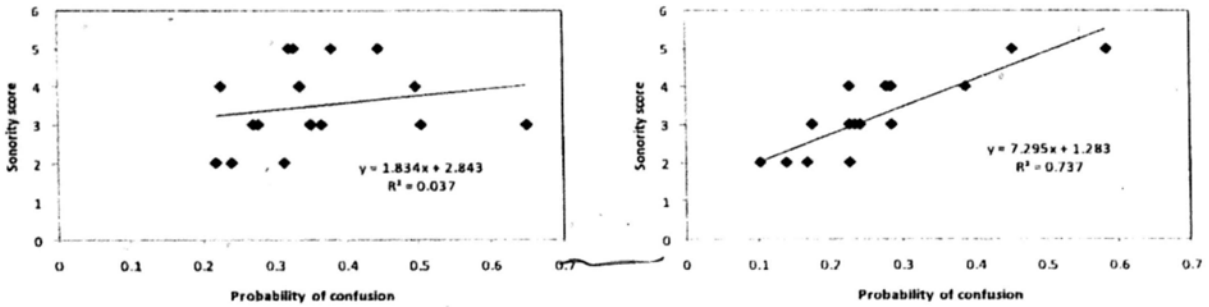
(b) vowel height

Figure 6.7: Scoring scheme for onset consonant voicing and height of the main vowel in the rhyme.

Figure 6.8 plots sonority score against tone confusion rate for the two tonal pairs. Linear regression results show a direct relationship between the two. In other words, characters which are more sonorous yielded higher tone confusion in the current questionnaire survey. The trend is more obvious for T_3/T_6, as revealed by the goodness-of-fit value R^2.

---

r-sounds < *voiced high vowels* < *voiced mid vowels* < *voiced low vowels* by Jespersen (1904) quoted in Clements (1990).



(a) T_2/T_5 (b) T_3/T_6

Figure 6.8: Effect of phonological sonority on tone confusion rate.

Discussions

Niyogi (2006), from a mathematical perspective, talked about Shen's (1997) Shanghainese study in his book on the computational nature of language evolution. In the chapter, he proposed four cases, varying in two dimensions (*forms used by an individual*, and *learning strategy*), to model language evolution observed in Shen's data. Assuming that there exist two forms of a word in a speech community, in cases 1 and 2, individual speakers only acquire and use one form of the word, while they acquire and use two variants of the word in cases 3 and 4. Along another dimension, individuals in case 1 and case 3 choose to learn the word from the entire speech community, while they only learn from their parents in cases 2 and 4.

From the mathematical formulation, the first two cases are unstable, with one of the two variants driven to extinction eventually given even a slight drift from the unstable balance. The last two allow variations to be maintained over time. Apparently, cases 1 and 2 are closer to the Shen's data, since the other two can accommodate variations forever. However, the sound change Shen studied still lacked some critical elements, like substantial language contact, and a triggering point to initiate sound change.

The ongoing tone merger we are studying seems to resemble case 1 and 2 more since results also show a trend of one of the word forms (tonal realization) disappearing. Different from Shen's data, Hong Kong is always full of speakers with different native languages, and thus language contact, which was lacking in Shen's case.

Successful modeling of empirical data can open a whole new horizon of possibilities and answers to the line of research on language evolution. To facilitate a more rigorous verification of empirical data against the models proposed, our questionnaire survey may further be modified such that (1) repetitions are presented to count the number of forms an individual uses; (2) subjects surveyed are grouped into generations from the same family, to investigate the source of individual's word form.

There are some points to notice regarding observations from the current questionnaire survey:

For participants of the questionnaire survey, a wider age spectrum should render results in this study more representative. However, it must be aware that, the questionnaire has to be formulated easy enough for participation from all ages, especially those young students and elderly people, such that factors other than their tone discrimination judgment (e.g. literacy, memory) does not affect the results.

Besides, the current study cannot accurately point out the directionality of sound change. Two scenarios are used here to illustrate the point, both involving the character triplet ($c_1 = T_x$, $c_2 = T_x$, $c_3 = T_y$). Suppose a subject responds 'all the same' in scenario 1. From the data alone, we have no way to determine whether it is the case where c_1 and c_2 has changed to T_y or it is c_3 which changed to T_x. In scenario 2, c_3 is chosen to sound different. This can mean (1) no change of pronunciation; (2) c_1 and c_2 changed to T_y while c_3 changed to T_x; (3) c_3 changed to a new tone T_z. To be more precise, other combinations can happen while yielding the same discrimination responses for the above two scenarios. This illustrates the limitation that the current study can at most tell the situation of lexical confusion among language users. To further determine the directionality of sound change, evidence from other perspectives must be recruited.

For the frequency analysis, the two types of corpora we used in Figure 6.5 and 6.6 have considerable difference in their size (nearly 140 thousands vs. more than three millions), as reflected in the smaller maximum value in the y-axes of Figure 6.6. Of course, a spoken corpus of a comparable size is favorable for a fairer comparison. However, it must be noted that laborious transcribing work is involved in spoken corpus construction, but not for written ones. Besides, the HKCAC corpus we used collected data from radio programs, supposedly with an above-average audio

recording quality, for more accurate data transcription. This inevitably introduced some inherent biases, however, like the filtering of irritating content, and the trained standardized pronunciation of the program hosts, which are lacking for the untrained.

For the sonority analysis, individual constituent segments have non-linear effect on acoustics of the syllable, so is its perceptibility. If this is taken into consideration in formulation of the sonority calculation, the result may better reflect the reality.

6.4 Summary of findings

In this chapter, we reported a large-scale questionnaire survey, revealing quantitatively the current status of the widely reported T_2/T_5 and T_3/T_6 tone mergers in Hong Kong Cantonese.

Evidence of tone-merger was found in both tone pairs, with age-dependent effect where younger subjects made more confusion in the tone discrimination task. Among the tone pairs, the rising tones T_2/T_5 were more severely confused by the participants than the level tones T_3/T_6. These results have been verified as statistically significant. Rate of tone merger is found to be higher for T_3/T_6. Besides, with regard to sex, male subjects made more tone errors than female subjects.

