Mechanism of Extreme Phonetic Reduction:

Evidence from Taiwan Mandarin

Chierh Cheng

A thesis submitted in fulfilment of requirements for the degree of

Doctor of Philosophy

to

Department of Speech, Hearing and Phonetic Sciences

Division of Psychology and Language Sciences

University College London (UCL)

2012

Declaration

I, Chierh Cheng, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Chierh Cheng

鄭 齊兒

“If, as has been said, the human brain is the most complicated thing in the universe, then language, as the principal expression of its intelligence, is the most wonderful thing in the universe.”

Ian Bruton-Simmonds

Abstract

Extreme reduction refers to the phenomenon where intervocalic consonants are so severely reduced that two or more adjacent syllables appear to be merged into one. Such severe reduction is often considered a characteristic of natural speech and to be closely related to factors including lexical frequency, information load, social context and speaking style. This thesis takes a novel approach to investigating this phenomenon by testing the *time pressure account of phonetic reduction*, according to which *time pressure is the direct cause of extreme reduction*. The investigation was done with data from Taiwan Mandarin, a language where extreme reduction (referred to as *contraction*) has been reported to frequently occur.

Three studies were conducted to test the main hypothesis. In Study 1, native Taiwan Mandarin speakers produced sentences containing nonsense disyllabic words with varying phonetic structures at differing speech rates. Spectral analysis showed that extreme reduction occurred frequently in nonsense words produced under high time pressure. In Study 2a, further examination of formant peak velocity as a function of formant movement amplitude in experimental data suggested that articulatory effort was not decreased during reduction, but in fact likely to be increased. Study 2b examined high frequency words from three spontaneous speech corpora for reduction variations. Results demonstrate that patterns of reduction in high frequency words in spontaneous speech (Study 2b) were similar to those in nonsense words spoken under experimental conditions (Study 2a).

Study 3 investigated tonal reduction with varying tonal contexts and found that tonal reduction can also be explained in terms of time pressure. Analysis of F_0 trajectories demonstrates that speakers attempt to reach the original underlying tonal targets even in the case of extreme reduction and that there was no weakening of articulatory effort despite the severe reduction. To further test the main hypothesis, two computational modelling experiments were conducted. The first applied the quantitative Target Approximation model (qTA) for tone and intonation and the second applied the Functional Linear Model (FLM). Results showed that severely reduced F_0 trajectories in tone dyads can be regenerated to a high accuracy by qTA using generalized canonical tonal targets with only the syllable duration modified. Additionally, it was shown that using FLM and adjusting duration alone can give a fairly good representation of contracted F_0 trajectory shapes.

In summary, results suggest that target undershoot under time pressure is likely to be the direct mechanism of extreme reduction, and factors that have been commonly associated with reduction in previous research very likely have an impact on duration, which in turn determines the degree of target attainment through the time pressure mechanism.

Acknowledgements

I would first and foremost like to express my heartfelt gratitude to my supervisor, Dr. Yi Xu, for all the help and guidance he has given me throughout my time at UCL. It has been a pleasure to work with him and learn from his expertise, and I thank him for giving me considerable academic freedom, which has made my time at UCL so enjoyable.

I owe a big thank you to all of the research, teaching and administrative staff at UCL who have been so helpful throughout my PhD. In particular, I would like to thank Dr. Mark Huckvale, Dr. Michael Ashby, Steve Nevard and my supervisor for giving me invaluable opportunities to learn from them as a teaching assistant. I would like to thank everyone who has taken part in the Speech Science Forum for insightful conversations and input, which have constantly inspired my own research and provided me with an opportunity to appreciate quality research. I would also like to thank my fellow PhD students who have created such a pleasant environment at UCL, and in particular the students with whom I carried out my PhD. Thanks to Nada Al-Sari, Wafa'a Al-Shangiti, Sam Evans, Sophie Gates, Sonia Granlund, Kota Hattori, Young Shin Kim, Jeong-Sug Kyong, Angelos Lengeris, Katharine Mair, Georgina Oliver, Melanie Pinet, Yasna Pereira Reyes, Csaba Redey-Nagy, Dong-Jin Shin, Yasuaki Shinohara, Wing Li Wu, Mark Wibrow and Kayoko Yanagisawa, who are or have been Room 326ers and who have continuously created a lovely atmosphere by sharing together ups and downs and a bit of laughter. I would also like to thank my internal examiner, Dr. Jill House, and external examiner, Dr. Fred Cummins (University College Dublin), for their insightful suggestions and an enjoyable discussion during the viva.

I am truly indebted and thankful to my department SHaPS (in particular to Dr. Andrew Faulkner and Prof. Valerie Hazan) and the UCL Graduate School for their financial assistance which has allowed me to carry out my study at UCL and to also attend some prestigious conferences. I, of course, owe a great deal of gratitude to my funding body, the Ministry of Education in Taiwan, who have funded me throughout my third year and during my writing-up period at UCL. I would also like to express my gratitude to Dr. Shu-Chun Tseng in Academia Sinica for her generosity in providing me with corpus data and research guidance

during my study leave in Taiwan. I would also like to thank Dr. Michele Gubian for teaching me how to think like an engineer, Dr. Juraj Simko and Dr. Shu-Chun Tseng for their interest in reading my thesis and various anonymous reviewers for giving me valuable feedback during my writing-up period. Similarly, I would like to express my gratitude to my undergraduate and graduate advisers in Taiwan, Prof. Ho-hsien Pan, Prof. Tse, Kwock-Ping John and Prof. Yuchau E. Hsiao for their constantly supporting and encouraging me to carry out my study in the U.K.

I also wish to warmly acknowledge the support of all of my friends, and in particular I would like to thank friends at Tzu Chi UK for great charity events during my PhD, and my tutees and their families for all of the entertainment in and around London.

I have a giant thank you to say to my fiancé, Rhodri Nelson, and his family for their patience, kindness and warmth (and for all the lovely Welsh meals and time we spent together over the past couple of Christmas vacations). Diolch yn fawr!

最後，我要感謝關心我的家人們，爸爸、媽媽、德芳姐、秉中和摯友明良。謝謝你們的一直以來的陪伴及鼓勵，我愛你們。^_^。

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

The variability of speech is one of the most challenging aspects of speech science (Keating, 1997; Perkell and Klatt, 1986) and a common form in which this variability manifests itself is phonetic reduction (Engstrand and Krull, 2001; Johnson, 2004; Kohler, 1990 and 1998). Various accounts have been proposed to explain the discrepancy between canonical and reduced forms. Early accounts regarding the sources of phonetic reduction can be dated back to Karlgren (1962) who proposed, based on Information Theory (Shannon, 1948), that reduction should be understood in terms of the transmission rate of content underlying the phonemic message: "…there is an equilibrium between information value on one hand, and duration along with similar qualities of the realization on the other" (Karlgren, 1962, p. 676). Karlgren did not explain what he meant by other "similar qualities of the realization", but focused mainly on duration and

postulated that reduction associated with more rapid speech is a measure of coding efficiency. Based on auditory transcriptions and a visual inspection of acoustic signals in six languages, Barry and Andreeva (2001) argued that variation in the time and effort invested in any given part of an utterance serves as a means to support the relative weight of elements within the information structure. They suggest that items containing a greater information load carry more weight and are therefore assigned greater effort and a longer duration in order to achieve better target attainment. Conversely, items with weak information loads are assigned a comparatively measly duration and effort, resulting in significant undershoot of these items.

Also focusing on temporal change and phonetic realisation, Lindblom (1963) observed the interplay between duration and formant realisation in a CVC structure and proposed a *duration-dependent undershoot* model: When speech rate is increased and vowel duration shortened, the extent of movement towards the vowel target is reduced. Lindblom attributed such reduction to articulatory constraints on the limit of the maximum speed of articulatory movement. In this model, Lindblom introduced the notion of an acoustic target being approached asymptotically and proposed that the determinants of undershoot are duration and locus-target distance (i.e. the displacement required to achieve a desired target). Lindblom's model was, however, questioned in subsequent studies (Engstrand, 1988; Fourakis, 1991; Gay, 1978; van Son and Pols, 1990 and 1992) which failed to find significant duration-dependent formant displacement effects. As a response to the criticisms, Moon and Lindblom (1994) showed that in an English /w _ l/ frame, where the locus-target distance is large, duration dependency could clearly

be observed. On the other hand, they also observed that the duration-dependency of formant shift was more limited in clear speech. Based on this observation, they suggested that articulatory effort could reduce duration-dependency. In their revision of Lindblom's (1963) original model, formant undershoot becomes a function of vowel duration, locus-target distance and rate of formant frequency change (which is used as an indicator of 'articulatory effort'). Lindblom (1990) further hypothesized that speakers can adapt to different speaking situations and choose appropriate production strategies (i.e., by changing kinematic parameters) to avoid or to allow reduction. This is known as the *Hyper- and Hypo-articulation (H&H) theory*, which characterizes the trade-offs between articulatory economy and perceptual comprehension. Importantly, H&H theory hypothesizes that the mechanics of speech production are similar to those of non-speech motor behaviours, which are constrained by the *principle of economy of effort* (Nelson, 1983).

H&H theory has influenced a number of recent studies regarding speech communication and is among the most dominant theories of phonetic reduction. As an example, studies of reduction based on consistent communicative contexts, such as lexical frequency effect, usually show both temporal and spectral reduction in high-frequency items (Aylett and Turk, 2006; Fosler-Lussier and Morgan, 1999; Myers and Li, 2009). It has been suggested that information regarding language redundancy, either because of context or word frequency, can influence the amount of effort exerted in articulation (Pluymakers et al., 2005).

This interpretation would lead to the prediction that, if a low probability word (supposedly initially allocated a comparably high amount of effort and thus a clear pronunciation) were to be pronounced at a fast rate (owing to a certain communicative function) speakers could offset this high time pressure (and potential undershoot) with an increased articulatory effort. Indeed van Son (1993:13-14) has suggested that unfamiliar or unknown lexical items such as nonsense words may lead to a speaking style that is clearer than normal. If this is the case then nonsense words, which by definition have the lowest possible frequency of occurrence, should be influenced the least by this frequency effect.

However, several perceptual studies have produced results that do not confirm the prediction that a very clear speech style (thus conceptually with more articulatory effort) can compensate for high time pressure. For example, Krause and Braida (2002) investigated alternative forms of clear speech by training professional speakers to produce clear and conversational speech at slow, normal and fast rates. The intelligibility advantage of clear speech was found at slow and normal rates. In particular, a form of clear speech was obtained at slow (approximately 0.5 second per syllable) and normal (approximately 0.25 second per syllable). However, the intelligibility advantage of clear speech was lost at the fast speech rate, that is, clear speech does not maintain an intelligibility advantage above a certain 'cut-off' speaking rate. A possible reason for this cut-off threshold is that there is a physical limit on how fast articulatory movements can be made, as assumed by Lindblom (1963). Adank and Janse (2009) compared the perceptual word processing speed of Dutch sentences that had been accelerated in two ways: (1) by asking the speakers to speak faster, and (2) by linearly time-compressing

sentences originally produced at a normal rate. Intelligibility of natural-fast speech turned out to be far worse than that of the time-compressed speech in terms of listener recognition accuracy. It seems that the human perceptual system can handle the more rapid acoustic changes in the synthetically accelerated speech, but naturally produced fast speech may already contain too much undershoot owing to various speed limits of articulation being reached, therefore making it difficult for listeners to decode information.

Support for the speed limit account can be found in studies of maximum speech rate. Sigurd (1973) examined the relationship between syllabic duration, syllabic structure and maximum speaking rate, and his data suggested that fast (or short) syllables are preferential in running text. That is, natural speech production tends to reorganize syllables with complex structures into simpler and thus articulatory faster ones. Further, Tiffany (1980) reported that for equivalent syllables, normal speech is no slower than the maximum rate of syllable articulation – both are approximately 13.5 phones per second. Tiffany's results indicate that, in terms of articulatory rate, there appears to be some form of highly rigid 'barrier', beyond which fully formed articulations cannot be achieved. This barrier concept is consistent with the notion of minimum duration of segments, which, according to Klatt (1976, p. 1215), is "an absolute minimum duration D_{min} that is required to execute a satisfactory articulatory gesture".

1.1 Extreme phonetic reduction

One way to examine if a 'speed barrier' is indeed a relevant mechanism in phonetic reduction is to look into cases where the barrier is most likely to be encountered. One such case is extreme phonetic reduction where an entire syllable is lost or merged into another syllable. An early observation of this phenomenon is seen in Stampe's discussion (1973) regarding variants of *divinity fudge* being shortened from their canonical forms [dəvínəti fʌ́ʤ] to severely reduced forms [dəvḯ̃ fʌ́ʤ] (quoted by Johnson, 2004). Similar examples of a sequence of two or more syllables being reduced into one are common in many other languages and are not exclusive to the segmental level. For instance, in Taiwan Mandarin[1], *wo zhi dao* [wo˨˩ ʈʂɨ˥ tau˥˩], 'I know' can be reduced into *wo zhao* [wo˨˩ ʈʂau]. Figure 1.1 illustrates the process, where the vowel /ɨ/ and the intervocalic consonant /t/ are omitted. The canonical tone shapes of H (˥, 55) in the syllable [ʈʂɨ] and F (˥˩, 51) in the syllable [tau] are also realised as a slightly sloping contour. In more extreme cases, trisyllables can also be reduced to monosyllabic units, such as *wo*

[1] Taiwan Mandarin here refers to the standard Mandarin natively spoken by people in Taiwan. It has four lexical tones: High (55, ˥), Rising (35, ˧˥), Low (21 or 214 if it occurs pre-pausally, ˨˩) and Falling (51, ˥˩). The digits in parenthesis are the conventional numeric notions for tonemes proposed by Chao (1930). Digit 5 indicates the highest pitch value and 1 the lowest within a speaker's normal pitch range. Owing to the constant influence of Southern Min, Taiwan Mandarin has developed its own stable linguistic system, which is distinct from the Mandarin spoken in Beijing.

bu zhi dao [wo˨˩ pu˧˥ tʂɨ˥ tau˧˥], 'I don't know' becoming *wo bao* [wo˨˩ pwau]. The present thesis is an investigation of extreme reduction in Taiwan Mandarin with the goal to identify some of the basic mechanisms of phonetic reduction in general.

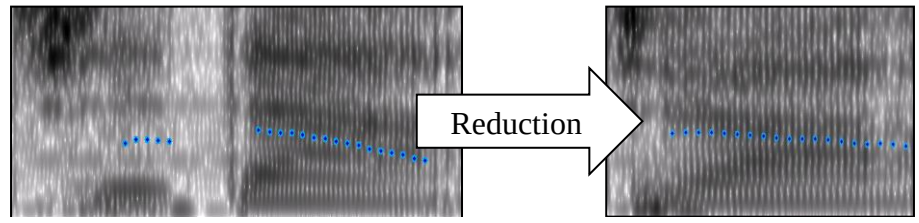


Figure 1.1: Spectrographic representation of [tʂɨ˥ tau˧˥] (on the left), '(I) know' being reduced to [tʂau] (on the right). Shorter duration and a reduced tonal range are also seen in the reduced token. Pitch values are shown as dots overlaid on the spectrograms.

Several terms have been used to refer to this severe form of phonetic reduction, including 'massive reduction' (Johnson, 2004), 'syllable fusion' (Wong, 2004 and 2006), 'syllable merger' (Duanmu, 2000) and 'syllable contraction' [2] or 'contraction' for short (Cheng, 2004; Chung, 2006; Hsiao, C. 1986; Hsiao, Y. C. 2002; Hsu, 2003; Kuo, 2010; Tseng, 2005a, b). Throughout this thesis, the term 'contraction' will be used to refer to such reductions on the grounds that it has

[2] The term 'syllable contraction', however, has been used to refer to two different phenomena. One is the extreme phonetic reduction (Tseng, 2008) that this research is concerned with. The other is the morphophonological process involving combinatory phonetic modifications of adjacent morphemes, e.g. English contracted forms *I'm* or *don't*, which might be arguably fossilized cases of phonetic reduction (Vance, 2008, p. 48; Suihkonen, 2005).

generally been used in research concerning severe forms of phonetic reduction in the Sinitic languages. More specifically, for the purpose of this study we define a 'contracted syllable' as a unit merged from its two source syllables in which the original intervocalic element cannot be easily detected. A more technical definition will be given in Chapters 2 and 3.

1.2 Main hypothesis and derived predictions

Studies regarding maximum speaking rate and the notion of minimum duration would suggest that the assumed additional effort in clear speech styles may not guarantee a full pronunciation, especially when duration is extremely short (e.g. at fast speech rate). In view of this, the following hypothesis is proposed.

Hypothesis:

- ***Time pressure is the direct cause of extreme reduction such as contraction.***

Here, 'direct cause' implies two things. First, from a biomechanical perspective, (compared to an articulatory effort perspective) duration is more directly related to the occurrence and severity of extreme reduction. That is, if the duration is too short there is simply no way for speakers to realise a target fully despite the extra effort that might have been applied. Secondly, the commonly recognized factors associated with phonetic reduction such as lexical frequency, information load, social context and speaking style, are very likely to impact directly on duration,

which in turn determines the degree of target attainment through time pressure. Two predictions can be derived from this hypothesis, which will be tested in this thesis.

Prediction 1:

- ***Extreme reduction such as contraction can occur in nonsense words if time pressure is sufficiently high*.**

As noted previously, nonsense words have the lowest possible lexical frequency and could therefore lead to a clearer speech style than real words due to a likely allocation of greater articulatory effort. If, however, time pressure is the direct cause of extreme reduction, extreme reduction would occur in nonsense words if speakers say them at a sufficiently high speed rate.

Prediction 2:

- ***When contraction occurs, articulatory effort is not decreased*.**

Consistent with Prediction 1, high articulatory effort is assumed to be exerted when producing nonsense words. This would mean that, if phonetic reduction did occur in nonsense words under high time pressure, it could not have been due to reduced articulatory effort.