Based on the confusion rate of individual tokens, two factors, *lexical frequency* and *phonological sonority* were engaged in an attempt to explain the differential progress of tone-merger across lexicon. Essentially, confusion rate varies directly with segmental sonority but inversely with lexical frequency. These preliminary results suggest the need for further investigation.

CHAPTER 7

CONCLUSIONS AND DISCUSSION

This dissertation demonstrates our effort to propose speaking rate fluctuation as one of the mechanisms underlying the directional bias of speech assimilation observed cross-linguistically.

First, speech production experiments (Experiment 1 and 5) were designed to elicit acoustic variations when language users produce speech utterances under a tight time constraint. The resultant speech samples were then presented to listeners in follow-up speech perception experiments (Experiment 2 and 6) to investigate perceptual consequences of the observed acoustic variations (e.g. formant and F_0 trajectories).

Next, to guarantee the general applicability of our findings obtained with tight experimental control, a large-scale spoken corpus was used to analyze speech confusion patterns. In addition, based on relative frequencies of various stop consonants, computer modeling efforts were made. All the results obtained pointed to the fact that the directionality bias obtained experimentally is in fact observable in the general speech community.

Lastly, focusing on the widely attested tone mergers in Hong Kong Cantonese, a large-scale questionnaire survey was conducted. Using the age-grading method, subjects of different ages were recruited to elicit synchronic variations in their categorization of tonal categories. Older subjects, who were assumed to reflect a pronunciation once existed further back in history, showed clearer tonal distinctions in both tonal pairs under study, compared to younger ones. The tone mergers were shown to progress in a lexical-diffusion manner. Several hypothesized underlying factors were proposed and tested.

Besides the experimental results, below lists some topics worth further discussions, as well as some directions for future follow-up explorations.

7.1 Experimental historical phonology

To pursue knowledge in a scientific manner, one has to work past the stage of relying on merely impressionistic judgment to set up testable and falsifiable[43] hypotheses, and proceed with experimentation. This also holds in linguistics studies if we treat the relevant research seriously.

Ohala (1974) put forward the idea of taking an experimental approach for doing historical phonology research. Instead of waiting for millennia for sound changes to proceed and complete[44], man-made experiments were suggested instead. By safely assuming that the mechanisms (e.g. articulatory and auditory constraints) driving sound changes are the same over history and at the present time, hypotheses can be set up and verified experimentally in the laboratory. Several examples were given in the paper to demonstrate experimentally how some production and perception constraints could cause confusions in speech communication.

Following the same philosophy, Hombert is another scholar pioneering this idea in the phonetics field. He studied the widely reported diachronic development of tones from onset consonant voicing (Hombert (1978), Hombert *et al.* (1979))[45]. In the studies, the whole speech communication chain was divided into several sub-processes: With production experiments, they first measured the magnitude and

---

[43] Karl Popper (1902-1994), a recent respected philosopher of science, considered a theory scientific if and only if it is falsifiable.

[44] It is not uncommon to find such time-consuming sound changes. E.g, the Great Vowel Shift in Middle English, and tone mergers in Middle Chinese, etc. are just too slow to be observed in a synchronic setting. Of course, there are also numerous scenarios (e.g. Shen (1997) and Experiment 7 in this dissertation) in which noticeable variations can already be observed within a speech community.

[45] Hombert (1977) was another attempt to investigate tone development from vowel height. Intrinsic vowel pitch is the tendency for low vowels to be produced at a lower F_0, and vice versa. Following a similar logic as tonogenesis from consonant voicing, vowel height feature should give rise to tone development. But it happens rarely, if at all, in reality. Perception results from the study gave an explanation by showing a perceptual compensation effect: low vowels like [a] were judged higher in pitch and high vowels like [i] or [u] were judged lower. Consequently, the perceived pitch difference

duration of F_0 perturbation due to consonant voicing feature. Next, with synthesized stimuli of the vowel [i], they elicited listeners' ability to discriminate pitch contours with differently sloped F_0 onsets. Subjects started to hear the difference when slope of the F_0 contour is 60ms long, considerably sooner than the end-point of the voicing-induced F_0 perturbation. Combining these results, the conditions under which tonogenesis likely occurs, were found.

Xu (2010) is a good piece of writing trying to defend for the use of laboratory speech in conducting phonetic research. Many stereotyped characterizations about experimentally elicited speech were discussed and shown to be untrue in the paper. Not totally rejecting the use of spontaneous speech in phonetic research, Xu argued that with careful experimental design, laboratory speech provides far more true insights about the nature of speech production.

Age-grading (previously introduced in Section 6.1) is yet another experimental approach to study the diachronic dynamics of sound changes from a synchronic snapshot of speech production / perception patterns. By sampling speech production / perception behavior of subjects from different age groups, the once-existing historical sounds can be recovered. Although the inter-subject variation is inevitably more significant due to various non-linguistic differences across age groups[46], it is already more precise than merely comparing literature records obtained from different historical periods[47]. By assuming that the same linguistic principle currently working should also work in the past and will work in the future, any insight we gain about language change can be used to project to a similar past event, which is no longer accessible (Labov, 1974).

With advancement of technologies, computer simulation has become another powerful tool to study historical sound changes (e.g. Niyogi (2006), Wang *et al.*

---

between vowels of different height was much smaller than the corresponding acoustic difference in F_0, minimizing the likeliness of perceivable F_0 distinction, and thus the chance of tonogenesis.

[46] Some examples of non-linguistic differences are memory capacity, visual and auditory acuity under the effect of biological aging.

[47] Wong (2009b) contains a more detailed discussion and as well a comparison between these two approaches.

(2004), among others). The basic idea is that, computer agents, in the form of software objects, perform linguistic interactions with one another according to a set of linguistic assumptions. The eventual evolution or emergence of linguistic behavior, even after an extended period of time, can be effectively projected given sufficient computational power. The advantage of computer modeling is that all the assumptions about the linguistic phenomenon being studied have to be explicitly made (Wang *et al.*, 2004). For example, our modeling study in Section 4.2 made clear assumption of the parameters of coda consonant evolution such as the coda diffusion and natural attrition rates. It is exactly these clear assumptions made which enabled us to examine even the slightest shift in the outcome with a little twist of the parameters.

With reference to the above literature, the current dissertation takes an experimental approach in studying speech sound evolution. We first reproduced experimentally (in Experiment 1 and 2) the target sound change (i.e. regressive assimilation) process in the laboratory. Provided sufficient speaking rate variations, participants showed perceptual confusions, especially for rapid-speech samples. Given that speech rate fluctuation is omnipresent in everyday speech communication scenarios, we then hypothesized that though iterations of oral communication, such observed minute probabilistic perceptual confusion would accumulate to cause a diachronic categorical (phonological) shift.

Apart from obtaining behavioral data (i.e. acoustic measurement and identification responses) under tightly controlled laboratory settings, we also tried to draw additional empirical support from real-world languages, in our case, Hong Kong Cantonese. Through both synchronic and diachronic analyses of spoken corpora, with supplementary simulation modeling, we were able to confirm that the regressive assimilation process is indeed an ongoing sound change, and that speaking rate variation is one of the driving forces behind.

Lastly, we followed Shen (1997) to apply the age-grading method to study ongoing T_2/T_5 and T_3/T_6 mergers in Hong Kong Cantonese. With subject age as the independent variable, we were able to unveil the historical trajectory of the target tone-mergers.

In sum, we conducted a multi-perspective study on Hong Kong Cantonese sound changes, collecting experimental and empirical evidence from methodologies like acoustics, perception, corpus analysis, simulation modeling, and age-grading.

7.2 Mechanism of sound change

There is a major divergence in the view on the process of sound change: Neo-grammarian hypothesis suggests the *regularity* of sound change, i.e. it applies simultaneously, according to exactly the same schedule, to all the words in which the conditioning context is found, implying the existence of a *homogeneous* speech community. In contrast to that, the lexical diffusion theory, pioneered by Wang (1969), emphasized *lexical gradualness*, i.e. sound changes originate in a small number of words, and later on spread to other words with a similar phonological makeup. Besides, *heterogeneity* is also recognized in different subgroups in the speech community.

The speaking-rate induced sound changes we were dealing with in this dissertation were context-dependent. For instance, Experiment 1 to 4 concerned coda consonants regressively assimilated by the following onset consonant. Compared to other characters with an [-n] coda, the labialization of [-n] to [-m] for *san1* is expected to progress at a relatively faster pace since one of its occurring contexts, *san1-man4* 'news', is a frequent item. Consequently, this coda labialization, if found to eventually spread to the whole lexicon, progresses at different paces across lexical items, supporting the notion of lexical diffusion[48].

The inclusion of a large pool of characters in Experiment 7 gives us an opportunity to study the query. In the lexical dimension, if sound change is regular, as proposed by the Neo-grammarians, the rate of confusion, and thus the degree of tone merger, should be approximately the same. Judging from our data, however, we found that characters of different segmental compositions progress on the tone merger pathway at significantly different paces (e.g. confusion probability of 10.22% for the triplet (變, 便, 辯) vs. 58.40% for (慰, 調, 衛)), depending on various factors (sonority

---

[48] The same argument holds also for the tone confusion between Mandarin tone-2 and tone-3 in front of another tone-3, as reported in Wang & Li (1967). Should there be a Mandarin tone-2 tone-3 merger initiated by such tone-3 sandhi rule, those tone-3 characters frequently occurring in front of another tone-3 should be the most susceptible to such change, progressing much faster in the tone merger evolutionary pathway.

and lexical frequency analyses in Section 6.3 were our two attempts to unveil such factors). In the population dimension, our results clearly show dependence of tone merger stage on subject age. The younger the subjects, the further stage of tone-merger they exhibited (in terms of tone confusion), even though they were members of the same speech community from a synchronic survey. Given the evidence of variations in both lexical and population dimensions, our questionnaire survey follows Bauer (1983)[49] as another identified case of lexical diffusion in Hong Kong Cantonese.

To be fair, it must be noted that, for every character in the lexicon, there are possibly more than one ongoing sound change. In case some of them proceeds in the opposite direction (e.g. tonal split in the course of tone merger), the differential tone merger schedule observed in Experiment 7 may result. Extending that logic, the superficial tone confusion rates can be interpreted as resulting from numerous simultaneous sound changes proceeding in a regular manner as predicted by the Neo-grammarian hypothesis. Thus, to rigorously pursue an answer to this possibility, factors other than the one under investigation (tone category in our case), though difficult, must be isolated.

Apart from the aforementioned variations across age groups in the questionnaire survey, heterogeneity exists even among subjects with similar linguistic (native Cantonese speakers) and education (all receiving education in Hong Kong, and mostly university degree holders) background in other phonetic experiments. For example, in Experiment 1, one of the subjects' (Subject 2) coda consonant production was quite stable, without showing obvious signs of rate-dependent regressive assimilation. The same speaker, when participating in tone production experiment (Experiment 5), also showed slight, if not none at all, F_0 peak shift under high speaking rate. This may be speaking style preference: more emphasis on accuracy rather than naturalness of speech production.

---

[49] The paper surveyed 75 subjects of various ages on two sound changes in progress in Hong Kong Cantonese: (1) syllabic velar nasal /ŋ/ becoming syllabic /m/, and (2) the loss of labialization of /k^w-/ to become /k-/ before low back round vowel /ɔ/.

The picture is further complicated when we consider the listeners' responses: perceptual confusions of coda and onset consonants varied a lot across stimuli from different speakers (e.g. 89 for stimuli from Subject 4 in Experiment 1 vs. 148 for that from Subject 3) [50]. Grouping the perceptual errors by listener, considerable inter-subject variation is also present: Subject 3 responded with only 66 confusions (coda + onset), while Subject 1 got 115 confused, about 74% more probable than the former. Summing the results, it is evident that successful speech communication depends on both the speaker and listener.

Considering the non-negligible cross-subject variations from experiments under tight laboratory control, variations are expected to be even more significant when one zooms out to the whole population. Hong Kong, for instance, has a well mixture of speakers with different linguistic skills (e.g. immigrants from mainland, European language speakers due to colonial history, Filipino maids). Sound change in a manner of lexical diffusion is much more probable and reasonable. A homogeneous speech community, as assumed by the Neo-grammarians, is basically non-existent.