Nevertheless, it is still possible that, due to some unknown mechanism, articulatory effort is indeed lowered under high time pressure as the reduced

intelligibility of fast speech seen by Krause and Braida (2002) and Adnak and Janse (2009) may suggest. There is therefore a need to have an assessment of articulatory effort that is independent of reduction itself. Such an assessment will be carried out in order to test the second prediction of this thesis.

1.3 The challenge of measuring articulatory effort

Currently, there is no standard accepted method of measuring articulatory effort. Malécot (1955, p. 36) described articulatory effort as "a kinaesthetically felt degree of force of articulation". That is, there seems to be a psychological referent of articulatory effort that speakers can 'feel in their head'. But this is not an *objective measurement* that directly corresponds to physiological reality (Parnell and Amerman, 1977; Tatham and Morton, 2006). Lindblom (1990) borrowed from Nelson (1983) the notion that 'peak velocity' is an indicator of 'articulatory effort'. Nelson (1983) characterised skilled movements using basic mechanical principles, and proposed that the peak velocity of an articulatory movement can be equated to the impulse cost measure (time integral of the magnitude of the force per unit mass) when there is negligible friction, as shown in the following equation:

$$\text{Impulse cost:} \quad I = \frac{1}{2}\int_0^T |u(t)|dt, \tag{1.1}$$

where *u(t)* is the applied force per unit mass (acceleration) and *T* is total movement time. In this equation, impulse cost (i.e. the equivalent of peak velocity) is proportional to movement time (duration), which means that the longer the

movement the greater the effort. However, this is different from the notion of economy of effort envisioned in the H&H theory, according to which effort is relatively independent of duration. Thus the use of peak velocity as a direct indicator of effort carries an intrinsic amalgam between time and force (see Kirchner, 1998 for a similar argument).

An empirical method used in a number of studies to assess articulatory effort is to examine the relation between peak movement velocity and movement amplitude (Kelso et al., 1985; Ostry et al., 1983; Ostry and Munhall, 1985; Perkell et al., 2002; Xu and Wang, 2009). These studies have consistently found that peak velocity is quasi-linearly related to movement amplitude. Such a quasi-linear relation means that peak velocity cannot be taken as an indicator of articulatory effort without knowing movement amplitude. That is, values of peak velocity are comparable only if they are from the same movement amplitude. It also follows that a steeper slope of peak velocity over movement amplitude may indicate greater muscle stiffness (Perkell et al., 2002), which would be related to articulatory effort. Therefore, if peak movement velocity is regressed over movement amplitude, the contribution of the movement amplitude can be normalised, making it possible to compare relative articulatory effort in movements of different sizes. This has been done in a number of previous studies (i.e. Perkell et al., 2002; Xu and Wang, 2009, as mentioned earlier).

In this thesis, the slope of regression of peak movement velocity over movement amplitude will be used to test the second prediction regarding articulatory effort.

To further justify the use of this measurement, it is necessary to show that a linear relationship between peak movement velocity and movement amplitude is present in the system of interest. This is shown in Sections 2.1.2 (Figure 2.9 for formant movement) and 3.1.1 (Figure 3.5 for F_0 movement).

Also, a critical issue concerning the methodology of the present study is the validity of using acoustic measurements to infer articulatory dynamics. This issue is addressed in some detail in the Appendix.

1.4 Structure of this thesis

This thesis is structured as follows: **Chapter 2** reports on the analysis of extreme segmental reduction. Two specific predictions derived from the main hypothesis that time pressure is the direct cause of extreme reduction are tested. Study 1 tested Prediction 1 to see whether extreme reduction could be elicited from nonsense words by simply increasing speech rate. Study 2a tested Prediction 2 by assessing articulatory effort of various degrees of phonetic reduction. Following these laboratory experiments, in Study 2b two sets of high frequency words extracted from three spontaneous speech corpora were analysed for cases of extreme reduction. An analysis of spontaneous speech corpora was carried out to verify the ecological validity of the experimental results and to examine the applicability of the experimental finding to the high-end extrema of the lexical frequency scale.

Chapter 3 concerns the analysis of tones in contracted syllables. To examine whether the nature of tonal reduction can also be explained by time pressure, Study 3 again tested Predictions 1 and 2 in order to further scrutinize the findings of Chapter 2. Additionally, a third prediction was tested: *When contraction occurs, speakers still attempt to approach each and every underlying tonal target.* This additional prediction was tested against the Edge-in model (Yip, 1988) which models the underlying target formation when extreme reduction such as contraction occurs.

Chapter 4 uses computational modelling to further test the time pressure account by checking whether extreme tone reduction can be reproduced by the computational model, and whether there is evidence that even under time pressure the canonical tonal targets are attempted. The main modelling method used is an articulatory-based model, the quantitative Target Approximation model (Xu and Wang, 2001; Prom-on et al., 2009). A supplementary method, Functional Linear Modelling (based on Functional Data Analysis) is also used to assess the nature of the durational effects on tonal variations.

Finally, the conclusions of this thesis and some possibilities for further work are discussed in **Chapter 5**. Some of the work presented in this thesis has been presented and published previously (Cheng and Xu, 2008a, b; Cheng and Xu, 2009; Cheng, Xu and Gubian, 2010; Cheng, Xu, Prom-on, 2011; Cheng and Gubian, 2011).

Chapter 2

Segmental reduction

In this chapter, two predictions based on the general hypothesis that *time pressure is the direct cause of extreme reduction such as contraction* are tested. As mentioned in the Introduction, in order to test this hypothesis data from two specifically designed experiments along with a corpus data set will be analysed. The methodology of both of these experiments (i.e. Study 1 and Study 2a) and discussion regarding the choice of data from the spontaneous speech corpora (Study 2b) will be presented first and then this is followed by their respective results and analyses. Following this the validity of the general hypothesis will be discussed.

2.1 Methodology

2.1.1 Study 1

The first experiment was designed to test Prediction 1 that *extreme reduction can occur in nonsense words if time pressure is sufficiently high*. To accomplish this, our strategy was to simply ask subjects to speed up their articulation and observe whether extreme reduction occurred. If this prediction is met, we will then try to identify a particular duration below which extreme reduction is regularly observed.

A. Stimuli

Testing materials for the experiment were constructed of 32 nonsense disyllabic sequences. Details of these stimuli are shown in Table 2.1. The target sequences were divided into four groups according to level of obstruction by intervocalic consonant: 1) zero obstruction – CV+V; CV+VN; CV+VV, 2) nasal consonant – CV<u>N</u>+V; CV+<u>N</u>V, 3) non-nasal consonant – CV+<u>C</u>V, where <u>C</u> is a fricative (fr.), plosive (pl.) or affricate (af.), and 4) nasal + non-nasal consonant – CV<u>N</u>+<u>C</u>V. Other combinations with non-nasal consonants as coda consonants (eg. CV<u>C</u>+NV or CV<u>C</u>+CV) were not considered owing to their absence in Mandarin phonotactics. All intervocalic consonants had a similar place of articulation (i.e. alveolar) but different manners of articulation to allow us to focus on the effect of the obstruction level. The vowels in these sequences were /i/, /a/ and /u/ in order to maximize variability in the amplitude of formant movement. Not all possible sequences of the selected vowels in all obstruction conditions were tested. In obstruction level 3, only non-nasal consonant – CV+<u>C</u>V, a balanced vowel

sequence, was used to examine articulatory demand and relative formant excursion size. All the disyllabic units had the same high-level tone in order to minimize potential tonal effects (Xu, 2001). [3]

Time pressure was controlled in two ways. The first was through the manipulation of durational variation related to position of the token in the sentence and in the phrase (Klatt, 1975) which has also been demonstrated for Chinese (Xu and Wang, 2009). This was achieved by devising a carrier sentence consisting of three phrases, each having a slot for the same target sequence (see Table 2.2). The first phrase consisted of eight underlying syllables, the second 13 and the third 15, all of which included the disyllabic target words. The second method was to elicit different speaking rates through direct instruction to the subjects (as detailed below).

[3] For readers less familiar with the Chinese language: Each Chinese character represents a monosyllabic morpheme. A syllable, even if with the same tone, may correspond to different morphemes. For example, the syllable /an/ with a falling tone can correspond to morphemes written as 暗, 岸 or 按, meaning ‘dark’, ‘shore’ or ‘to press’. Another morpheme, written as 案, is of the same pronunciation and carries multiple meanings including ‘a project’, ‘a long table’ or ‘a legal case’. Which of the meanings it takes depends on its combination with other morphemes to form words or compounds and on the phrasal and sentential semantic context. With such semantic flexibility, disyllabic nonsense words used in this study were designed to be as semantically unexpected as possible.

Table 2.1: Stimuli used in Study 1. The shaded rows indicate the obstruction level of the intervocalic consonants from low to high.

Disyllabic structure		Phonetic presentation and characters		
1. Zero obstruction				
CV+V		/ti/+/i/ 滴依	/ta/+/a/ 搭阿	/tu/+/u/ 督巫
CV+VN		/ti/+/in/ 滴因	/ta/+/an/ 搭安	/tu/+/un/ 督溫
CV+VV		/ti/+/ai/ 滴哀	/ti/+/au/ 滴凹	/tu/+/ai/ 督哀
		/tu/+/au/督凹		
2. Nasal consonant				
CV<u>N</u>+V		/tan/+/i/ 單依	/tan/+/u/ 單巫	
CV+<u>N</u>V		/ta/+/ni/ 搭妮*	/ta/+/nu/ 搭奴*	
3. Non-nasal consonant				
	fricative	/ta/+/ɕi/ 搭悉	/ta/+/su/ 搭蘇	/ta/+/sa/ 搭撒
	plosive	/ta/+/ti/ 搭滴	/ta/+/tu/ 搭督	/ta/+/ta/ 搭搭
CV+<u>C</u>V where C is a	plosive^{h}	/ta/+/t^{h}i/ 搭踢	/ta/+/t^{h}u/ 搭禿	/ta/+/t^{h}a/ 搭他
	affricate	/ta/+/tɕi/ 搭激	/ta/+/tsu/ 搭租	/ta/+/tsa/ 搭紮
	affricate^{h}	/ta/+/tɕ^{h}i/ 搭戚	/ta/+/ ts^{h}u/ 搭粗	/ta/+/ts^{h}a/ 搭擦
4. Nasal+ non-nasal consonant				
CV<u>N</u>+<u>C</u>V		/ɕin/+/ti/ 新滴	/sun/+/ti/ 孫滴	/san/+/ti/ 三滴

*Note that the characters in the second syllable of these two stimuli are indicative of an R tone. However, during the recording speakers were told to pronounce them as H tones (as with all other stimuli), so as to make up for the lack of actual morphemes.

Table 2.2: Carrier sentence used in Study 1.

Pinyin	ni shuo de shi ____ shi ba! wo dangran bu chi ____ shala nazhong dongxi, yinwei wo zui bu xihuan ta jia chu de ____ shala.
Character	你說的是_____是吧！我當然不吃_____沙拉那種東西，因為我最不喜歡他家出的_____沙拉!
English	You meant ____ , didn't you! Of course I won't eat ____ salad that kinda stuff, because I dislike ____ salad made by his family the most!

Note that in these carriers the nonsense words all occupy position of nouns, and are surrounded by high-frequency verbs, function words and nouns. This guarantees that not only lexically, but also semantically they are treated as high-information-load words.

B. Subjects and recording procedure

Six male Taiwan Mandarin speakers were recorded. They were aged between 21 and 28 and had no self-reported speech disorders or professional vocal training. The speakers were all postgraduate students studying in London whose prior education was in Taiwan. They had been in England for less than two years at the time of recording. Only male speakers were used because their formants are easier to track than those of female speakers. The recordings were conducted in an anechoic chamber at University College London. Speech was recorded with a Shure SM10A microphone placed approximately 30 centimeters from the subjects' mouth. The speech signals were recorded to a computer using the software package Adobe Audition v.1.5 with a sampling rate of 44.1 kHz. All stimuli were presented to the subjects in traditional Chinese characters and carrier sentences with the embedded stimuli were shown one at a time on the screen in front of the seated subject.

Subjects were instructed to articulate the material at three speaking rates: (1) slow and clear as if reciting in class, (2) in a natural manner as if conversing with a friend, and (3) as fast as possible. During each trial the speaker read out the sentences at the three speeds in the above order. No explicit instructions were given as to whether syllables can or should be contracted. However, if a speaker's pronunciation was too slurred, he was asked to repeat the entire trial (i.e., the carrier sentence displayed on the monitor at slow to fast speech rates). The exact speed of articulation was left to the subjects' discretion. The mean speech rates of slow, natural and fast across the six subjects were 4.9, 6.8 and 9.3 underlying syllables per second, respectively. To increase the size of the data sets, three randomized blocks of the above 32 sentence sequences were used. In total, the number of target sequences produced was 32 (stimuli) × 3 (positions in the carrier) × 3 (speech rates) × 6 (subjects) × 3 (blocks) = 5,184. Among these sequences, 31 (6%) were discarded from further analysis due to inadequate voice quality such as creaky voice or speaker errors.

C. Segmentation and measurements

The segmental labelling and measurements were conducted in Praat (Boersma and Weenink, 2010). Figures 2.1-2.3 display example spectrograms of target sequences produced in Study 1. It can be seen that, as the duration of target sequence decreases, the spectrographic patterns become increasingly simplified until little or no trace of the intervocalic consonant is left when the duration is at its minimum. Such spectral reduction is apparently much more severe than the cases of moderate reduction considered in studies of undershoot such as Lindblom

(1963). As one of the measurements, target sequences were classified as non-contracted (NC), semi-contracted (Semi) and contracted (Cntr) based on their degree of intervocalic segmental weakening or loss. *Non-contracted* units were those with clear interruption of formants by the intervocalic consonant, presence of nasal murmur or a clearly lowered F1. *Contracted* units were those with continuous F1, without interruption by either intervocalic consonants or nasal murmur. Units classified as *semi-contracted* were those for which the above segmentation criteria were difficult to apply and no straightforward delimitation of the spectrogram could be made.

All sound files were segmented and labelled by the author, a native Taiwan Mandarin speaker. The consistency of contraction type labels was double checked one month following the initial labelling. Uncertainty in the labelling occurred only very occasionally, and in all such cases the uncertainty was related to semi-contracted units. A handful of tokens were relabelled from non-contracted or contracted to semi-contracted upon rechecking. There were no tokens of non-contracted relabelled as contracted or vice versa. It is important to note that in the zero obstruction level, according to the current labelling criteria, most tokens were marked as contracted because the disyllabic sequence consisted of an open CV syllable followed by a syllable with a vowel onset. Hence, unless there was a clear sign of glottal stop or glottalization between the two vowels, as shown in Figure 2.1, they were marked as contracted.

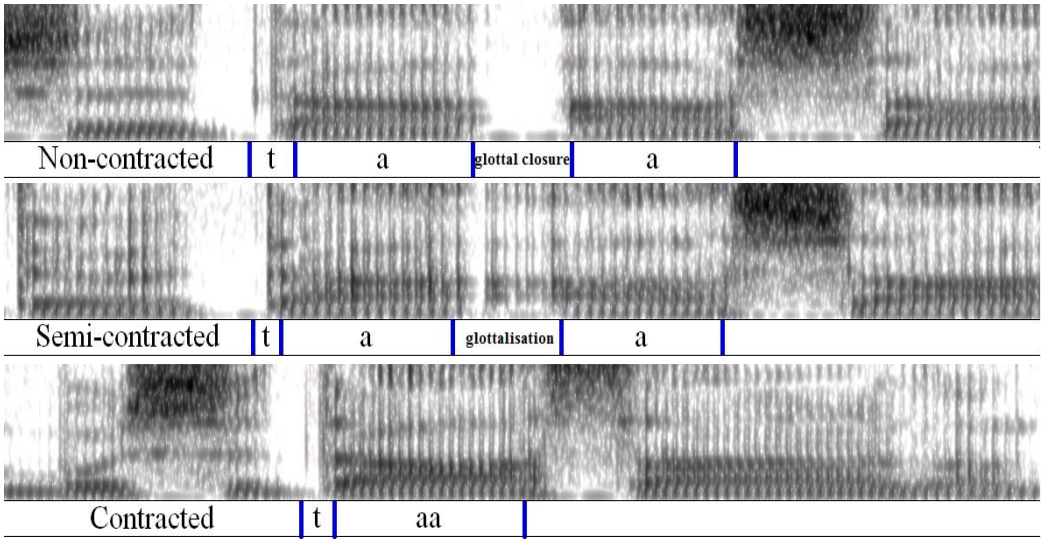


Figure 2.1: Examples of labelling /ta/+/a/ (zero intervocalic obstruction). From top to bottom, non-contracted (realized with a full glottal stop), semi-contracted (realized with glottalization), and contracted (with continuous formants). The time domains are of similar window length from 0 to 1 second and formant frequency from 0 to 5000 Hz.

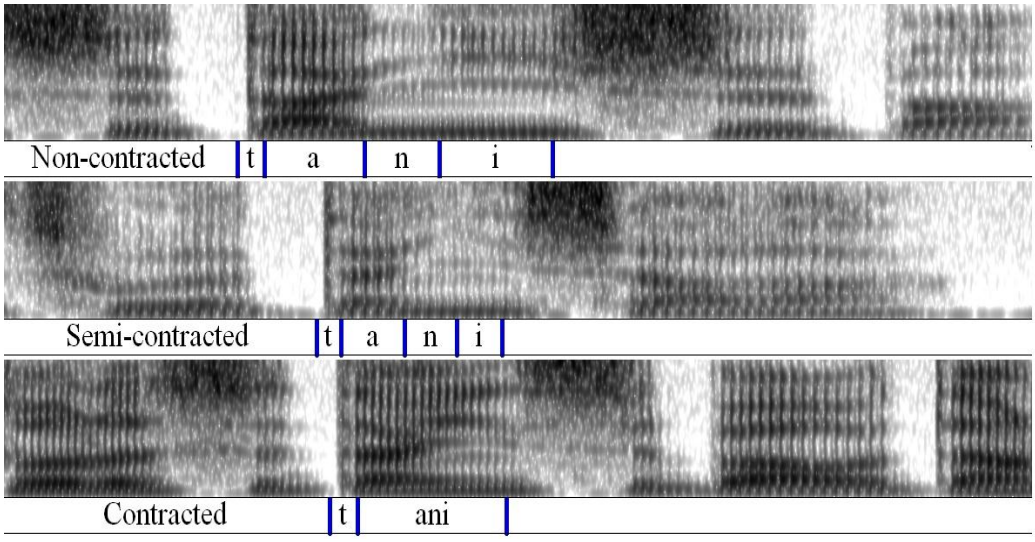


Figure 2.2: Examples of labelling /tan/+/i/ (intervocalic nasal consonant). From top to bottom, non-contracted, semi-contracted, and contracted. The time domains are of similar window length from 0 to 1 second and formant frequency from 0 to 5000 Hz.