To conclude this discussion, sound change is observed to originate from variations / heterogeneities, both across the lexicon and speakers of the same speech community. Sound change is initiated by minute communication errors resulting from such variations, and proceeds and propagates to the whole lexicon and speech community through iterations of communication, with co-operative effort from the heterogeneous speech community members.

7.3 Contextual variability

Like compound formation from elements in chemistry study, syllables are combined together, limited by phonotactic constraints, to produce utterances for speech

---

[50] Given that Subject 2 show the least variation across speaking rates in Experiment 1, stimuli from him was expected to yield the smallest number of coda confusions in Experiment 2. This, interestingly, was not the case. Instead, stimuli from Subject 4, whose speech showed obvious rate-dependent acoustic variations, led to the smallest number of coda confusions. This gives the hint that some other factors, other than the F_2 trajectory we investigated, also influence the perception of consonant POA.

communication. This is probably one of the reasons why many speech production and perception experiments took syllable as their target unit of investigation.

However, it must be noted that spoken language communication is not accomplished in a syllable-by-syllable manner, but through sound streams consisting of concatenated syllables. Such concatenation involves *non-linear* distortion to the acoustic signals. For instance, F_0 shows carryover and anticipatory effects (Xu, 1997) when tones appear successively in a speech stream, and formants are deflected according to POA setting of the neighboring consonants (Öhman, 1966).

The above examples illustrate contextual variabilities in one acoustic dimension due to another speech unit occurring nearby. There are also contextual variabilities across acoustic dimensions like *vowel height on pitch* (i.e. intrinsic F_0 of vowels (Whalen & Levitt, 1995[51])), *speaking rate on VOT* (Miller *et al.*, 1986[52]), and *consonant voicing on pitch*[53] (e.g. House & Fairbanks, 1953; Lehiste & Peterson, 1961).

These contextual variabilities must be removed before successful recovery of speakers' intended messages is possible. To counter-act it, our perceptual system makes use of phonetic context to compensate for the variabilities. To recover identity of a target speech unit, contextual acoustic cues like *phoneme* (Warren, 1970), *formant height* (Ladefoged & Broadbent, 1957), *F_0 level* (Lin & Wang, 1985), and *F_0 contour* (Wong, 2008) are all possibly utilized by the listeners under different circumstances.

It is exactly the complicated contextual variabilities arising from acoustic interaction between neighboring segments during production, and the complexities

---

[51] Whalen & Levitt presented a typological survey of 31 languages from different language families and types (i.e. tonal vs. pitch accent vs. stress). They included the three corner vowels in their survey, and the major result was that a consistent relationship between vowel-height and pitch was observed: The higher a vowel is, the higher its F_0 becomes.

[52] Essentially, when speaking rate is lower, and thus longer syllable duration, the VOT values for consonants become longer accordingly.

[53] It has been widely reported that, consonant pairs (e.g. /p/ vs. /b/, /k/ vs. /g/) always have the voiced one associated with a lower F_0.

involved in de-contextualization during perception which lead to un-negligible confusions, initiating sound changes.

Our Experiment 1 and 2 illustrated how consecutive contextualized consonants influence each other under high speaking rate, and how the resultant acoustic variations become so great that a categorical shift of the perceived consonant happens.

Our experiments dealt with continuous speech. Contrasting them with those taking isolated syllables as the target of investigation, the stimuli elicited in the production experiments contained more variations (other than the manipulated ones). Similarly, our corpus analyses involved spontaneous speech samples, with even larger amount of variations (e.g. emotion, speaker gender, etc.). Being more deviant from the ceteris paribus principle, other factors might play a more-than-expected influential role in leading to the investigated sound changes. However, since continuous speech is the most frequent form of speech we encounter, it must not be under-estimated when study the underlying mechanism of human speech communication. It is always a pursuit to maintain a balance between experimental control and general applicability of results.

7.4 Further work

In this dissertation, although effort has been put into demonstrating, from multiple perspectives, a pathway for diachronic assimilation process, the following follow-ups are expected to further improve our understanding in our research problem.

In Experiment 1, for simplicity, we investigated only the rate-induced variations of only F_2 trajectories. A full spectrogram, however, contains many more than that. F_1, for instance, normally has a higher amplitude (and is thus more audible), due to the -6dB/octave spectral filtering of the oral cavity (Clark & Yallop, 1995, p246-247). Besides, the first few formants are argued to conspire to correspond to consonant specification, as suggested by the acoustic loci theory (Delattre *et al.*, 1955). F_3, being weaker in energy level, plays a significant role in indicating lip rounding. If we are to generalize the finding of regressive segmental feature (currently only *place* feature in stop consonants) assimilation under high speaking rate to other segmental features (e.g. *frication*, *voicing*, *nasality*) in other consonant types, a fuller spectrum of acoustic parameters have to be considered.

In the original design of Experiment 2, natural samples elicited from Experiment 1 were used for a perceptual task. This was primarily to ensure that all the acoustic cues co-varying with F_2, our target cue in Experiment 1, were preserved such that a natural communication scenario could be re-generated as far as possible in the laboratory. Listeners' failure in stop identification given that abundance of cues solidly confirmed speaking rate as one of the underlying factor. If more quantitative details, like the F_2 drift threshold for perceptual confusion, are required, we probably have to resort to synthetic stimuli[54], which remove cross-cue co-variances, allowing precise control of individual acoustic parameters.

Experiment 3 attempted to obtain empirical support for results from the previous two phonetic experiments, by studying a large-scale transcribed spoken corpus. While it currently focused only on onset / coda variation and assimilative influence from their neighboring segments, we also noticed other patterns like *n-/l-* merger, *ŋ-* /zero initial alternation (Zee, 1999), alveolarization of coda consonants (Bauer, 1979), de-labialization of rounded initials (e.g. from k^w- to *k-*), etc., during data analyses. Further in-depth study of their context of occurrence, as well as frequency distribution may give new insights on the mechanism of sound change.

The preliminary modeling attempt in Experiment 4 was made to give meaning to the onset / coda frequency distribution calculated from a spoken corpus. More sophisticated computer modeling approaches, like those outlined in Wang *et al.* (2004) and Niyogi (2006), should enable us to obtain more detailed parameters of the evolutionary trajectories like the *duration of change*, the *onset of change*, and the *completion time*, which are hardly determined otherwise.

Results from Experiment 5 indicated a failure of our speech production organs to properly synchronize between tonal and segmental components under rapid speech. Apparently, this is due to the physiological limits as shown by previous literature on maximum performance for speech production (Kent *et al.*, 1987; Xu & Sun, 2002). An alternative hypothesis (Wang, personal communication) suggested that such

---

[54] There is always a trade-off between naturalness and precision of acoustic parameter control in experimentation. However, given the results of Liberman *et al.* (1954, 1957), the manipulation of only F_2 is justified.

de-synchronization may have its root in the vastly different length of the nerves innervating the corresponding speech organs: the recurrent laryngeal nerve[55] (RLN) runs a lot longer, and thus take longer time for neural signals, to innervate the larynx for pitch control than those for segmental feature control (e.g. tongue, mandible, lip, etc.)[56]. To verify this hypothesis, electromyographic (EMG) techniques may be required to directly tap the relative timing of control signal arrival.

Experiment 6 aimed at eliciting perceptual consequences of the F_0-peak delay caused by high speaking rate. The rather disappointing results indicated that, unlike the case of segmental place feature, the numerous co-varying cues were suspected to help listeners successfully recover the intended tones. Consequently, the query of the underlying mechanism for rightward tone spreading remains unsolved, awaiting follow-up investigations. On the other hand, the set of acoustic cues co-varying with F_0, if uncovered, should be useful in applications like automatic pitch generation, tone recognition, etc.

Compared with previous phonetic experiments, the questionnaire survey (Experiment 7) on tone merger guarantees a higher general applicability by considering a much larger sample pool. Yet, a more even frequency distribution of subject age should minimize effect of outliers, and thus facilitate more representative statistical analysis results. Besides, tone has been shown to surface with considerable variations depending on the surrounding tonal context (Wong, 2006b, 2007). Such context-dependent variations certainly enhance or reduce tonal confusion, and thus rate of tone change. Careful examination into the most frequently occurring context of the characters used in our survey may give some hints to the mechanism underlying

---

[55] The nerve is prefixed as 'recurrent' because its takes a circuitous route to first descend into the thorax before reaching the neck. This rather clumsy configuration probably comes from biological evolution of human.

[56] A complication here is that nerve length is not the sole determining factor for nerve conduction speed. Conduction speed, for instance, is in general also proportional to *nerve fiber thickness* (see discussions in Walker, 1994). A good example is that, the left RLN is longer than the right one (the two pass around the aorta in the chest and the subclavian artery respectively), however, their signal

tone merger, rather than merely a simple statement of 'due to close resemblance in F_0 shape'.

On a global level, studies in the current dissertation only focus on observations / data at the acoustic and perception stages. Other stages like *physiological*, which imposes limits on the range of possible acoustics and perceptual acuity, and *neurological*, where a complete speech communication pathway initiates and terminates at, are left uncovered. Queries like speed limit of neural coupling between production of segmental and tonal components, and neural mechanism of perceptual confusion, can only be convincingly solved with sophisticated techniques like electromyography (EMG) and neural imaging. A fuller understanding of speech sound change is achievable with relevant research into these areas.

---

conduction latencies are claimed to be more or less the same due to the fact that the left RLN is thicker than the right, compensating for the path length difference.

BIBLIOGRAPHY

Abramson, A. S. (1972) Tonal experiments with whispered Thai. In Albert Valdman (Ed.), *Papers in linguistics and phonetics in memory of Pierre Delattre*, 31-44.

Bauer, R. S. (1979) Alveolarization in Cantonese: A case of lexical diffusion, *Journal of Chinese Linguistics*, 7(1): 132-141.

Bauer, R. S. (1983) Cantonese sound change across subgroups of the Hong Kong speech community, *Journal of Chinese Linguistics*, 11(2), 303-356.

Bauer, R., & Benedict, P. K. (1997) *Modern Cantonese phonology*, Berlin: Mouton de Gruyter.

Bauer, R. S., Cheung, K. H., & Cheung, P. M. (2003) Variation and merger of the rising tones in Hong Kong Cantonese, *Language Variation and Change*, 15, 211-225.

Boersma, P., & Weenink, D. (2005) Praat: Doing phonetics by computer (version 4.3.27) [Computer program]. Retrieved on 7th October, 2005 from http://www.praat.org/.

Burquest, D. A., & Payne, D. L. (1993) *Phonological analysis: A functional approach*, Dallas, TX: Summer Institute of Linguistics.

Chang, C. Y. (2003) *Intonation in Cantonese*. LINCOM Studies in Asian Linguistics, vol. 49, Munich: Lincom Europa.

Chao, Y. R. (1930) A system of tone letters, *Le Maitre Phonetique*, 30, 24-27.

Chao, Y. R. (1931) 反切語八種 "Eight varieties of secret language based on the principle of fan-qie", *Bulletin of the Institute of History and Philology*, 2, 312-354.

Chao, Y. R. (1933) Tone and intonation in Chinese, *Bulletin of the Institute of History and Philology*, 4, 121-134.

Chao, Y. R. (1947) *Cantonese primer*, Cambridge, MA: Harvard University Press.

Chen, M. Y. (2000) *Tone Sandhi: Patterns across Chinese Dialects*, UK: Cambridge University Press.

Chen, M., Wang, W. S-Y. (1975) Sound change: Actuation and implementation. *Language*, 51: 255-281.

Chuenwattanapranithi, S., Xu, Y., Thipakorn, B. & Maneewongvatana, S. (2006) Expressing anger and joy with the size code. In *Proceedings of International Conference on Speech Prosody 2006*, Dresden, Germany.

Clark, J., & Yallop, C. (1995) *An introduction to phonetics and phonology*. Blackwell.

Clements, G. N. (1990) The role of the sonority cycle in core syllabification, in *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*, John Kingston, and Mary E. Beckman (eds.), 283-333. CUP.

Delattre, P. C., Liberman, A. M., & Cooper, F. S. (1955) Acoustic loci and transitional cues for consonants, *Journal of the Acoustical Society of America*, 27(4), 769-773.

Dobzhansky, T. (1973) Nothing in biology makes sense except in the light of evolution, *The American Biology Teacher*, 35, 125-129.

Fok, C. Y.-Y. (1974) *A perceptual study of tones in Cantonese*. Center of Asian Studies. University of Hong Kong.