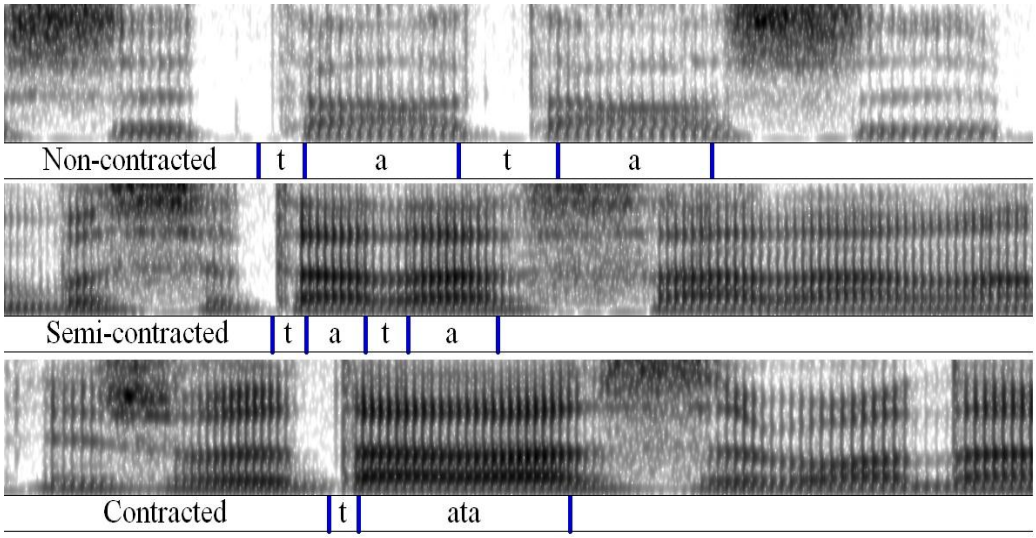


Figure 2.3: Examples of labelling /ta/+/ta/ (non-nasal intervocalic consonant). From top to bottom, non-contracted, semi-contracted, and contracted. The time domains are of similar window length from 0 to 1 second and formant frequency from 0 to 5000 Hz.

2.1.2 Study 2a – Laboratory data

The second experiment was aimed at examining the continuous reduction process in greater detail than in Study 1 and testing whether articulatory effort was strengthened or weakened when contraction occurred. The method was to track the formant trajectories and velocity profiles of two near symmetric articulatory movements so as to determine the relative contributions of duration and articulatory effort.

A. Stimuli

Two nonsense disyllabic sequences, /ta/+/ja/ and /ta/+/wa/ with intervocalic glides /j/ and /w/, were devised to allow observation of formant trajectories without

interruption by obstruent intervocalic consonants (see Table 2.3). The use of glides also avoids the issue of articulatory overlap between C and V. This is due to the fact that glides, being semivowels, are specified for the entire shape of the vocal tract rather than predominantly at a particular place of articulation as in the case of obstruents (Moon and Lindblom, 1994; Xu and Liu, 2007). This would help make the interpretation of the relation between articulatory effort and duration more straightforward. The same carrier sentence as in Study 1 was used (see Table 2.2).

Table 2.3: Stimuli used in Study 2a.

Disyllabic structure	Phonetic presentation and characters	
CV+GV	/ta/+/ja/ 搭壓	/ta/+/wa/ 搭挖

B. Subjects and recording procedure

Four of the six subjects from Study 1 were re-recruited to participate in this experiment. Two other male subjects with very similar linguistic backgrounds to those of Study 1 were added. The same procedure as Study 1 was followed. The total number of target sequences produced in this experiment was 2 (stimuli) × 3 (positions in the carrier) × 3 (speech rates) × 6 (subjects) × 3 (blocks) = 324 tokens.

C. Segmentation and measurements

Segmental labelling was carried out in a manner similar to that of Study 1 but with some slight modifications. When producing /ta/+/ja/ and /ta/+/wa/ within a carrier sentence, the speaker's vocal tract was always open for the semivowels /j/ and /w/. Thus there were few non-contracted cases going by the previous criteria, that is, a pause between the first and second vowels (as seen in Figure 2.1 for the zero-obstruction level in Study 1). Therefore, the labelling of degrees of reduction in Study 2a used F1 dip as a primary indicator and F2 peak or valley as a secondary indicator. In producing intervocalic glides /j/ and /w/ the articulators need to move to the position of the glide from the position of the preceding vowel /a/ and then to the position of the following vowel /a/. Since F1 was very low in both /j/ and /w/, cases with a fall followed by a rise in F1 were marked as *non-contracted*. In contrast, cases in which both formants (F1, F2) were nearly flat (and no obvious curve could be seen) were marked as *contracted*. Units marked as *semi-contracted* were cases where the preceding two classifications were not straightforwardly applicable. Such cases commonly showed a slight F1 dip along with a slight F2 rise for /j/ and a slight F2 fall for /w/. Examples of the labelling are shown in Figures 2.4-2.5.

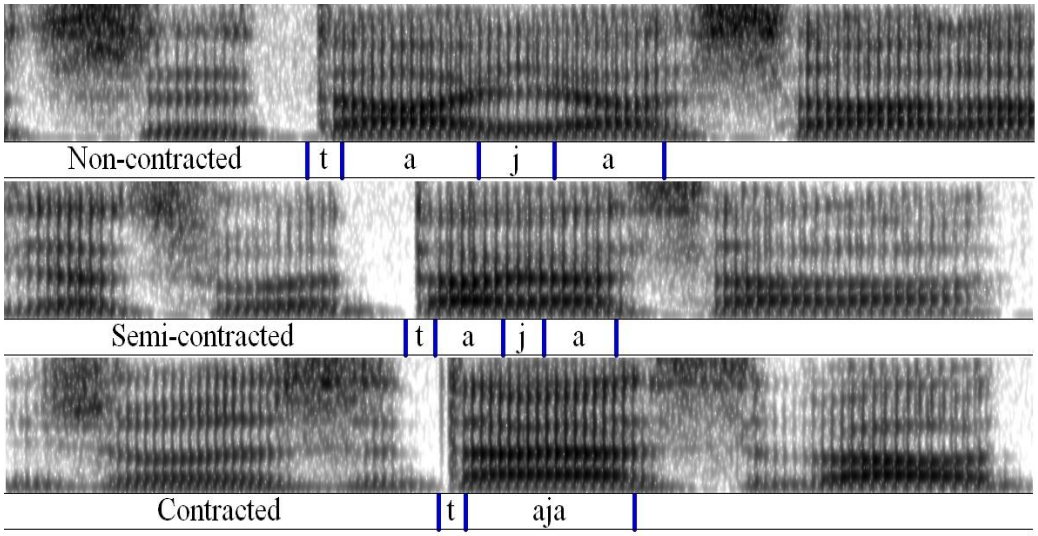


Figure 2.4: Examples of labelling /ta/+/ja/ (intervocalic glide /j/ in-between). From top to bottom, non-contracted, semi-contracted, and contracted. The time domains are of similar window length from 0 to 1 second and formant frequency from 0 to 5000 Hz.

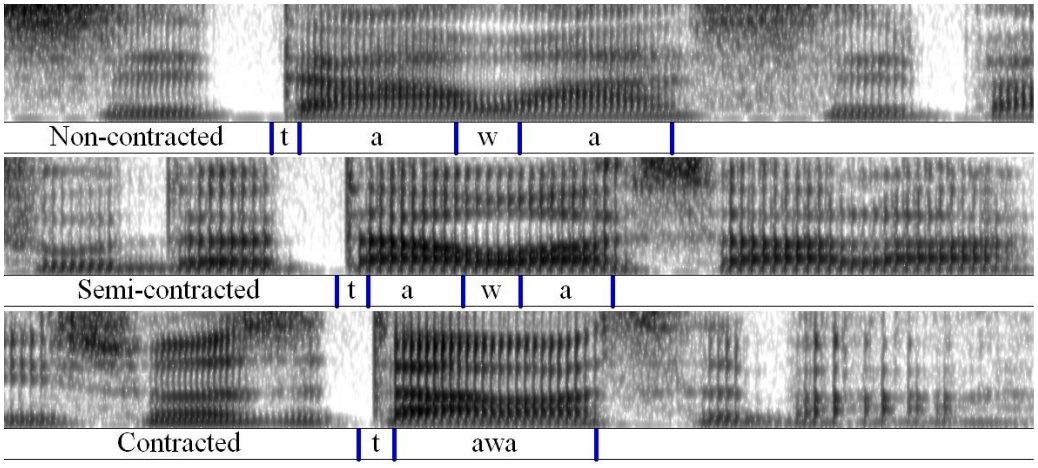


Figure 2.5: Examples of labelling /ta/+/wa/ (intervocalic glide /w/ in-between). From top to bottom, non-contracted, semi-contracted, and contracted. The time domains are of similar window length from 0 to 1 second and formant frequency from 0 to 5000 Hz.

In order to assess articulatory effort, a set of kinematic measurements, including movement duration (the elapsed time between two formant turning points), movement amplitude (difference in semitones between two turning points), and peak velocity (highest absolute value in the velocity profile corresponding to a movement), were taken using a Praat script specifically written for this experiment. The script uses the Berg algorithm to extract continuous formants and applies a trimming algorithm (originally designed for processing F_0 contours, cf. Xu, 1999) to remove excessive and sudden bumps in the formant trajectories. It then computes the velocity (i.e., the first derivative) of the formants using a two-point central differentiation algorithm (Bahill et al., 1982). To illustrate this process, Figure 2.6 shows the F1 movement of syllable /ta/+/ja/ and its velocity profile. The script first searched for the F1 minimum (point B in Figure 2.6) in the LPC formant track generated by Praat. It then finds peak velocities from within each of the two intervals (interval 1: from A to B and interval 2: from B to C). Similar procedures were applied to F2 movements. In cases where formant trajectories either became effectively flat (as illustrated in Figures 2.7-2.8, in particular for the contracted cases) or had a direction different to that of the canonical form, the kinematic measurements would become erroneous. Such cases (146 out of 324 for F1 and 112 out of 324 for F2) could not be used to generate valid measurements using this algorithm and thus were not processed to estimate articulatory effort.[4]

[4] Note that, had the algorithm in the script been written in such a way that all problem tokens were included, we would have greatly increased the number of data points clustered along the left extreme of the y-axis in Figure 2.9. This is because the problem tokens mostly had extremely small movement amplitudes but

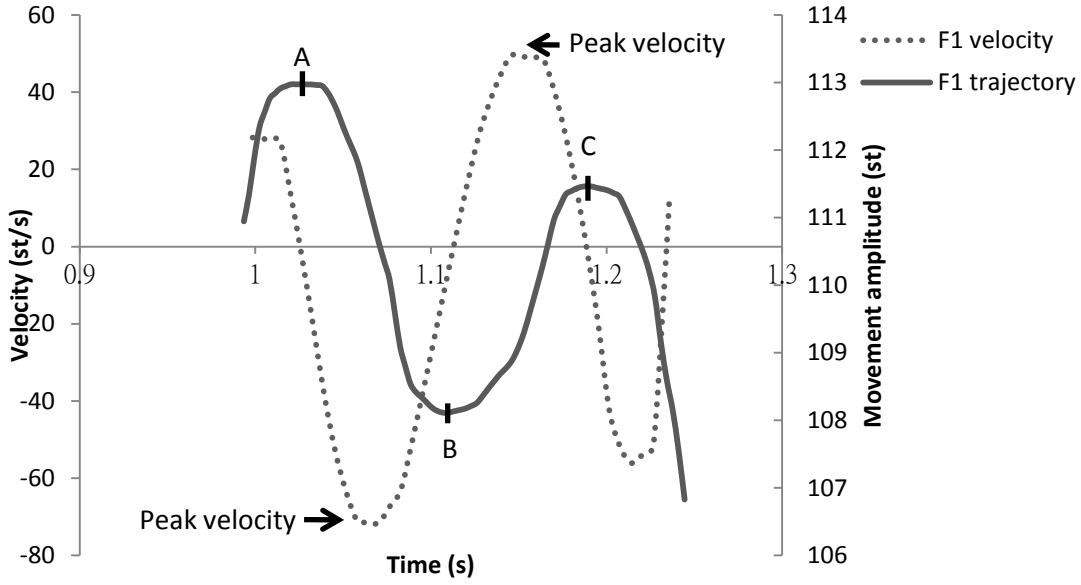


Figure 2.6: F1 trajectory (solid line) and its velocity profile (dotted line): A, B and C mark the turning points of F1 trajectory and delineate two intervals (interval 1: A-B and interval 2: B-C). Two peak velocities were obtained, one within each interval. The x-axis is the time domain in seconds. The y-axis on the right hand side is in units of semitones for the F1 trajectory and the y-axis on the left is in the units of semitone per second for the F1 velocity.

---

highly variable peak velocities. Such data would have been uninformative. Note also that such a high exclusion rate is due to the intrinsic characteristic of extreme reduction, which by its very nature, necessarily involves virtual destruction of the integrity of the underlying articulation. Thus there is an unavoidable trade-off between being able to simulate and systematically analyse extreme reduction in an experimental setting and not having to throw out a substantial amount of uninformative data (i.e., only examining cases well short of extreme reduction). The present study has given priority to the former in the interest of pushing the boundaries of our understanding of speech on this inherently difficult issue.

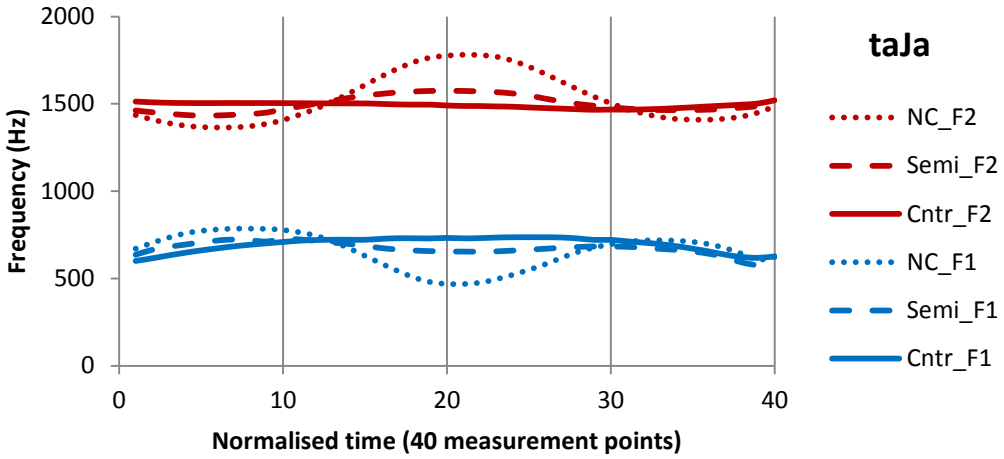


Figure 2.7: Formant trajectories (F1 in blue and F2 in red) of /ta/+/ja/ sequences averaged across all six subjects. The x-axis shows time-normalised 40 measurement points and the y-axis is formant frequency in Hz. The legend indicates both formants of different contraction types.

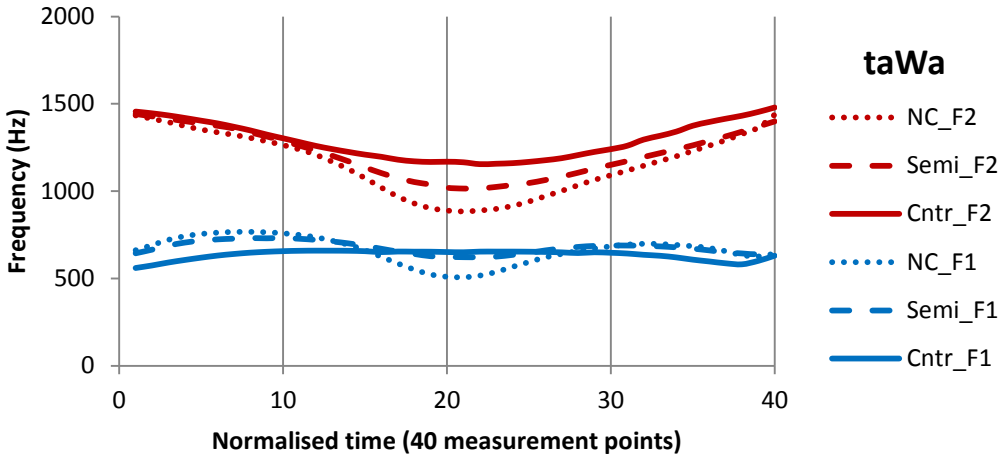


Figure 2.8: Formant trajectories (F1 in blue and F2 in red) of /ta/+/wa/ sequences averaged across all six subjects. The x-axis shows time-normalised 40 measurement points and the y-axis is formant frequency in Hz. The legend indicates both formants of different contraction types.

As mentioned in Chapter 1 and explained in detail in the Appendix, at least theoretically, acoustic measurements such as formant frequencies are not inferior to measurements of individual articulators. Further confirmation can be seen in Figure 2.9 where formant data show similar kinematic patterns to those of articulatory movement. The figure displays scatter plots of F1 peak velocity as a function of F1 movement amplitude computed with data from all six subjects at all three speech rates. The relation between F1 peak velocity and F1 movement amplitude was highly linear ($r = .879$, $p = .001$). A similar linear relationship was also seen in F2 ($r = .859$, $p < .001$). Such linear relations are consistent with previous findings regarding articulatory movements, which have been considered to directly reflect the stiffness of the articulatory movements (Kelso et al., 1985; Ostry et al., 1983; Ostry and Munhall, 1985; Perkell et al., 2002). On this basis, the ratio of peak velocity and movement amplitude was used as an indicator of the articulatory effort applied.