Fok, C. Y. Y. (1979) The frequency of occurrence of speech sounds and tones in Cantonese. In L. Robert (Ed.), *Hong Kong Language Papers*. Hong Kong: Hong Kong University Press: 150-157.

Fujimura, O., Macchi, M. J., & Streeter, L. A. (1978) Perception of stop consonants with conflicting transitional cues: A cross-linguistic study, *Language and Speech*, 21, 337-346.

Gandour, J. T., Potisuk, S., & Dechongkit, S. (1994) Tonal coarticulation in Thai, *Journal of Phonetics*, 22, 477-492.

Garrett, K. L., & Healey, E. C. (1987) An acoustic analysis of fluctuations in the voices of normal adult speakers across three times of day, *Journal of the Acoustical Society of America*, 82, 58-62.

Gauchat, L. (1905) L'unite phonetique dans le patois d'une commune. In *Aus Romanischen Sprachen und Literaturen: Feschrift Heinrich Mort*. Halle: Max Niemyer. 175-232.

Gay, T. (1978) Effect of speaking rate on vowel formant movements, *Journal of the Acoustical Society of America*, 63(1), 223-230.

Greenberg, J. H. (1965) Some generalizations concerning initial and final consonant clusters, *Linguistics*, 18, 5-34.

Han, M. S. & Kim, K. O. (1974) Phonetic variation of Vietnamese tones in disyllabic utterances, *Journal of Phonetics*, 2, 223-232.

Hashimoto, O. K. Y. (1972) *Studies in Yue Dialects 1: Phonology of Cantonese*, Cambridge, England: Cambridge University Press.

Hermann, E. (1929) Lautveranderungen in der individualsprache einer Mundart. *Nachrichten der Gesellschaft der Wissenschaften zu Gottingen, philosophisch-historische Klasse* 11: 195-214.

Hill, A. A. (1955) Consonant assimilation and juncture in English: A hypothesis, *Language*, 31(4), 533-534.

Hombert, J. M. (1977) Development of tones from vowel height? *Journal of Phonetics*, 5, 9-16.

Hombert, J. M. (1978) Consonant types, vowel quality and tone, in Fromkin V. (Ed.), *Tone: A Linguistic Survey*, New York, Academic Press, 77-111.

Hombert, J. M., Ohala, J. J., & Ewan, W. G. (1979) Phonetic explanations for the development of tones, *Language*, 55, 37-58.

House, A. S., & Fairbanks, G. (1953) The influence of consonant environment upon the secondary acoustical characteristics of vowels, *Journal of the Acoustical Society of America*, 25, 105-113.

Hyman, L. M. & Schuh, R. G. (1974) Universals of tone rules: evidence from West Africa, *Linguistic Inquiry*, 5, 81-115.

Jakobson, R. (1966) Implications of language universals for linguistics, in J. H. Greenberg (Ed.), *Universals of language*, 263-278. MA: M.I.T. Press.

Jespersen, O. (1904) *Lehrbuch der Phonetik*, Leipzig and Berlin.

Jun, J. (2004) Place assimilation. In B. Hayes, R. Kirchner, & D. Steriade (Eds.) *Phonetically Based Phonology*. Cambridge University Press.

Katamba, F. (1989) *An introduction to phonology*, Longmans, London.

Kelso, J. A. S. (1984) Phase transitions and critical behavior in human bimanual coordination, *American Journal of Physiology: Regulatory, Integrative and Comparative Physiology*, 246, R1000-R1004.

Kennedy, G. A. (1953) Two tone patterns in Tangsic, *Language*, 29(3), 367-373.

Kent, R. G. (1936) Assimilation and dissimilation, *Language*, 12(4), 245-258.

Kent, R. D., Kent, J. F., & Rosenbek, J. C. (1987) Maximum performance tests of speech production, *Journal of Speech and Hearing Disorders*, 52, 367-387.

Krakow, R. A. (1999) Physiological organization of syllables: a review, *Journal of Phonetics*, 27, 23-54.

Krámský, J. (1959) A quantitative typology of languages, *Language and Speech*, 2, 72-85.

Labov, W. (1966) *The social stratification of English in New York City*, Washington, DC: Center for Applied Linguistics.

Labov, W. (1974) On the use of the present to explain the past, in L. Heilmann (Ed.), *Proceedings of the Eleventh International Congress of Linguistics*, Bologna: Mulino.

Ladefoged, P., & Broadbent, D. E. (1957) Information conveyed by vowels, *Journal of the Acoustical Society of America*, 29, 98-104.

Laver, J. (2000) Linguistic phonetics. In Aronoff, M. & Rees-Miller, J. (Eds.) *The Handbook of Linguistics*. Blackwell. 150-179.

Leather, J. (1983) Speaker normalization in perception of lexical tone, *Journal of Phonetics*, 11, 373–382.

Lehiste, I., & Peterson, G. E. (1961) Some basic considerations in the analysis of intonation, *Journal of the Acoustical Society of America*, 33(4), 419-425.

Leung, M. T., & Law, S. P. (2001) HKCAC: The Hong Kong Cantonese adult language corpus. *International Journal of Corpus Linguistics*, 6: 305-325.

Leung, M. T., Law, S. P., & Fung, S. Y. (2004) Type and token frequencies of phonological units in Hong Kong Cantonese, *Behavior Research Methods, Instruments, & Computers*, 36(3), 500-505.

Liberman, A. M., Delattre, P. C., Cooper, F. S., & Gerstman, L. J. (1954) The role of consonant-vowel transitions in the perception of the stop and nasal consonants, *Psychological Monographs*, 68(8).

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957) The discrimination of speech sounds within and across phoneme boundaries, *Journal of Experimental Psychology*, 54, 358-368.

Lindblom, B. (1990) Explaining phonetic variation: A sketch of the H&H theory. In Hardcastle, W. J. and Marchal, A. (Eds.), *Speech Production and Speech Modeling*. The Netherlands: Kluwer Academic, 403-439.

Lin, M. (1988) 普通話聲調的聲學特性和知覺徵兆, *中國語文*, 204, 182-193.

Lin, T. & Wang, W. S-Y. (1985) 聲調感知問題, *中國語言學報*, 2, 59-69.