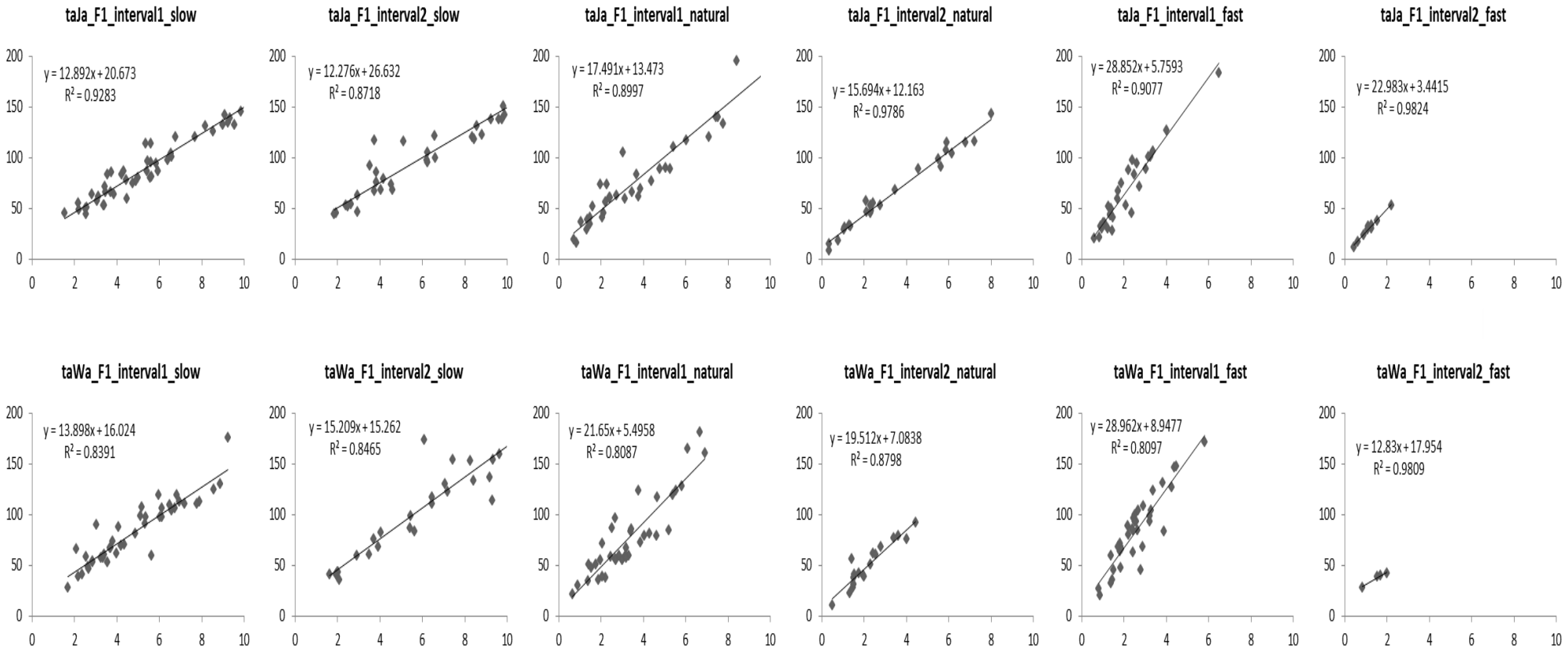


Figure 2.9: Linear relation of F1 peak velocity (y-axis in semitones/seconds) to F1 movement amplitude (x-axis in semitones) across slow, natural and fast speech rates.

2.1.3 Study 2b – Corpus data

Two sets of high frequency words from spontaneous speech corpora were examined for cases of extreme reduction. The goal of this study was to compare the kinematic measurements of nonsense words in Study 2a to those of high frequency words in spontaneous speech. Spontaneous speech materials were taken from three corpora provided by Academia Sinica. A brief summary of the corpora is given in Table 2.4, adapted from Tseng (2008, p. 3, Table 1).

Table 2.4: Summary of Academia Sinica corpora.

Corpus	Mandarin Conversational Dialogue Corpus (MCDC)	Mandarin Topic-Oriented Conversation Corpus (MTCC)	Mandarin Map Task Corpus (MMTC)
Scenario	Free conversation between strangers	Subjects knew each other well	Subjects knew each other well
Purpose	Disfluency	Dialogue acts	Phonetic variations
Period	2001.03-2001.07	2002.01-2002.03	2002.01-2002.03
Transcription	All orthographically transcribed and annotated		

(Adapted from Tseng, 2008)

From the three corpora, two sets of words were selected for analysis. Each set contained one stem form and one compound form as shown in Table 2.5. The stem syllables were *zheyang* ([tʃʰɤ˥˩ jaŋ˥˩], 'this') and *nayang* ([na˥˩ jaŋ˥˩], 'that') where all syllable sequences have the falling tone. The compound form was an added suffix *zi* ([tsɨ]) which has the neutral tone. These tokens were selected

because, firstly, they are of similar phonetic structure to the stimuli used in Study 2a (VGV), and secondly, both the canonical and reduced forms of the four words were frequent and giving enough tokens to allow reliable comparisons. In terms of lexical frequency, *zheyang*, *zheyangzi* and *nayang* rank 48th, 66th and 547th out of the 11,728th places in the Sinica corpus, respectively. (No ranking information of *nayangzi* was listed).[5] A total of 262 tokens from 17 speakers (eight males and nine females) were extracted. These speakers were in their twenties and had similar language backgrounds to those in Studies 1 and 2a.

Table 2.5: Selected units used in Study 2b.

Phonetic structure		Characters	Pinyin	Meaning	Count
1.	tʃʰɤ̲ jaŋ̲	這樣	zhe yang	this	148
	tʃʰɤ̲ jaŋ̲ tsɨ̶	這樣子	zhe yang zi	such this	103
2.	na̲ jaŋ̲	那樣	na yang	that	6
	na̲ jaŋ̲ tsɨ̶	那樣子	na yang zi	such that	5

All tokens extracted from the corpora were again labelled according to their degrees of reduction using the same criteria as in Study 2a. For each token, only the segments exhibiting relevant formant trajectories (for calculating articulatory effort, marked as underscored in Table 2.5) were subjected to further analysis. Three kinematic measurements (movement duration, movement amplitude and

[5] Other references regarding Chinese word statistics such as frequency rank and cumulative percentage can be found at http://elearning.ling.sinica.edu.tw/eng_teaching.html.

peak velocity) were obtained. As in Study 2a, erroneous measurements due to flattened or inaccurate formant trajectories were excluded (155 out of 262 for F1) and (136 out 262 for F2). Note that these corpus tokens were not as symmetric as those used in Study 2a in terms of their articulatory movements (i.e. they did not have the same vowels in both syllables), which lead to a smaller success rate in generating valid measurements (i.e., reducing from 55% for F1 and 65% for F2 in Study 2a to 41% for F1 and 48% for F2 in Study 2b, the corpus data).

2.2 Analysis and results

From Figures 2.1-2.3 we can see various effects of gradual reduction on the spectrographic integrity of the target sequences. In Figure 2.2, for example, /n/ in /ani/ gets weakened in semi-contracted tokens and virtually disappears as a separable segment in contracted tokens. Simultaneously, there is severe undershoot of the vowel /a/, as its F1 and F2 become more similar to those of the surrounding consonants. The same is true of the sequence /ata/ in Figure 2.3. Thus there seem to be two processes involved in continuous reduction: (1) disintegration of the intervocalic consonants and (2) undershoot of the flanking vowels. Study 1 is designed to mainly scrutinize process (1) by examining the conditions under which intervocalic consonants become severely reduced in a variety of VCV sequences. Process (2) will be more closely examined in Study 2a.

2.2.1 Study 1

A. Contingency of contraction type

Figure 2.10 shows the distribution of the three contraction types. Non-contracted and contracted occurred more frequently (43.63% and 47.04%, respectively) than semi-contracted (9.33%), leading to a binomial distribution of contraction types.

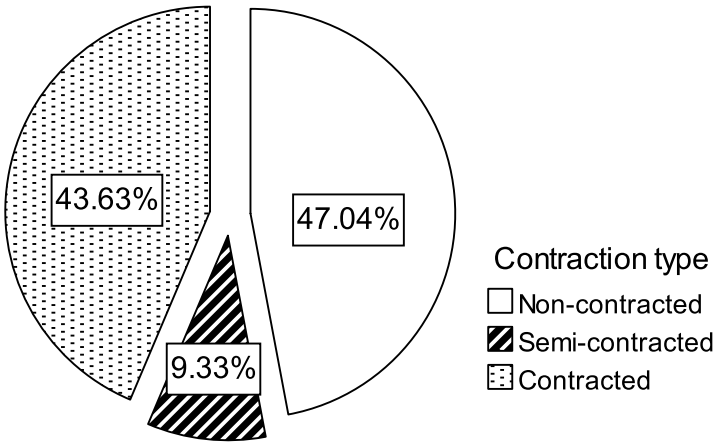


Figure 2.10: Distribution of the three contraction types in Study 1.

B. Speed and contraction type

A multinomial logistic regression was performed with contraction type as the ordinal dependent variable, and speed, obstruction level and position in the carrier sentence as predictors. Results showed that speed was positively correlated to contraction type (Coef. = 1.59, S.E. = 0.05, $p < .000$). For a unit increase in speed,

the expected ordered log odds increased by 1.59 as one moved to the following higher category of contraction (i.e. from non-contracted, semi-contracted to contracted). On the other hand, obstruction level was negatively related to contraction type (Coef. = -1.86, S.E. = 0.05, $p < .000$). For a unit increase in obstruction level, the expected ordered log odds decreased by 1.86 as one moved to the following higher category of contraction. Position had no effect on contraction type (Coef. = 0.02, S.E. = 0.04, $p = 0.58$).

Figure 2.11 shows the effect of speed on contraction type. In non-contracted types, a decline in frequency count is seen as speed increases. Conversely, in both the semi-contracted and the contracted, as speed increases frequency count also increases. The largest distributions in each contraction type are slow speed in non-contracted (23.79%), fast speed in semi-contracted (4.42%) and fast speed in contracted (20.98%). This is in agreement with the above statistics (i.e. Ordinal Logistic Regression), that is, a significant positive relationship is seen between speed (from slow to fast) and contraction type (from non-contracted to contracted).

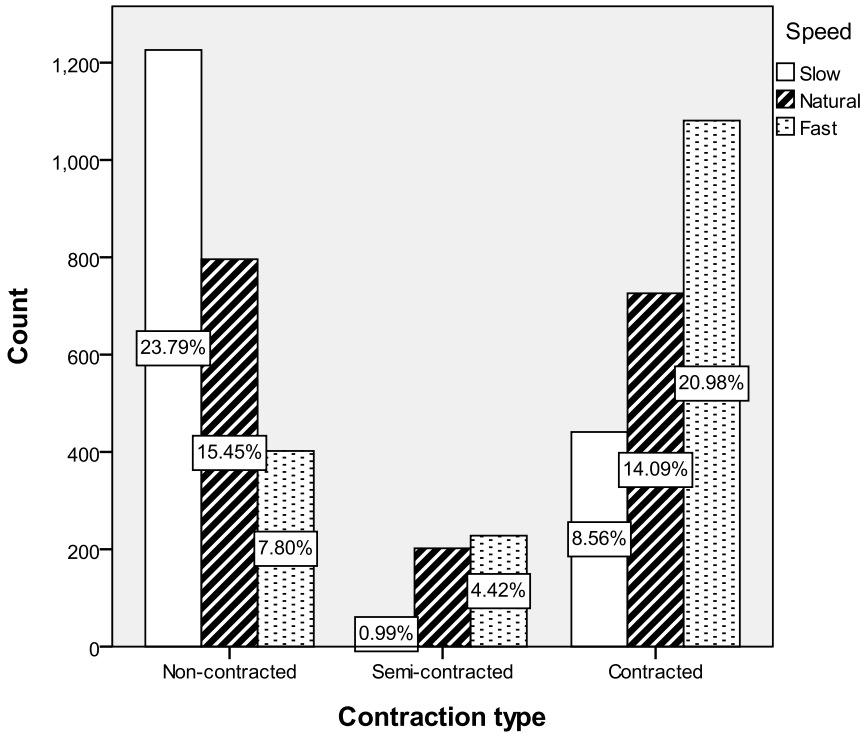


Figure 2.11: Contingency of contraction type at different speeds in Study 1. The x-axis shows three different contraction types and the y-axis shows frequency count.

Note that 8.56% of the contracted cases occurred at slow speed. A major contributor here is the zero-obstruction group that involves vowels as syllable onset in the second syllable and was therefore often labelled as contracted. To avoid this effect, a follow-up logistic regression was conducted with all zero-obstruction cases removed. Results remained comparable to those of the OLR report, i.e. a positive relation was observed between contraction type and speed (Coef. = 1.62, S.E. = 0.06, $p < .000$), a negative relation between contraction type and obstruction level (Coef. = -0.93, S.E. = 0.07, $p < .000$), and position had no effect on the contraction type (Coef. = -0.02, S.E. = 0.05, $p = .71$).

C. Phonetic structure and extreme reduction

Table 2.6 summarizes rates of extreme reduction (i.e. *contracted* cases) in different phonetic structures in terms of percentage of occurrences. (Semi-contracted items were not included owing to their ambiguous status as *extreme reduction*). In the zero-obstruction level, the rate of contracted cases was nearly 90%. This is a natural consequence of the lack of canonical consonantal obstruction to interrupt the vowel-to-vowel formant movements as mentioned earlier in section 2.1.1.C. A mean rate of 42.04% was seen in the nasal consonant level. It appears that it is easier to lose coda nasals (62.35%) than initial nasals (21.67%) under time pressure. As for the non-nasal consonant level, rates of contracted cases varied with manner of consonant articulation. Unaspirated obstruents (plosive and affricate) had higher rates of extreme reduction (mean: 22.42%) than their aspirated counterparts (mean: 17.72%). In the nasal + non-nasal consonant level, the highest intervocalic obstruction yielded the lowest rate of extreme reduction (10.54%).

Table 2.6: Contracted cases (%) at different levels of intervocalic obstruction. The left column indicates the obstruction level of the intervocalic consonants from low to high. The middle column shows percentage of contracted cases occurred in respect to each phonetic structure, and the rightmost column the means of contraction rates for different levels of obstruction.

Disyllabic structure		Contracted cases (%)	Mean (%)
1. Zero obstruction			
CV+V		88.41%	
CV+VN		93.15%	89.77%
CV+VV		88.27%	
2. Nasal consonant			
CV<u>N</u>+V		62.35%	42.04%
CV+<u>N</u>V		21.67%	
3. Non-nasal consonant			
CV+<u>C</u>V where <u>C</u> is a	fricative	18.71%	19.80%
	plosive	20.87%	
	plosive^{h}	19.21%	
	affricate	23.97%	
	affricate^{h}	16.22%	
4. Nasal+ non-nasal consonant			
CV<u>N</u>+<u>C</u>V		10.54%	10.54%

Table 2.6 demonstrates that as the level of obstruction increases, the rate of contracted cases decreases. This inverse relation implies that contracted cases are dependent on the level of articulatory demand, but time pressure may actually be a more likely determining factor. As the CV<u>N</u>+<u>C</u>V group indicates, reduction rate is related to the time allocated to the consonant: When there are two adjacent consonants (of similar articulatory demands, i.e. /n/ and /t/), presumably twice as much time is allocated to the closing gesture. When duration is shortened

proportionally under time pressure, these two consecutive obstruents are the last ones whose combined allocated time is reduced to the point when no closure of the vocal tract is possible.

D. Minimum duration

The above interpretation is further supported by Figure 2.12 which shows mean segmental duration of the *non-contracted* items in the /ta/+CV sequences. As can be seen, in the non-contracted units, the duration of the intervocalic consonants varies with their level of obstruction. Moreover, the duration of the second vowel varies compensatorily with the onset duration ($r = -0.97$, $p < .01$).

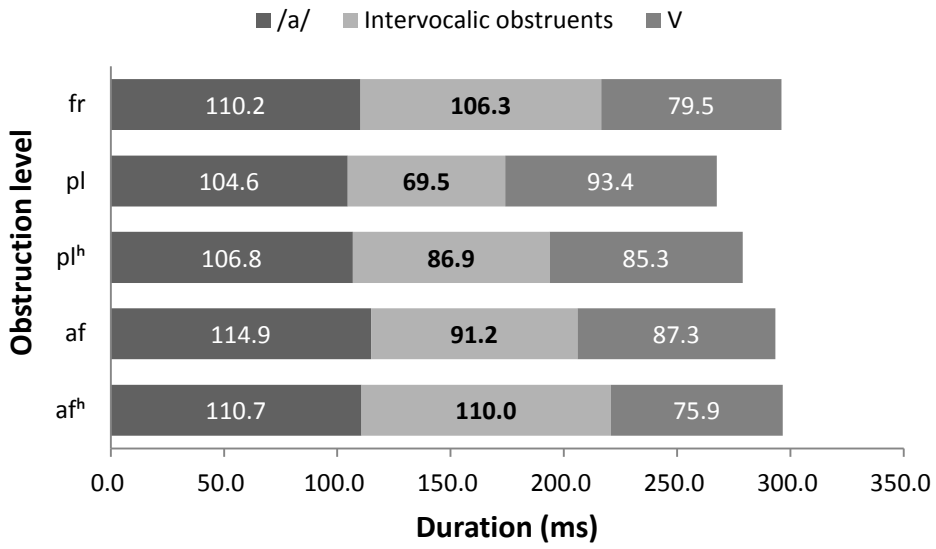


Figure 2.12: Varying segmental durations of the *non-contracted* /ta/+CV sequences. The x-axis indicates the consecutive durations of the preceding /a/, the intervocalic consonant and the following vowel. The y-axis shows the different intervocalic consonants in terms of manner of articulation, from top to bottom, fr: fricatives, pl: plosives, pl^h: aspirated plosives, af: affricates, and af^h: aspirated affricates.

To see the time demand of varying articulatory functions in a more straightforward manner, mean consecutive segmental durations of all three contraction types in the /ta/+CV sequences are plotted in Figure 2.13. Here the durations of non-contracted units may reflect the amount of time used in canonical articulations. In semi-contracted units, the duration of all segments is reduced with the most severe reduction in the intervocalic obstruents (44.4 ms). In contracted units, the overall duration of disyllabic words is compressed to the point where no intervocalic consonantal closure is possible. Therefore, 44.4 ms in the semi-contracted case appears to be the mean minimum duration below which intervocalic consonants are virtually 'lost'.

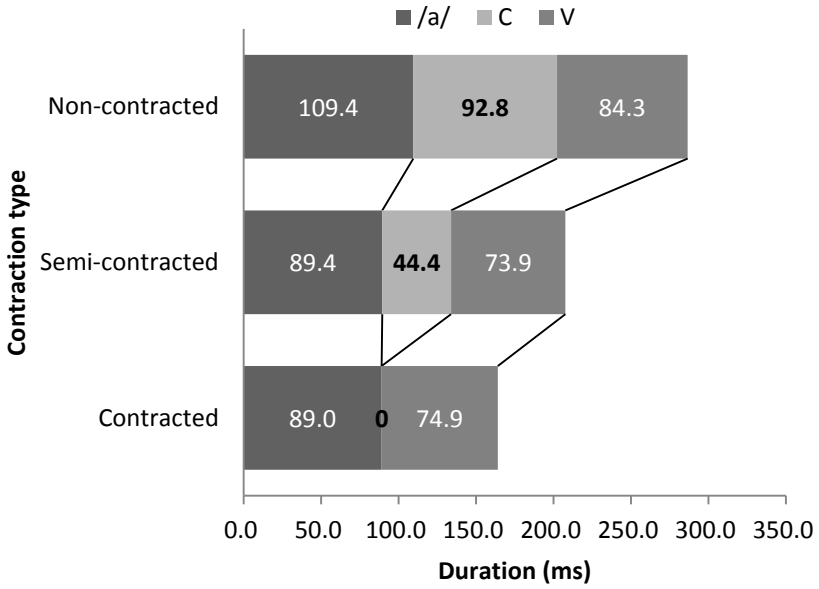


Figure 2.13: Segmental durations at different contraction types in all /ta/+CV sequences.

E. Duration and excursion size

As the degree of vocal tract constriction is reduced, the magnitude of formant displacement in the vowels is also reduced. It is therefore possible to further examine the effect of duration on phonetic reduction by observing the relationship between duration and formant excursion size as shown in Figure 2.14: A scatter plot of all items in the /ta/+CV sequences.

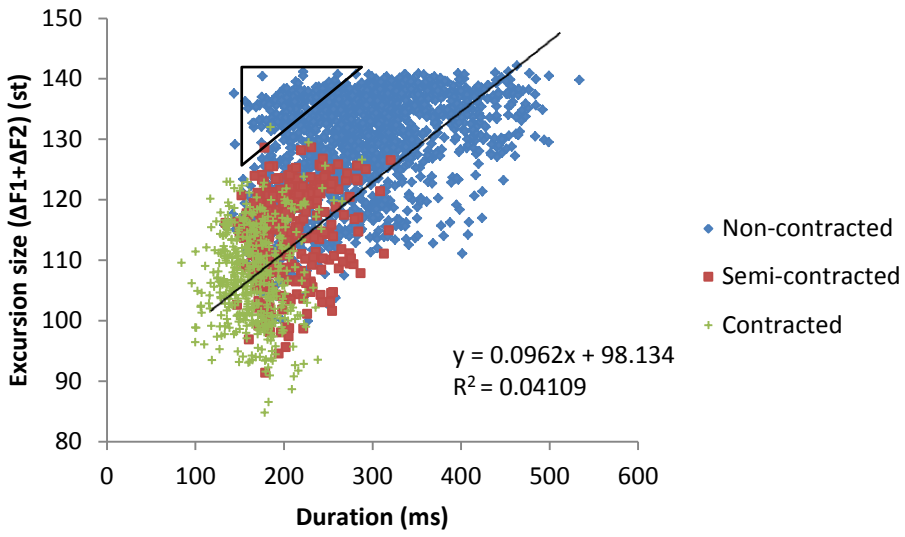


Figure 2.14: Scatter plot of formant excursion size over formant duration: /ta/+CV. The x-axis represents the combined duration of /a/ and the second syllable in /ta/+CV sequences. The y-axis represents the sum of F1 and F2 displacements within this interval.

The scatter plot demonstrates a positive relation between duration and formant displacement in /ta/+CV sequences: the longer the duration, the larger the displacement ($R^2 = 0.41$, $p < .001$). The plot also shows an orderly distribution of

the reduction levels as a function of duration, that is, most non-contracted cases had relatively long durations; when the duration became extremely short (less than about 200 ms) extreme reduction occurred. Note that there are also some cases with a duration of 200 ms labelled as *non-contracted* (as indicated within a triangle). This might suggest that extra effort was exerted by the speakers to avoid undershoot, as it will be examined further in Study 2a and 2b. In general, it is interesting to note that the boundary between non-contracted and contracted was approximately 200 ms, suggesting that the mean duration for the integrity of disyllables is around 200 ms (without the duration of the onset consonant /t/ in /ta/+CV).

2.2.2 Study 2a – Laboratory data

A. Contingency of contraction type

Figure 2.15 displays the distribution of the three contraction types in Study 2a. Contracted and semi-contracted cases occurred with similar frequencies (27.27% and 19.48%, respectively) and non-contracted occurred most frequently (53.25%).

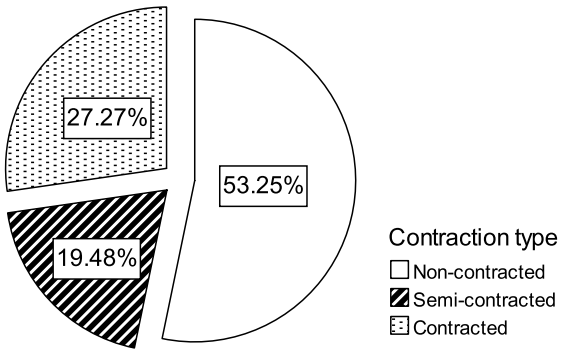


Figure 2.15: Distribution of the three contraction types in Study 2a.

B. Speed and contraction type

A multinomial logistic regression was performed with contraction type as the ordinal dependent variable and speed as well as position in the carrier sentence as the predictor variables. Results showed that speed was positively related to contraction type (Coef. = 2.51, S.E. = 0.23, $p < .000$). For a unit increase in speed, the expected ordered log odds increased by 2.51 as one moved to the adjacent higher category of contraction (i.e. from non- to semi- to contracted). Position had no effect on contraction type (Coef . = -0.002, S.E. = 0.16, $p = 0.99$). Thus the statistical results of Studies 1 and 2a both showed similar patterns of dominant speed effect on the occurrences of contraction.

C. Duration and excursion size

In Figure 2.16 a scatter plot displaying the relationship between duration and formant excursion size for all tokens of /ta/+/ja/ and /ta/+/wa/ is shown. Similarly to Figure 2.14, Figure 2.16 also shows a positive relation between duration and

excursion size (R^2 = 0.67, p < .001) indicating that the two measurements are strongly related. A mean duration for the integrity of disyllables was again observed: Semi-contracted units of /ta/+/ja/ and /ta/+/wa/ sequences cluster around 200 ms (without the duration of the onset consonant /t/ in /ta/+GV). It should be noted that in Figure 2.14, the data points lying within the highlighted triangle are largely absent of which points in Figure 2.16. A possible explanation could be a phonetic structure (V.GV) that is less conducive to coarticulation (i.e. speakers need to implement segments one after another) than that of V.CV structure, for which the data in Figure 2.14 is shown, (where speakers can raise their tongue tip while moving the tongue body for the vowel sequence) when time pressure is high.

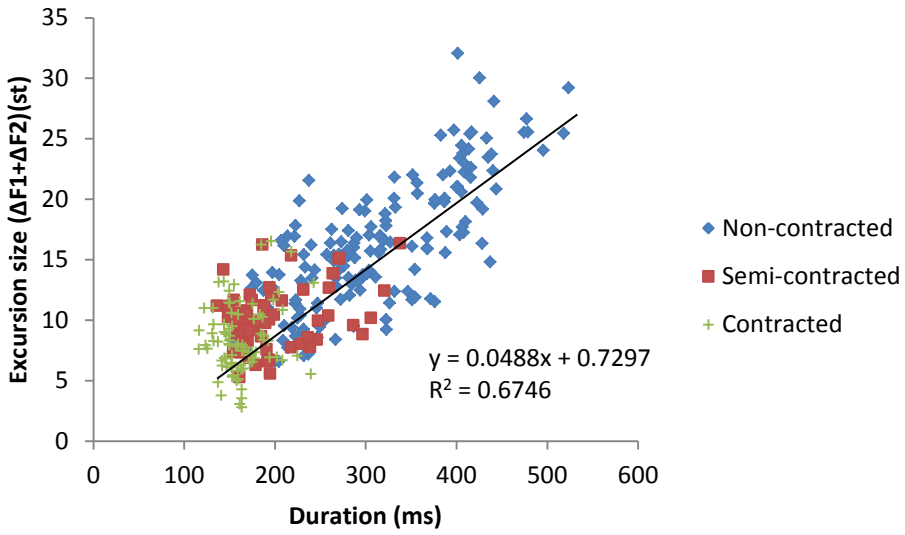


Figure 2.16: Scatter plot of formant excursion size over formant duration: /ta/+/ja/ and /ta/+/wa/.

D. Articulatory effort

The design of Study 2a allowed us to further examine the relationship between duration, formant displacement and articulatory effort for the three contraction types. Table 2.7 shows a set of one-way ANOVAs performed with contraction type as the independent variable and duration, formant displacement and slope of the regression line (peak velocity over movement amplitude) as dependent variables. All dependent variables were averaged with respect to /ta/+/ja/ and /ta/+/wa/ items.

Table 2.7: Mean duration (ms), formant displacement size (st) and slope of the regression line of formant peak velocity over formant movement amplitude of the three contraction types – Study 2a, laboratory data.

Type	Duration	F1 size	F1 slope	F2 size	F2 slope
Non-	313.8	20.31.	21.28	15.21	19.00
Semi-	199.0	8.86	26.95	8.82	23.95
Contracted	160.7	8.14	32.82	6.40	27.25
F value	$F_{(2,3)} = 360.7$	$F_{(2,3)} = 29.7$	$F_{(2,3)} = 167.7$	$F_{(2,3)} = 1.9$	$F_{(2,3)} = .31$
p value	$p < .001$*	$p < .05$*	$p < .001$*	$p = .289$	$p = .756$

The ANOVAs showed that contraction type had significant effects on duration, F1 size and F1 slope. Post hoc (LSD) analysis of *duration* showed that all three contraction types were significantly different from each other ([NC > Semi, *Sig.* = .000; [NC > Cntr], *Sig.* = .000; [Semi > Cntr], *Sig.* = .008). Post hoc (LSD) analysis of *formant displacement size* showed that F1 size of the non-contracted units was significantly larger than that of semi-contracted and contracted units but the difference was insignificant between semi-contracted and contracted (F1 size:

[NC > Semi], *Sig*. = .008; [NC > Cntr], *Sig*. = .006). Post hoc (LSD) analysis of *regression slope* indicated that F1 slopes increased along with the degree of reduction (F1 slope: [NC < Semi], *Sig*. = .003; [NC < Cntr], *Sig*. = .000; [Semi < Cntr], *Sig*. = .003). The same kind of variation was not significant for F2 in terms of its formant displacement size and regression slope.

In summary, the above results show that for F1, the displacement of both semi-contracted and contracted units was smaller than that of the non-contracted units. However, the slope of regression of peak velocity over amplitude was greater in contracted than in non-contracted units, indicating that articulatory effort is at least not reduced as the level of contraction increases. For F2, although the difference followed the same trend as F1, no difference was statistically significant across the different contraction types, again indicating no reduction in articulatory effort for increased levels of contraction.

2.2.3 Study 2b – Corpus data

As with Table 2.7 for Study 2a, Table 2.8 lists a set of ANOVA results for corpus data – contraction type as the independent variable and duration, displacement and slope of the regression line (peak velocity over movement amplitude) as dependent variables. All the dependent variables were averaged in respect to four selected high frequency words, namely *zheyang, zheyangzi, nayang, nayangzi*.

Table 2.8: Mean duration (ms), formant displacement size (st) and slope of the regression line of peak formant velocity over formant movement amplitude of the three contraction types – Study 2b, corpus data.

Type	Duration	F1 size	F1 slope	F2 size	F2 slope
Non-	359.1	31.15	20.568	15.32	23.04
Semi-	230.8	23.00	25.99	8.99	23.37
Contracted	221.2	23.22	33.44	11.27	25.58
F value	$F_{(2,8)} = 11.31$	$F_{(2,4)} = 1.45$	$F_{(2,7)} = 3.03$	$F_{(2,1)} = .68$	$F_{(2,7)} = .21$
p value	$p = .005$*	$p = .336$	$p = .113$	$p = .651$	$p = .819$

Only the difference in duration was significant, showing a progressive decrease from non-contracted to contracted units. Post hoc (LSD) analysis of *duration* showed that non-contracted units were significantly different from both semi-contracted and contracted units (Duration: [NC > Semi], *Sig.* = .006; [NC > C], *Sig.* = .002). Planned comparisons of slope and displacement size indicated that only F1 slope of contracted was marginally steeper than that of non-contracted (F1 slope: [NC < C], *Sig.* = .043). F2 slope and F2 displacement showed no significant difference across contraction types.

Comparing these results to those of Study 2a (*cf.* Table 2.7 and its respective post hoc analysis results), similar patterns can be seen, in particular duration is significantly different across contraction levels, and the slope of regression of peak velocity over amplitude is either similar across contraction levels, or steeper in the contracted tokens than in the non-contracted ones in both nonsense (Study 2a) and high frequency words (Study 2b).

2.3 General discussion and conclusions

In this chapter, we tested two predictions based on the general hypothesis that *the direct cause of extreme reduction is time pressure*. The first prediction tested was that *extreme reduction such as contraction can occur in nonsense words if time pressure is sufficiently high*. This prediction was strongly supported through analyses of experimental data. It transpired that, eliciting contraction from nonsense words from all subjects was easily achieved by having them speak at a fast rate. A highly consistent relationship between contraction rate and speech rate was found (Figure 2.11). Further analyses showed that contraction rate was closely related to time pressure: The shorter the duration, the more likely extreme reduction was to occur. In other words, a very short duration constitutes a sufficient condition for extreme reduction to occur. The positive relation between reduction and time pressure is seen more clearly in Figures 2.14 and 2.16. Additionally, a critical duration of approximately 200 ms (without the duration of C_1) was found for disyllables to virtually lose their intervocalic consonants.

The second prediction, that *articulatory effort is not decreased when contraction of nonsense words occurs*, was also supported. Data suggested that contraction was not accompanied by a decrease in articulatory effort because the slope of the regression line between peak velocity and movement amplitude was no shallower, in fact the slope was often steeper in tokens labelled as contracted compared to those labelled as semi- or non-contracted. That articulatory effort seems to be increased in reduced nonsense words is in partial support of H&H theory's

prediction that greater peak velocity may be applied so as to offset the effect of time pressure. But the fact that clear duration dependency was observed suggests that the effect of the compensation is sometimes not sufficient and thus Lindblom's (1963) earlier and simpler duration-dependent undershoot model seems to be largely supported. What is shown more clearly by the data presented here than that presented in Lindblom (1963) is that when duration is severely shortened, e.g., by half, as seen in Table 2.7, there is simply no way for speakers to maintain the integrity of a segment or a syllable. As a result, the intervocalic consonants are not just reduced, but shortened to the point when no vocal tract closure can be achieved. Target undershoot due to time pressure is thus *inevitable* when the time allocated to a segment is less than its minimum duration, as predicted by Klatt (1973, 1976).

2.3.1 Structural complexity and contraction rate

The examination of syllable sequences with different intervocalic obstruents in Study 1 allowed us to make observations regarding the minimum time required to execute articulatory gestures. In general, when greater time pressure is present, sequences in which the intervocalic consonants are allocated more time in their canonical forms are less likely to exhibit extreme reduction. For example, aspirated obstruents are less frequently reduced than their unaspirated counterparts. This could be attributed to the greater articulatory complexity of aspirated sounds in comparison to unaspirated ones (Tseng, 2005a). However, this is also consistent with a time pressure account: Segments with greater allocated duration

in their canonical forms, e.g., aspirated stops, are less likely to be severely reduced because their allocated duration is less likely to be shortened to the point where no closure of any kind is possible. This account is supported by the fact that semi-contracted units exhibit a constant duration at which intervocalic consonants start to get 'lost' (Figure 2.13). Further evidence in support of this is seen in the case of CVN+CV, where N and second C are both alveolar and hence involve minimum articulatory movements as far as the tongue is concerned. These sequences nevertheless exhibited the smallest degrees of reduction. An even more direct test would be to look at geminates, which were not included in the present study (e.g. /an/+/na/). Geminates would double the allocated articulation time without doubling articulatory demand. This is a possible topic for future research.

In Table 2.6, for the nasal-consonant group, the onset vs. coda asymmetry in reduction rates for CVN+V (62.35%) vs. CV+NV (21.67%) is noticeable. This agrees with the well-known fact that languages prefer CV structures more than VC structures (Levelt et al., 1999, McCarthy, 2007; Hall, 2010). Additionally, in several reduced cases of the nasal-consonant group, some remaining nasal features are still detectable (as can be seen in Figure 2.2). This observation is also related to the description in Cheng (2004) that nasality becomes an overlay of vowels in contracted syllables in rapid speech. This indicates that phonetic reduction is irrespective of segmentability, a dynamic articulatory process (Niebuhr and Kohler, 2011).