Liu, J. (2001) *Tonal behavior in some tone languages*. Ph.D. dissertation. City University of Hong Kong.

Lisker, L. (1986) "Voicing" in English: A catalogue of acoustic features signaling /b/ versus /p/ in trochees, *Language and Speech*, 29(1), 3-11.

LSHK (1997) *Hong Kong Jyutping Characters Table* (*粵語拼音字表*). Linguistic Society of Hong Kong Press (香港語言學會出版).

Maddieson I. (1978) Universals of tone. In J. H., Greenberg (Ed.), *Universals of Human Language, vol. 2, Phonology*, 335-365. Stanford: Stanford University Press.

Miller, J. L., Green, K. P. & Reeves, A. (1986) Speaking rate and segments: A look at the relation between speech production and speech perception for the voicing contrast, *Phonetica*, 43, 106-115.

Mok, P. K., & Wong, P. (2010) Perception of the merging tones in Hong Kong Cantonese: preliminary data on monosyllables, in *Proceedings of Speech Prosody 2010*, 100916:1-4. Chicago.

Moore, C. B. & Jongman, A. (1997) Speaker normalization in the perception of Mandarin Chinese tones, *Journal of the Acoustical Society of America*, 102, 1864-1877.

Niyogi, P. (2006) *The computational nature of language learning and evolution.* Cambridge, MA: MIT Press.

Ohala, J. J. (1974) Experimental historical phonology, In: Anderson, J. M., & Jones, C. (Eds.), *Historical linguistics II. Theory and description in phonology*. [Proc. of the 1st International Conference on Historical Linguistics. Edinburgh, 2 - 7 Sept. 1973.] Amsterdam: North Holland, 353 - 389.

Ohala, J. J. (1981) The listener as a source of sound change. In C. S. Masek, R. A. Hendrick, & M. F. Miller (Eds.), *Papers from the Parasession on Language and Behavior*, 178-203. Chicago: Chicago Linguistic Society.

Ohala, J. J. (1990) The phonetics and phonology of aspects of assimilation. In J. Kingston & M. Beckman (Eds.), *Papers in Laboratory Phonology I: Between the grammar and the physics of speech*. Cambridge: Cambridge University Press, 258-275.

Öhman, S. E. G. (1966) Coarticulation in VCV utterances: Spectrographic measurements, *Journal of the Acoustical Society of America*, 39(1), 151-168.

Peng, G. (2006) Temporal and tonal aspects of Chinese syllables: A corpus-based comparative study of Mandarin and Cantonese, *Journal of Chinese Linguistics*, 34(1), 134-154.

Peng, G., & Wang, W. S-Y. (2005) Tone recognition of continuous Cantonese speech based on support vector machines, *Speech Communication*, 45, 49-62.

Peterson, G. E., & Barney, H. L. (1952) Control methods used in a study of the vowels, *Journal of the Acoustical Society of America*, 24(2), 175-184.

Potisuk, S., Gandour, J., & Harper, M. P. (1997) Contextual variations in trisyllabic sequences of Thai tones, *Phonetica*, 54(1), 22-42.

Sadler, J. D. (1973) Assimilation and dissimilation, *The Classical Journal*, 68(3), 267-271.

Shen, X. (1990) Tonal coarticulation in Mandarin, *Journal of Phonetics*, 18, 281–285.

Shen, Z. (1997) *Exploring the Dynamic Aspect of Sound Change*. Journal of Chinese Linguistics Monograph Series No. 11.

So, L. K. H. (1996) Tonal changes in Hong Kong Cantonese. *Current Issues in Language and Society*, 3(2), 186-189.

So, L. K. H., & Varley, R. (1991) *Cantonese Lexical Comprehension Test*. Department of Speech and Hearing Sciences: University of Hong Kong.

Steriade, D. (2001) Directional asymmetries in place assimilation: a perceptual account. In Hume, E. & Johnson, K. (eds.) *The role of speech perception in phonology*. San Diego: Academic Press, 219-250.

Trask, R. L. (1998) *Historical linguistics*. Hodder Arnold.

Tse, K.-P. J. (1979) Mang Kung Wa: A game language of Cantonese, *Studies in English Literature and Linguistics*, 4, 97-106.

Tuller, B., & Kelso, J. A. S. (1990) Phase transitions in speech production and their perceptual consequences. In M. Jeannerod (Ed.), *Attention and Performance XIII*, 429-452. Hillsdale, NJ: Erlbaum.

Vance, T. J. (1976) An experimental investigation of tone and intonation in Cantonese, *Phonetica*, 33, 368-392.

Walker, S. F. (1994) The possible role of asymmetric laryngeal innervation in language lateralization: Points for and against, *Brain and Language*, 46, 482-489.

Wang, W. S.-Y. (1959) Transition and release as perceptual cues for final plosives, *Journal of Speech and Hearing Research*, 2, 66-73.

Wang, W. S.-Y. (1967) Phonological features of tone, *International Journal of American Linguistics*, 33(2), 93-105.

Wang, W. S.-Y. (1969) Competing changes as a cause of residue, *Language*, 45, 9-25.

Wang, W. S.-Y. (1972) The many uses of F_0. In Valdman, A. (Ed.), *Papers in Linguistics and Phonetics to the Memory of Pierre Delattre*, Mouton, 487-503.

Wang, W. S.-Y. (1974) How and why do we study the sounds of speech? In Mitchell, L. (Ed.) *Computers in the Humanities*. Edinburgh: Edinburgh University Press, 39-53.

Wang, W. S.-Y., Ke, J. Y., & Minett, J. W. (2004) Computational studies of language evolution. In Huang, C.R. & Lenders, W. (Eds.), *Computational linguistics and beyond*. Academia Sinica: Institute of Linguistics, 65-106.

Wang, W. S.-Y., & Li, K. P. (1967) Tone 3 in Pekinese, *Journal of Speech and Hearing Research*, 10(3), 629-636.

Warren, R. M. (1970) Perceptual restoration of missing speech sounds, *Science* 167, 392-393.

Whalen, D. H., & Levitt, A. G. (1995) The universality of intrinsic F_0 of vowels, *Journal of Phonetics*, 23, 349-366.