2.3.2 Direct versus indirect mechanisms of extreme reduction

The present finding that time pressure has a clear role in reduction does of course not rule out the possible contributions of other factors; such as lexical frequency (Bybee, 2002; Myers and Li, 2009), information load (Karlgren, 1962), listener considerations (Lindblom, 1990), speech style (Dankovicŏvá and Nolan, 1999), semantic relatedness to the discourse topic (Gregory et al., 1999) and repetition time in the same discourse (ibid, see also Fowler and Housum, 1987), etc. Instead, what the present results demonstrate is that time pressure is likely to be the most *direct* cause of reduction. This is supported by the close relation between duration and formant displacement shown in Figures 2.14 and 2.16, which is in sharp contrast to the slightly positive relation between lexical frequency and spectral reduction found by Myers and Li (2009, Figures 3 and 5) for Taiwanese Southern Min. Further evidence is seen in studies that have found clear effects of lexical frequency when duration itself is used as a measure of reduction (Jurafsky et al., 2001; Pluymaekers et al., 2005).

Segmental duration, however, is not controlled only by the factors mentioned above. Many other factors have been identified to have significant duration effects, including phrase boundary, stress, within-syllable location, within-word location, within-phrase location, lexical tone, focus and syllable structure (Berkovits, 1994; Dankovičová, 1997; Gahl and Garnsey, 2004; Klatt, 1975, 1976; van Santen, 1994; van Santen and Shih, 2000; Xu, 2009, to cite only a few). A case in point is the contraction of the Mandarin word *jiao ta che* [tɕiao ta̲ tʃʰɤ] ‘bicycle’ into [tɕiao a̲

tʃʰɤ] despite the fact that it is a noun and is not particularly high frequency (Chung, 2006). Chung (2006) points out that the second syllable is easily elided in tri- or tetra syllabic items. Additionally, Chen (2006) shows that the middle syllables of four-syllable words in Mandarin are drastically shortened. Xu and Wang (2009) further demonstrated that tonal reduction in these syllables is directly attributable to shortened duration. These findings may explain why listeners have the impression of reduction in the medial position of tri- or tetra syllabic items.

The time pressure account also does not rule out the effect of fossilized lexicon. It is unquestionable that contracted forms such as *don't, aren't* and *isn't* in English are fossilized. In Mandarin *beng* [pɤŋ] as a contracted form of *bu yong* [pu jioŋ] 'no need to' is even written as a single character ('甭'), indicating that it is supposed to be spoken as a monosyllabic rather than disyllabic word ('不用'). These fossilized forms can remain monosyllabic even when spoken slowly, which is in contrast with the nonsense words examined in this study (which show clear variability with duration). Therefore, there is a need to test each suspected case of fossilized reduction by directly controlling duration in future research.

Finally, given the present finding that the slope of peak velocity over movement amplitude is often negatively rather than positively related to duration, it is difficult to maintain that (non-fossilized) reduction is related to weakened articulatory effort. Instead, the strong duration dependency of reduction found in the present study actually suggests that it is more likely that duration is used to control the level of target attainment. That is, items that need to be uttered clearly

are given more time, so that their targets are more fully attained; those that do not need to be uttered clearly are given less time, often to the extent of being allowed to go below the minimum duration, resulting in severe reduction. The indirect relationship between information load and reduction can be further seen in the similarity of the results of Study1, 2a and 2b. The target sequences in Studies 1 and 2a are nonsense words (and thus of high information load), while the target words in the corpus data (Study 2b) are highly frequency words (and thus of low information load). Despite this, very similar reduction patterns are seen between these data sets when duration is used as the control factor. Therefore, the results with respect to the durational account are very much the same, giving more weight to the evidence in support of our main hypothesis.

2.3.3 Conclusion

To the best of our knowledge, no prior research has systematically examined variations in spectral patterns in relation to possible articulatory mechanisms underlying extreme reduction in any language. In the present study, contraction was successfully elicited from speakers in the laboratory reciting nonsense disyllabic words at high speech rates, and the rate of contraction as a function of time pressure was found to be similar to that of high frequency words in a number of spontaneous speech corpora. This indicates that extreme reduction is neither a characteristic of only casual speech nor directly related to lexical frequency. For both experimental and spontaneous data, spectral analyses show that extreme reduction regularly occurs when segmental duration is shortened beyond a certain

threshold, and regression analyses of peak velocity of formant movement as a function of formant movement amplitude suggest that articulatory effort is not weakened when contraction occurs. We thus interpret our results as evidence that the direct cause of extreme reduction is target undershoot under time pressure (Lindblom, 1963), while other factors mainly contribute through their effects on duration.

Chapter 3

Tonal reduction

In Chapter 2 it was shown that a severe form of segmental reduction, known as syllable contraction, can occur with nonsense words in Taiwan Mandarin if sufficient time pressure is exerted. It was also shown that time pressure is a direct cause of syllable contraction. The present chapter investigates tonal reductions that occur together with syllable contraction in Taiwan Mandarin. The hypothesis tested is that the nature of the tonal reduction can also be explained by time pressure. In addition to the two predictions tested in Chapter 2, a third prediction is also tested: *Speakers still attempt to approach each and every underlying tonal target under high time pressure.* For this third prediction, tone shapes of contracted syllables are also checked against the Edge-in model (Yip, 1988), which concerns the underlying target formation when extreme reduction such as contraction occurs.

3.1 Methodology

3.1.1 Study 3

The method used in Studies 1 and 2a was again employed to elicit syllable contraction, that is, subjects were asked to produce the target sentences at three speech rates, slow, normal and fast. Similarly to Studies 2a and 2b (Chapter 2), articulatory effort was assessed by measuring the slope of the regression line of F_0 peak velocity over movement amplitude.

A. Stimuli

Disyllabic /ma/+/ma/ nonsense sequences with a total of 16 tone dyads (4 tones x 4 tones) embedded in two carrier sentences were constructed as testing materials. To observe continuous F_0 contours and facilitate segmentation, the target tone-bearing syllables were /ma/+/ma/, written as '媽', '麻', '馬', '罵' in traditional Chinese characters for High (H), Rising (R), Low (L), and Falling (F) tone carriers respectively. To create different tonal contexts for the target sequences, two carrier sentences with an H or an L tone preceding the target sequence were composed (see Table 3.1). The tone following the target sequence is always H. The reason for not varying this following tone is that previous research has shown that contextual tonal variations are predominantly due to carryover effects (Gandour et al., 1994; Xu, 1997). Each carrier sentence consists of three phrases. The first phrase contains 9 underlying syllables, the second 13, and the third 17. The same target sequence was embedded in each phrase and thus was produced three times within each carrier sentence.

Table 3.1: Carrier sentences used in Study 3.

Carrier sentences with a *H*igh/*L*ow preceding tone	
Characters	你想*吃/買*____沙拉是吧！我當然不*吃/買*____沙拉那種東西，因為我不喜*歡/欣賞*____沙拉那種酸酸的醬料！
Pinyin	ni xiang chi*H*/mai*L* ____ sha*H*la shi ba! wo dangran bu chi*H*/maiL ____ sha*H*la nazhong dongxi, yinwei wo bu xihuan*H*/xinshang*L* ____ sha*H*la nazhong suansuande jiangliao!
English	You want to eat/buy ____ salad, didn't you! Of course I won't eat/buy ____ salad that kinda stuff, because I dislike the sour source of ____ salad.

B. Subjects and recording procedure

The same subjects from Study 2a were recruited to participate in this experiment. The same procedure was followed, that is, six subjects were instructed to recite the sentences at three speaking rates, slow and clear as if reciting in class, in a natural manner as if conversing with a friend, and as fast as possible. The mean speech rates of the slow, natural and fast conditions across the six subjects were 4.5, 6.1 and 9.3 underlying syllables per second, respectively. Three repetitions of the entire block were recorded, each with a different randomization order. In total, the number of target sequences produced in this experiment was 16 (tone dyads) × 2 (preceding tones) × 3 (positions in the carrier) × 6 (subjects) × 3 (speech rates) × 3 (blocks) ＝5,184. Out of these, 14 (0.2 %) were discarded from further analysis due to inadequate voice quality such as creaky voice or speaker errors.

C. Segmentation and measurements

All sound files were manually segmented by the author. First, all tokens of the target sequence /ma/+/ma/ were isolated and then a boundary between the two syllables was marked whenever a clear second nasal was identified and the token was labelled as *non-contracted.* When the second nasal was absent, no boundary was marked and the token was labelled as *contracted*. Intermediate cases were labelled as *semi-contracted* with two intervals. Examples of this labelling are shown in Figures 3.1-3.4. Consistency of the labelling was double checked one month following the initial labelling. A small amount of tokens were relabelled from non-contracted or contracted to semi-contracted upon rechecking, but no non-contracted tokens were relabelled as contracted or vice versa.

Extraction of the F_0 contours was carried out using a modified version of ProsodyPro, a general purpose Praat script for large scale F_0 analysis (Xu, 2005-2011). The script extracts F_0 by displaying the vocal cycle marking generated by the Praat programme (Boersma and Weenink, 2010) and allowing users to perform manual rectifications through adding missing vocal pulse marks and removing redundant marks. The script then converts the vocal periods into F_0 values and applies a trimming algorithm (Xu, 1999) to remove sudden jumps and dips. It then generates various output data for further analysis. One output the script produces is time-normalized F_0 contours which are then analysed graphically (see Figures 3.10-3.13 for example). In order to compare contracted or non-contracted tokens, time-normalisation was conducted as follows. For tokens marked with two intervals, that is, classified as non-contracted and semi-

contracted where the intervocalic nasal was present, each interval was discretised into 20 evenly spaced points. For contracted tokens where the intervocalic nasal was absent, 40 evenly spaced points were taken within the single interval. This process allows for averaging across repetitions and speakers and allows direct comparisons of different contraction types.

To assess articulatory effort, measurements were taken from two unidirectional F_0 movements (rising and falling). Because increased speech rate could lead to 'flattened' F_0 contours, making the detection of F_0 turning points impossible, measurements were taken from only tone dyads where a unidirectional rising or falling movement could be guaranteed. Tonal contexts (H)#*R*F and (H)#*R*L (where # delimits a preceding H tone and the target tone dyads) were cases where a unidirectional *rising* movement could be identified. In (H)#*R*F (see Figure 3.1), the trajectory of F_0 needs to first move down from the preceding H ending tone (toward the minimum) in order to realise a rising movement for the R tone in the first syllable. The presence of a maximum could be further guaranteed by the required falling movement of the F tone in the second syllable. The modified ProsodyPro script automatically located the key measurement points. It first located the F_0 minimum ('min') in the early part of /mama/. It then searched for the F_0 maximum ('max') between the 'min' point and the end of /mama/. In-between 'min' and 'max', the velocity of this rising movement was calculated and the location of peak velocity was identified ('v'). Figures 3.1 and 3.2 show the measurement points for a *rising* movement of the R tone in (H)#*R*F and (H)#*R*L, respectively. Similar measurements were also taken from the (L)#*F*R and (L)#*F*H

sequences for a unidirectional *falling* movement (see Figures 3.3 and 3.4). In total, 383 out of 648 tokens from the four selected tone sets were found to be valid for assessing articulatory effort (a success rate of 59.1%). This low success rate is indication that the rate of articulation of the speakers had indeed been pushed to the limit, as will be seen more clearly in Sec. 3.2.1.E.

Other possible tonal contexts such as (H)#*R*R for a *rising* movement and (L)#*F*F for a *falling* movement were also considered but later excluded from the analysis. This is because in cases of severe reduction a maximum in the second R in an (H)#RR sequence was less guaranteed, and neither was a minimum in the second F in an (L)#FF sequence. This is owing to the high articulatory demand resulting from consecutive dynamic tones, i.e. RR or FF, in an incompatible tonal environment (Kuo et al., 2007; Xu and Wang, 2009).

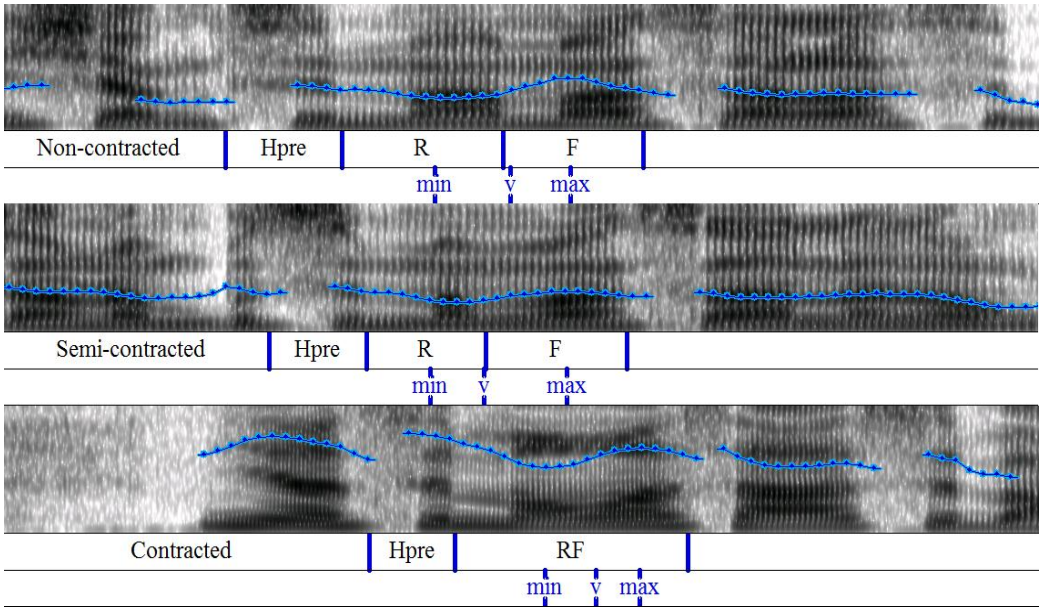


Figure 3.1: Rising movement and labelling examples in cases of H#*R*F. The time domains are of similar window length from 0 to 1 second and formant frequency from 0 to 5000 Hz. F_0 values are shown as dots overlaid on the spectrograms, scaling from 50 to 300 Hz.

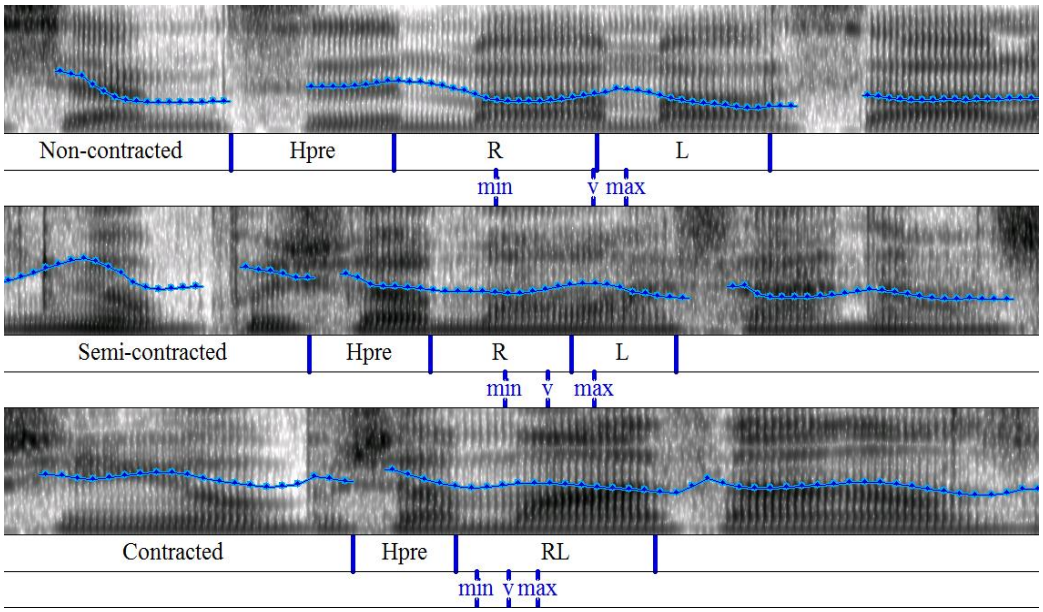


Figure 3.2: Rising movement and labelling examples in cases of H#*R*L. The time domains are of similar window length from 0 to 1 second and formant frequency from 0 to 5000 Hz. F_0 values are shown as dots overlaid on the spectrograms, scaling from 50 to 300 Hz.

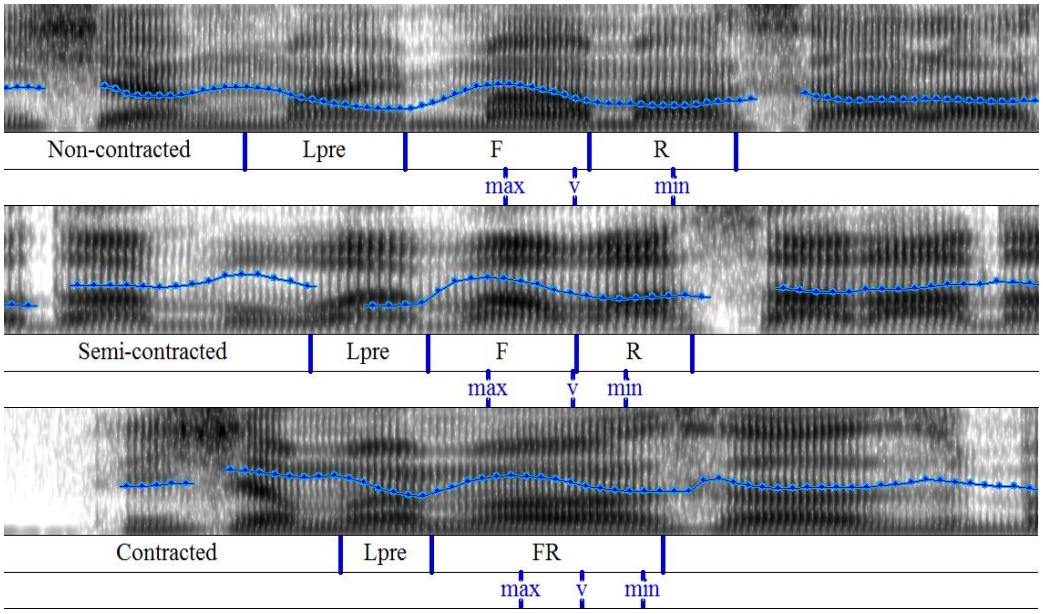


Figure 3.3: Falling movement and labelling examples in cases of L#*F*R. The time domains are of similar window length from 0 to 1 second and formant frequency from 0 to 5000 Hz. F_0 values are shown as dots overlaid on the spectrograms, scaling from 50 to 300 Hz.

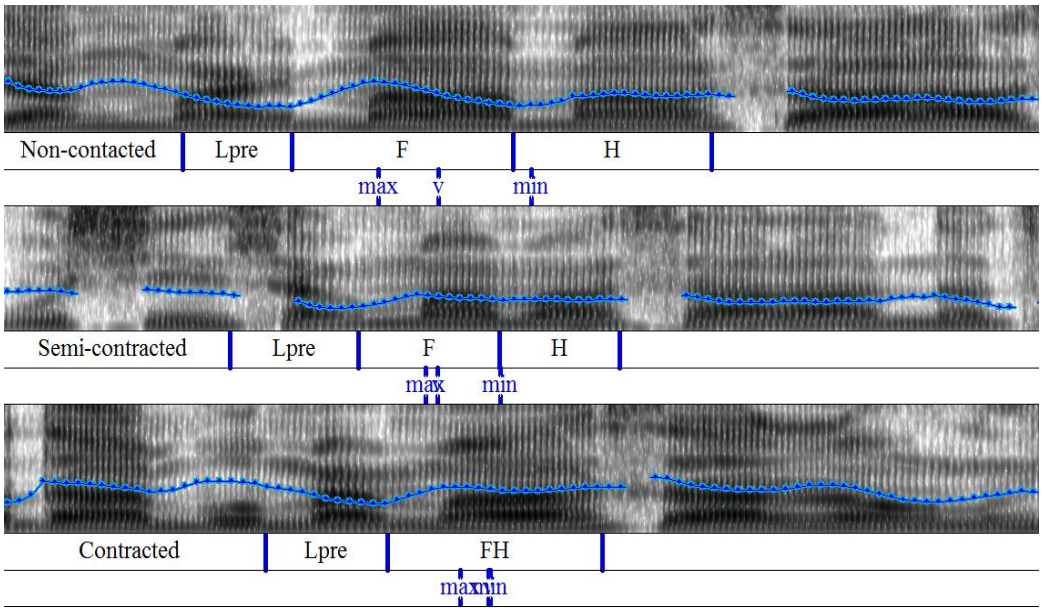


Figure 3.4: Falling movement and labelling examples in cases of L#*F*H. The time domains are of similar window length from 0 to 1 second and formant frequency from 0 to 5000 Hz. F_0 values are shown as dots overlaid on the spectrograms, scaling from 50 to 300 Hz.

To assess articulatory effort, a similar approach to the analysis in Chapter 2 was followed. That is, three kinematic measurements were taken for each unidirectional movement: (1) F_0 movement duration – time difference between adjacent max F_0 and min F_0 in seconds, (2) F_0 movement amplitude – F_0 difference between adjacent max F_0 and min F_0 in semitones, and (3) F_0 peak velocity – maximum absolute value in the first derivative of a unidirectional F_0 movement, in semitones/second. Figure 3.5 displays an overall scatter plot of F_0 peak velocity as a function of F_0 movement amplitude for all selected rising and falling movements. The relationship between F_0 peak velocity and F_0 movement amplitude was highly linear ($r = .743$, $p < .001$), which is consistent with the formant analyses in Chapter 2 and movement analyses of previous studies (acoustic movements: Xu and Su, 2002; Xu and Wang, 2009; articulatory movements: Hertich and Ackermann, 1997; Kelso et al., 1985; Ostry and Munhall, 1985; Vatikiotis-Bateson and Kelso, 1993). Given the highly linear relationship between F_0 peak velocity and F_0 movement amplitude; this experiment also used the slope (i.e. gradient) of their regression line to assess articulatory effort.

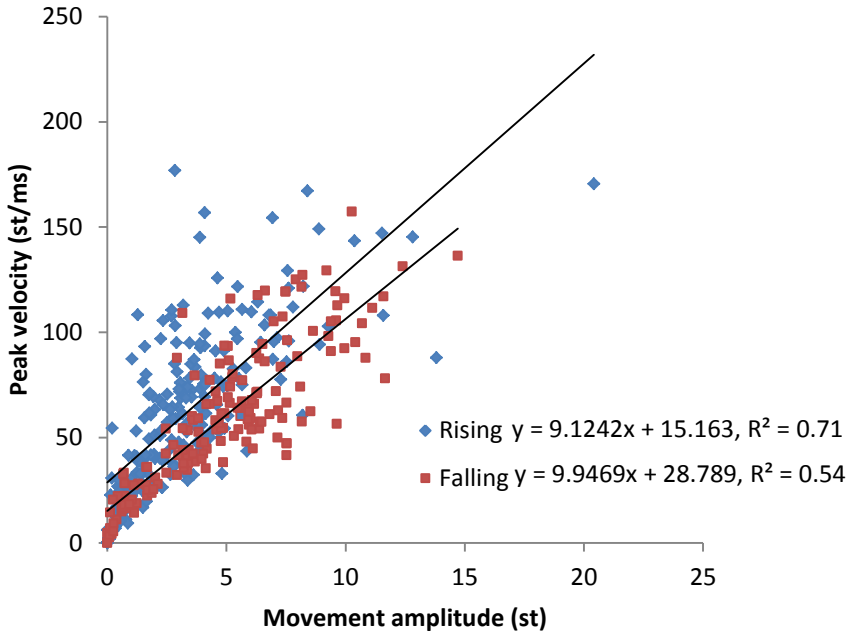


Figure 3.5: Linear regressions of F_0 peak velocity (y-axis in semitones/seconds) over F_0 movement amplitude (x-axis in semitones) for both rising and falling movements (in absolute values of peak velocity). In total, 216 data points were valid for a *rising* movement of the R tone in (H)#*R*F and (H)#*R*L and 167 data points for a *falling* movement of the F tone in (L)#*F*R and (L)#*F*H. The valid data points were distributed across all subjects and conditions.

3.1.2. Edge-in model

To explain the mechanism of tonal contraction, a phonological account known as the Edge-in model has been proposed (Yip, 1988). The model, which presupposes two successive pitch targets for each tone, hypothesises a phonological process that operates in an outside-in fashion during tonal contraction in a manner such that the two adjacent targets in a disyllabic sequence are suppressed, leaving only the two targets on the outer edges intact. Figure 3.6 shows an example of the proposed Edge-in process.

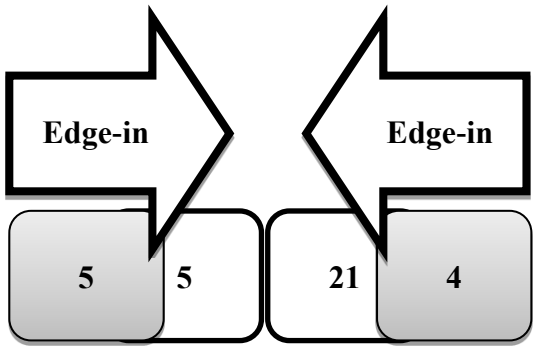


Figure 3.6: An Edge-in model for deriving the output tone 54 from two source syllables, [kən55] + [pən214] → [kəm54], meaning 'basically'. The bilabial plosive /p/ gives rise to a realisation of coda /m/.

This model offers a simple and efficient mechanism of generating possible patterns of phonetic reduction. However, given that it is based mostly on transcribed data and a somewhat ad-hoc formalisation, one may wonder whether a phonological grammar such as this truly reflects what happens in contraction. Recent experimental studies have shown that exceptions to this formal generalisation are not uncommon and that traditional phonology cannot straightforwardly account for the graded nature of reduction (Cheng, 2004; Myers and Li, 2009; Zhang and Lai, 2010). In the analysis that follows, the Edge-in model will be tested against the third prediction of this thesis, namely, *speakers still attempt to approach each and every underlying tonal target when contraction occurs*.

3.2 Analysis and results

3.2.1 Time pressure and articulatory effort – Predictions 1 and 2

Two predictions were tested to see whether tonal reduction can also be explained by time pressure. These two predictions are parallel to those tested in Chapter 2: (1) *extreme reduction such as contraction can occur in nonsense words if time pressure is sufficiently high;* and (2) *when contraction occurs, articulatory effort is not decreased.*

A. Contingency of contraction type

Figure 3.7 displays distributions of the three contraction types obtained in Study 3. Non-contracted occurred most frequently, taking up more than half the cases of this experiment (66.98%), followed by contracted (21.53%). The semi-contracted showed the least frequency of 11.49%.

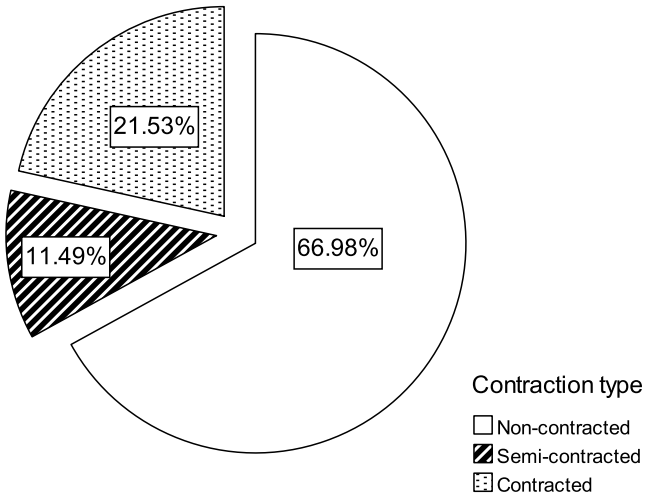


Figure 3.7: Distribution of the three contraction types in Study 3.

Table 3.2 shows the percentage of each contraction type for different tone combinations. Tone dyads preceded by an L tone generally showed a greater percentage of the contracted type (than those preceded by an H tone), on average from 28.07% (L#) to 16.05% (H#).

Table 3.2: Percentage of each contraction type occurred across different tone combinations. Column represents the tones (H, R, L, F) in the first syllable and row the second syllable.

2nd sylb. / 1st sylb.		H		R		L		F	
		(H)#	(L)#	(H)#	(L)#	(H)#	(L)#	(H)#	(L)#
H	NC	73.46%	62.11%	78.62%	62.35%	76.54%	63.58%	80.86%	63.52%
	Semi	10.49%	11.80%	8.18%	12.35%	11.11%	12.96%	7.41%	15.72%
	Cntr	**16.05%**	**26.09%**	**13.21%**	**25.31%**	**12.35%**	**23.46%**	**11.73%**	**20.75%**
R	NC	66.67%	60.49%	64.42%	56.17%	73.46%	56.79%	75.31%	56.79%
	Semi	11.73%	11.73%	10.43%	10.49%	9.26%	10.49%	12.35%	14.81%
	Cntr	**21.60%**	**27.78%**	**25.15%**	**33.33%**	**17.28%**	**32.72%**	**12.35%**	**28.40%**
L	NC	77.78%	55.26%	74.07%	57.55%	63.29%	50.70%	75.31%	57.98%
	Semi	8.64%	14.04%	12.96%	10.38%	14.56%	12.68%	6.17%	13.45%
	Cntr	**13.58%**	**30.70%**	**12.96%**	**32.08%**	**22.15%**	**36.62%**	**18.52%**	**28.57%**
F	NC	75.93%	67.90%	67.28%	58.39%	63.58%	62.96%	75.31%	57.41%
	Semi	9.88%	7.41%	17.28%	15.53%	19.75%	12.96%	11.11%	14.20%
	Cntr	**14.20%**	**24.69%**	**15.43%**	**26.09%**	**16.67%**	**24.07%**	**13.58%**	**28.40%**

B. Speed and contraction type

A multinomial logistic regression was performed with contraction type as the ordinal dependent variable and position in the carrier sentence, as well as speed, as predictor variables. Position in the carrier sentence displayed no relation to contraction type (Coef = -0.010, S.E. = 0.041, p = .802). However, speed was positively related to contraction type (Coef = 1.961, S.E. = 0.053, p < .0001). For a unit increase in speed, the expected ordered log odds increased by 1.96 as one moved to the adjacent higher category of contraction type (i.e. from non- to semi- to contracted). To better display this effect, Figure 3.8 shows frequency of occurrence of contraction type according to speaking rate.

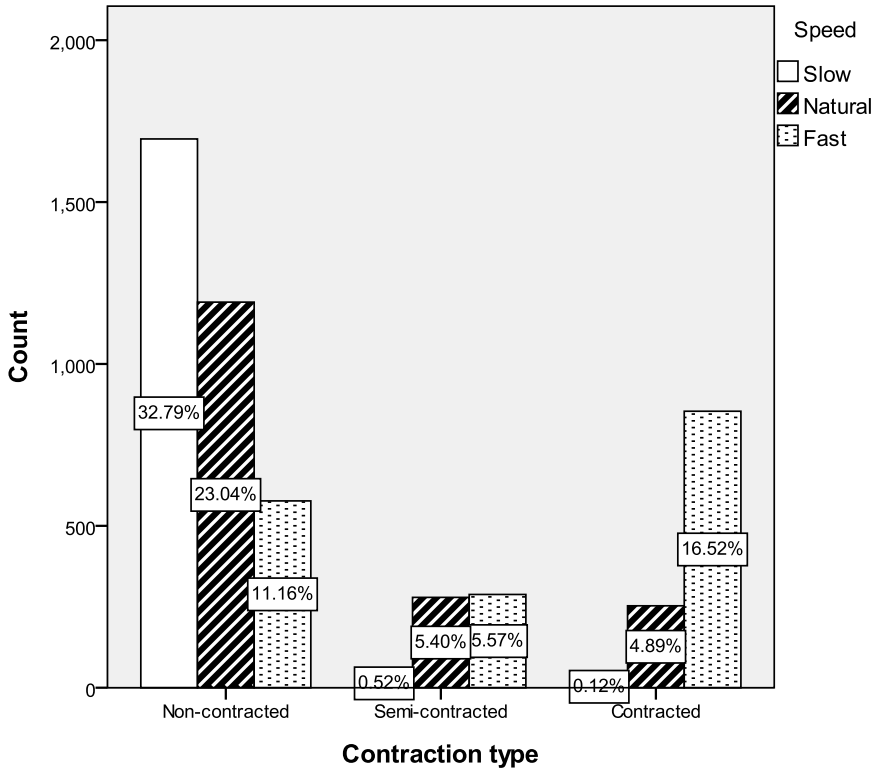


Figure 3.8: Contingency of contraction type at different speeds obtained in Study 3. The x-axis shows three different contraction types and the y-axis shows frequency count.

The frequency of occurrence of non-contracted units decreased as speaking rate increased. Conversely, for both semi-contracted and contracted types, frequency increased with speaking rate. For each contraction type, the highest percentage observed was: 32.79% non-contracted at slow speech rate, 5.57% semi-contracted at fast speech rate, and 16.52% contracted at fast speech rate. This is in agreement with the statistics that also show a positive relationship between speed (from slow to fast) and contraction type (from non-contracted to contracted). The distribution of contraction types shown here are consistent with the results of Study 1 (Chapter 2).

C. Duration and excursion size

To examine the relationship between duration and excursion size, a Pearson correlation test (2-tailed) was carried out between duration and F_0 excursion size. The correlation was significant (r = .464, p < .001). Further, two one-way ANOVAs were conducted using contraction type as independent variable and duration and F_0 as dependent variables. Table 3.3 shows the results of the statistical analysis.

Table 3.3: Mean duration (ms) and F_0 excursion size (st) of the three contraction types – All tone dyads.

Contraction type	Duration	F_0 size
Non-contracted	372.6	41.7
Semi-contracted	276.3	31.6
Contracted	236.0	28.6
F value	$F_{(2,93)} = 694.6$	$F_{(2,93)} = 11.3$
p value	$p = .000$*	$p = .000$*

Contraction type had significant effect on both duration and F_0 excursion size. Post hoc (LSD) analysis of *duration* showed that all three contraction types were significantly different from each other (all *Sigs.* < .001). Post hoc (LSD) analysis of *F_0 excursion size* indicated that the non-contracted type had significantly larger movement amplitude in comparison to both semi-contracted and contracted (both *Sigs.* < .001), whereas no significant difference was found between the semi-contracted and contracted types (*Sig*. = .295).

D. Articulatory effort

Similar to the design of Study 2a, the two unidirectional pitch movements (rising and falling) selected from the four tone combinations allowed us to further examine the relationship between duration, F_0 displacement and articulatory effort for the three contraction types. Table 3.4 shows a set of one-way ANOVAs performed with contraction type (non-contracted, semi-contracted and contracted) as the independent variable and duration, F_0 excursion size and slope of the regression line of F_0 peak velocity over F_0 movement amplitude as dependent variables. All dependent variables were averaged with respect to the selected (H)#*R*F, (H)#*R*L, (L)#*F*R and (L)#*F*H items.

Table 3.4: Mean duration (ms), F_0 excursion size (st) and slope of the regression line of F_0 peak velocity over F_0 movement amplitude of the three contraction types – Study 3.