Wong, P. C. M. (1999) The effect of downdrift in the production and perception of Cantonese level tone. In *Proceedings of the XIVth International Congress of Phonetic Sciences*, San Francisco, 3, 2395-2398.

Wong, P. C. M. & Diehl, R. L. (2003) Perceptual normalization of inter- and intra-talker variation in Cantonese level tones, *Journal of Speech, Language, and Hearing Research*, 46, 413-421.

Wong, Y. W. (2006a) Temporal distribution of tonal information in continuous Cantonese speech. In *Proceedings of Second International Symposium on Tonal Aspects of Languages*, La Rochelle, France.

Wong, Y. W. (2006b) Contextual tonal variations and pitch targets in Cantonese. In *Proceedings of International Conference on Speech Prosody 2006*, Dresden, Germany.

Wong, Y. W. (2007) *Production and Perception of Tones in Cantonese Continuous Speech*. Unpublished M.Phil thesis. Chinese University of Hong Kong.

Wong, Y. W. (2008) The role of carryover tonal variations in Cantonese tone perception. Paper presented at *the 5th Postgraduate Research Forum on Linguistics*, Hong Kong, China.

Wong, Y. W. (2009a) 語速與音段特徵同化方向的關係. Paper presented at *Conference in Evolutionary Linguistics I*, Guangzhou, 2009.

Wong, Y. W. (2009b) Report: Conference in Evolutionary Linguistics I, Guangzhou, 2009. *Journal of Chinese Linguistics*, 37(2), 386-396.

Wong, Y. W. (2010) Variations of onset and coda consonants in Hong Kong Cantonese: A spoken corpus study. In Proceedings of *Conference in Evolutionary Linguistics II*, Tianjin, 2010.

Xu, Y. (1997) Contextual tonal variations in Mandarin, *Journal of Phonetics*, 25, 61-83.

Xu, Y. (1999). Effects of tone and focus on the formation and alignment of F_0 contours, *Journal of Phonetics*, 27, 55-105.

Xu, Y. (2001) Fundamental frequency peak delay in Mandarin. *Phonetica*, 58, 26-52.

Xu, Y. (2010) In defense of lab speech, *Journal of Phonetics*, 38, 329-336.

Xu, Y., & Sun, X. (2002) Maximum speed of pitch change and how it may relate to speech, *Journal of the Acoustical Society of America*, 111(3), 1399-1413.

Xu, Y. & Wang, Q. E. (2001) Pitch targets and their realization: Evidence from Mandarin Chinese, *Speech Communication*, 33, 319-337.

Yue-Hashimoto, A. O. (1987) Tone sandhi across Chinese dialects. In Chinese Language Society of Hong Kong (Ed.), *Wang Li Memorial Volumes: English Volume*, 445-474. Hong Kong: Joint Publishing Co.

Zee, E. (1985) Sound change in syllable final nasal consonants in Chinese. *Journal of Chinese Linguistics*, 13: 291-330.

Zee, E. (1991) Chinese (Hong Kong Cantonese), *Journal of the International Phonetic Association*, 21(1), 46-48.

Zee, E. (1999) Change and variation in the syllable-initial and syllable-final consonants in Hong Kong Cantonese. *Journal of Chinese Linguistics*, 27(1), 120-167.

Zee, E. (2002) The effect of speech rate on the temporal organization of syllable production in Cantonese. In *Proceedings of International Conference on Speech Prosody 2002*.

Zhang, J. (2005) Contour restrictions and faithful alignment in Chinese tone sandhi systems. Paper presented at *the 3rd Workshop on Theoretical East Asian Linguistics*.

APPENDIX

廣東話字音辨認測試

以下有一系列的三字組，它們的讀音有些相同，另一些則不同。請根據你的 **廣東話** 讀音，判斷並圈出讀音不同的那一個字。如果你認爲三個字的讀音一樣，請在該一字組右側「一樣」方格內劃號表示。如果你不肯定題目內某字的讀音或不肯定答案，請於「不肯定」格內劃號表示。如果你明白以上描述，現在就請開始測試。

				一樣	不肯定					一樣	不肯定
1.	軟	丸	院	☐	☐	19.	椅	議	耳	☐	☐
2.	史	屎	市	☐	☐	20.	幣	閉	陛	☐	☐
3.	富	負	庫	☐	☐	21.	社	寫	捨	☐	☐
4.	來	朱	豬	☐	☐	22.	討	禱	肚	☐	☐
5.	便	辯	變	☐	☐	23.	印	刃	孕	☐	☐
6.	紙	指	只	☐	☐	24.	忌	寄	技	☐	☐
7.	逝	勢	細	☐	☐	25.	啞	雅	瓦	☐	☐
8.	雨	瘀	與	☐	☐	26.	善	線	腳	☐	☐
9.	刊	罕	捍	☐	☐	27.	秘	避	鼻	☐	☐
10.	舞	帽	武	☐	☐	28.	吊	釣	掉	☐	☐
11.	真	珍	陳	☐	☐	29.	膚	佬	魯	☐	☐
12.	受	秀	獸	☐	☐	30.	粉	奮	憤	☐	☐
13.	斧	婦	苦	☐	☐	31.	願	怨	縣	☐	☐
14.	洞	凍	動	☐	☐	32.	他	流	留	☐	☐
15.	齒	似	始	☐	☐	33.	穩	允	尹	☐	☐
16.	謂	慰	衛	☐	☐	34.	委	偉	毀	☐	☐
17.	陣	鎮	震	☐	☐	35.	有	誘	柚	☐	☐
18.	謝	借	蕉	☐	☐	36.	卸	瀉	射	☐	☐

測試完成，謝謝你的幫助！

（姓名：_____ 年齡：_____ 性別：_____ 居港年數：_____ 教育程度：_____）

Question and answer sheet used in the questionnaire survey reported in Chapter 6.

```
void trim(float pitch[], ...)
{
    for each pitch point pitch[i]
    {
        pitch_diff1 = pitch[i] - pitch[i-1];
        pitch_diff2 = pitch[i] - pitch[i+1];

        if(an upward/downward jump is found)
        {
            pitch[i] = linearly-interpolated value from pitch[i-1] & pitch[i+1];
        }
    }
}
```

Pseudo code for the trimming algorithm used in Experiment 5.