Contraction type	Duration	F_0 size	Slope of F_0 peak velocity over amplitude
Non-	383.3	42.6	8.5
Semi-	278.1	32.8	12.2
Contracted	239.7	28.1	12.4
F value	$F_{(2,9)} = 73.5$	$F_{(2,9)} = 8.5$	$F_{(2,9)} = 6.1$
p value	$p = .000$*	$p = .008$*	$p = .021$*

The ANOVAs showed that contraction type had significant effects on all three dependent variables. Post hoc (LSD) analysis of *duration* again showed that all contraction types were significantly different from each other (Duration: [NC > Semi], *Sig.* = .000; [NC > C], *Sig.* = .000; [Semi > Cntr], *Sig.* = .012). Post hoc (LSD) analyses of *F_0 excursion size* and *regression slope* indicated that significant

differences existed between all but the semi- and contracted types (F_0 size: [NC > Semi], *Sig.* = .023; [NC > C], *Sig.* = .003; [Semi > Cntr], *Sig.* = .221; F_0 slope: [NC > Semi], *Sig.* = .033; [NC > C], *Sig.* = .027; [Semi > Cntr], *Sig.* = .899).

In summary, the degree of reduction, as reflected by contraction type, was negatively related to duration and F_0 excursion size but positively related to the slope of the regression line, indicating that there is an increase in effort from non-contracted to semi-contracted to contracted. The higher level of contraction with an increased regression slope suggests that the duration-dependent undershoot cannot be fully offset by effort. The results of Study 3 have thus far largely been in agreement with those presented in Chapter 2.

E. Maximum speed of pitch change

In view of the insufficient compensation from an increased articulatory effort for items of limited duration (Table 3.4), it appears that the speakers may have reached their physiological limit for changing pitch within a reduced duration, in particular when duration was as short as that of contracted syllables. Therefore, it may also be helpful to assess whether speakers do indeed approach their maximum speed of pitch change. According to Xu and Sun (2002), the minimum amount of time required to raise or lower pitch at the maximum speed of voluntary pitch change obeys a quasi-linear relationship with the amplitude of F_0 movement, which can be approximated by the following two formulae:

$T = 100.4 + 5.8\ d$ (pitch lowering), (3.1)

$T = 89.6 + 8.7\ d$ (pitch raising), (3.2)

where T is the minimum movement time in milliseconds and d is the F_0 movement amplitude in semitones. Equations 3.1 and 3.2 were applied to the current data. Figure 3.9 shows the observed duration and the theoretical minimum time needed to generate the same F_0 movement amplitude together with their time differences.

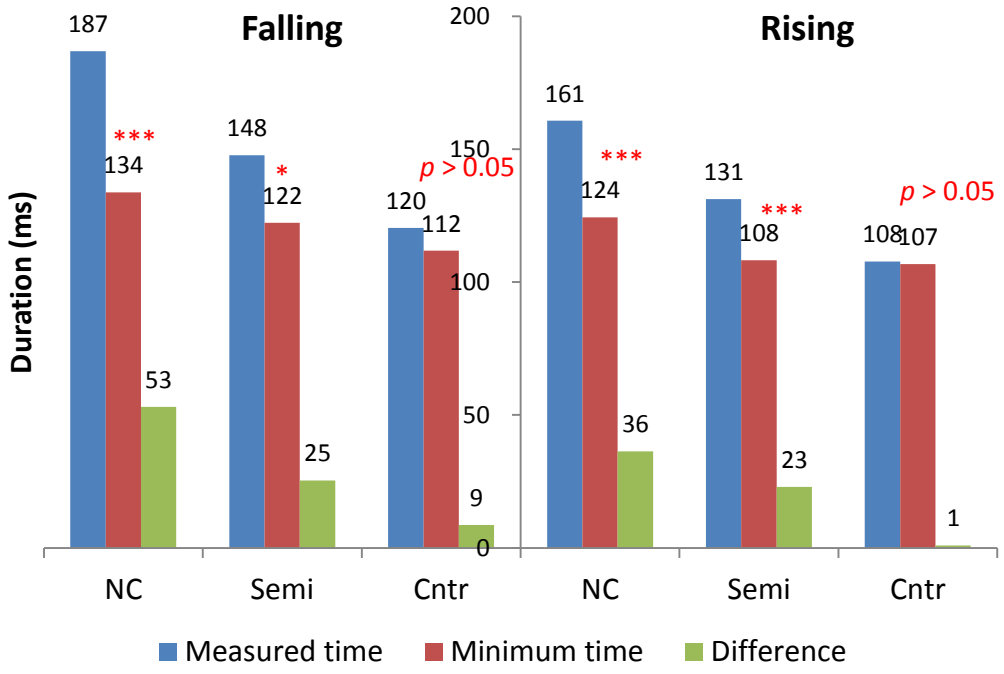


Figure 3.9: Measured time (blue) at different contraction types and movement directions compared to the minimum time (red) required for the same amount of F_0 movement amplitude computed by Equations 3.1 and 3.2. The green bars indicate the differences between these two time intervals. The red asterisks and *p* values indicate the statistical significance as described in the text.

As can be seen in Figure 3.9, the measured time of non-contracted and semi-contracted units were both significantly longer than the corresponding minimum duration according to Welch two sample t-test, indicating that the time interval was ample and that speakers were not required to reach their maximum speed (Falling NC: t = 8.64, df = 136.6, Falling Semi: t = 2.67, df= 23.1, Rising NC: t = 8.39, df = 232.6, and Rising Semi; t = 3.93, df = 47.4). However, in the most severely reduced cases, the times observed were virtually the same as the minimum time needed to execute both the falling and rising movements. This may indicate that speakers had reached their physiological limit of pitch change (Falling Cntr: t = 0.92, df = 17.2 and Rising Cntr: t = 0.09, df= 24.3) and extreme reduction was therefore inevitable.

3.2.2 Evidence of underlying targets in contracted tones – Prediction 3

A. Tonal contours of different contraction types

To see the tone shapes in a straightforward manner, mean F_0 contours of the sixteen tone dyads in different contraction regimes are first displayed in Figures 3.10-3.13. F_0 contours were averaged across three positions within the same sentence and across the three repetitions of the same sentence. These values were then converted to semitones and averaged across the speakers. Note that the tone dyad LL (Figure 3.12c) was always realized as RL due to an obligatory tone sandhi rule in Mandarin Chinese that modifies the first L in a LL sequence to R (Chao, 1968).

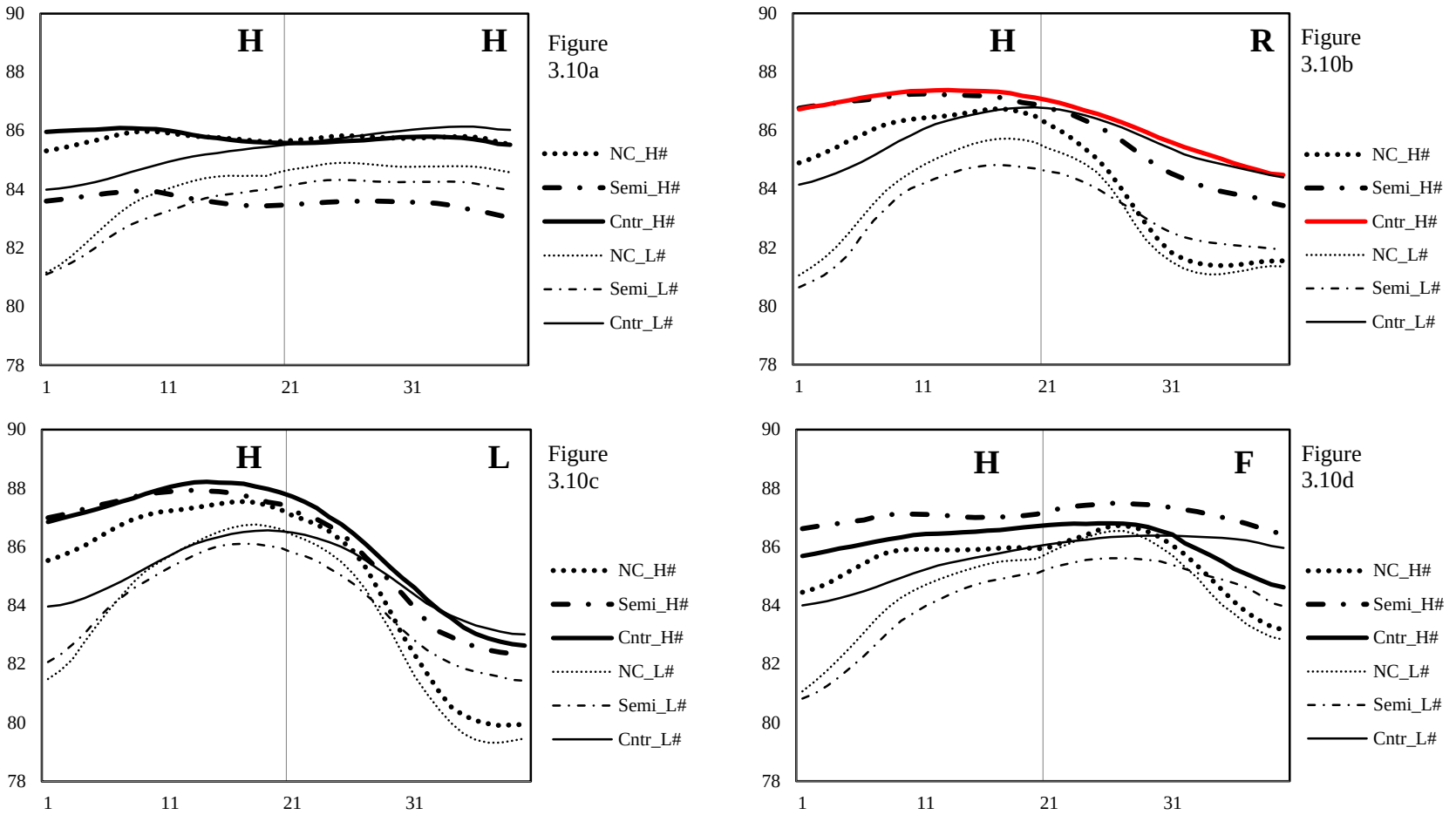


Figure 3.10: F_0 contours of tone dyads HH (a), HR(b), HL(c) and HF(d). Tones preceding the tone dyads are indicated by line thickness and contraction types line style, as shown in the legend. The x-axis is 40 evenly spaced measurement points and the y-axis is in semitones.

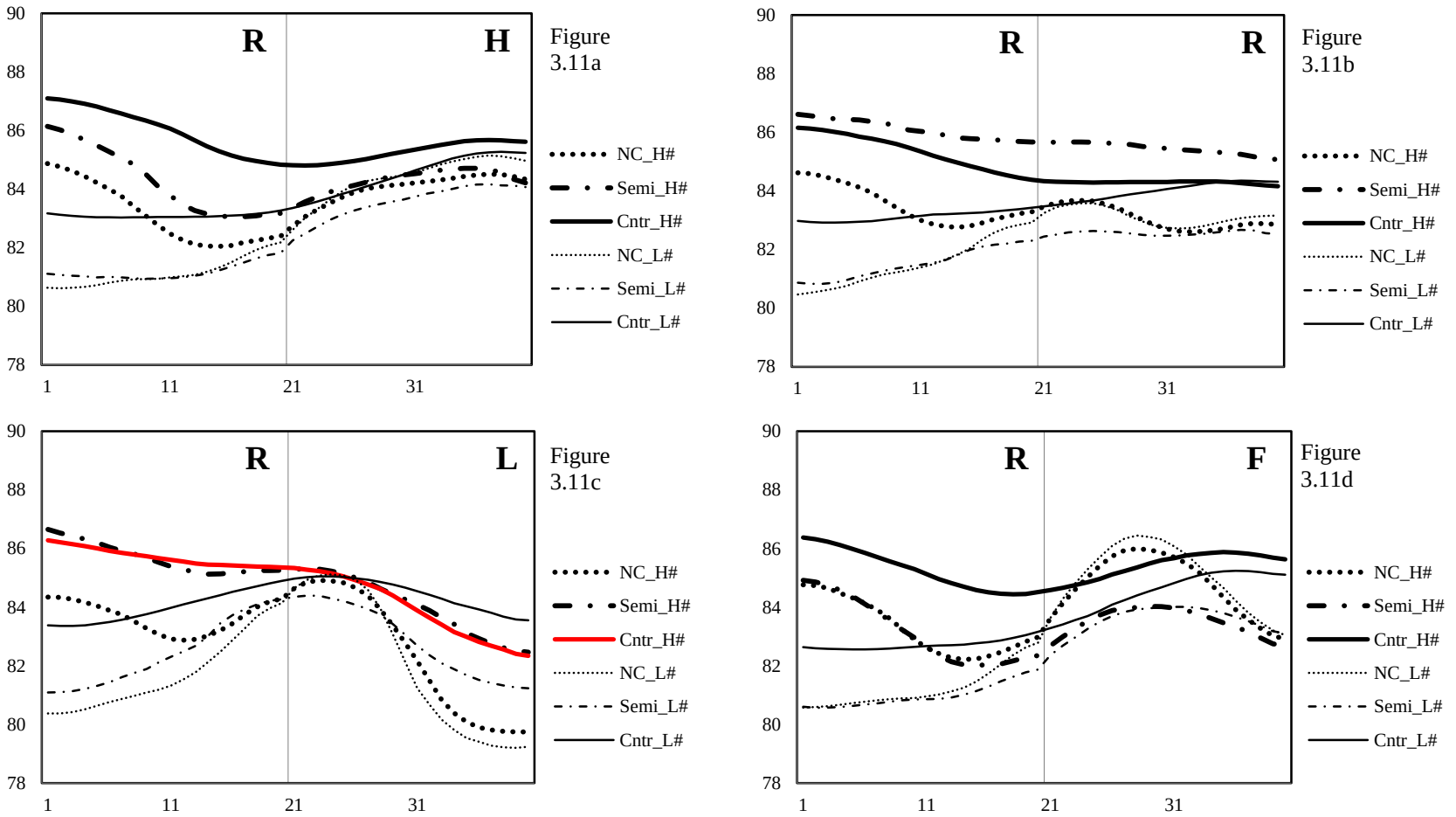


Figure 3.11: F_0 contours of tone dyads RH (a), RR(b), RL(c) and RF(d). Tones preceding the tone dyads are indicated by line thickness and contraction types line style, as shown in the legend. The x-axis is 40 evenly spaced measurement points and the y-axis is in semitones.

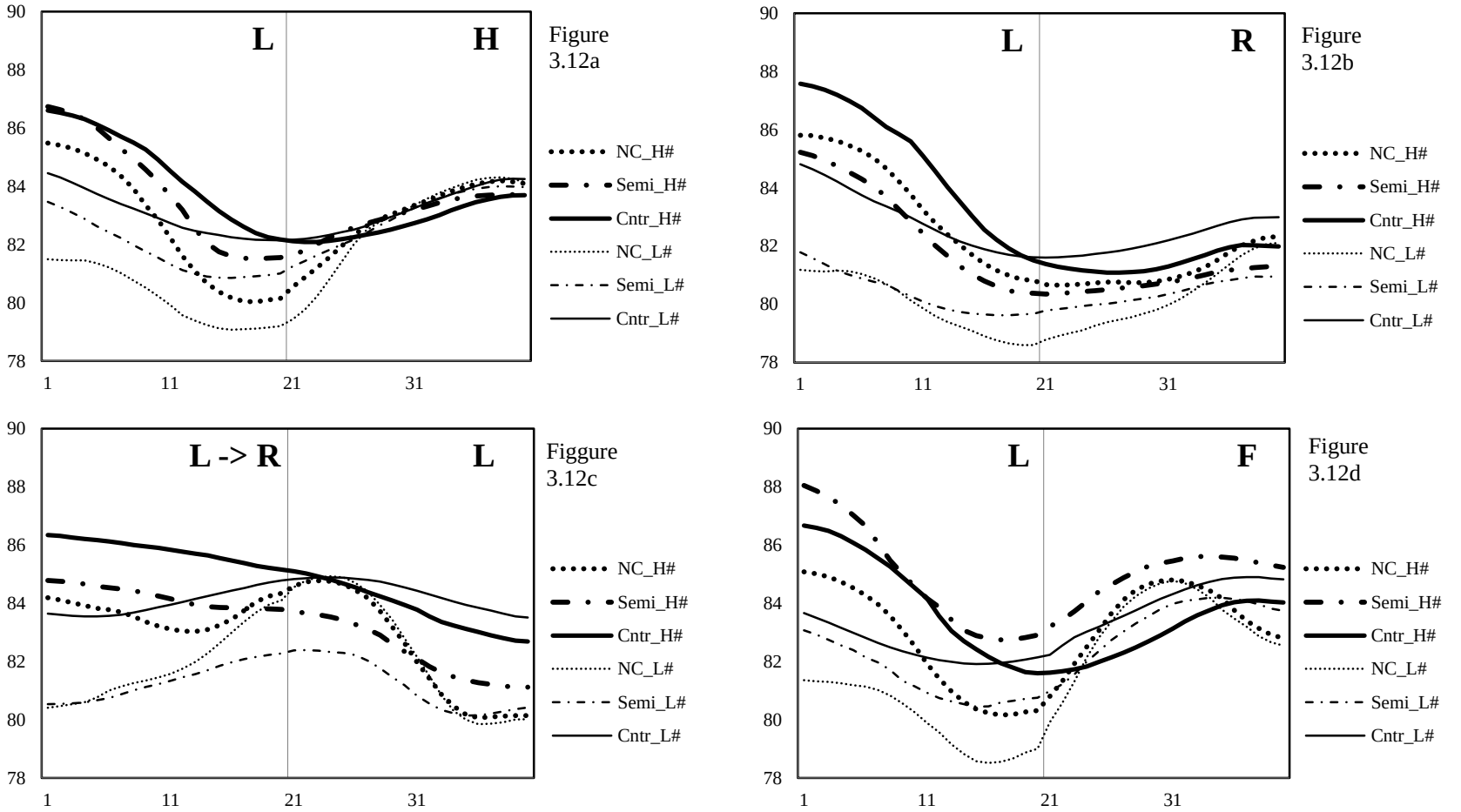


Figure 3.12: F_0 contours of tone dyads LH (a), LR(b), LL -> RL(c) and LF(d). Tones preceding the tone dyads are indicated by line thickness and contraction types line style, as shown in the legend. The x-axis is 40 evenly spaced measurement points and the y-axis is in semitones.

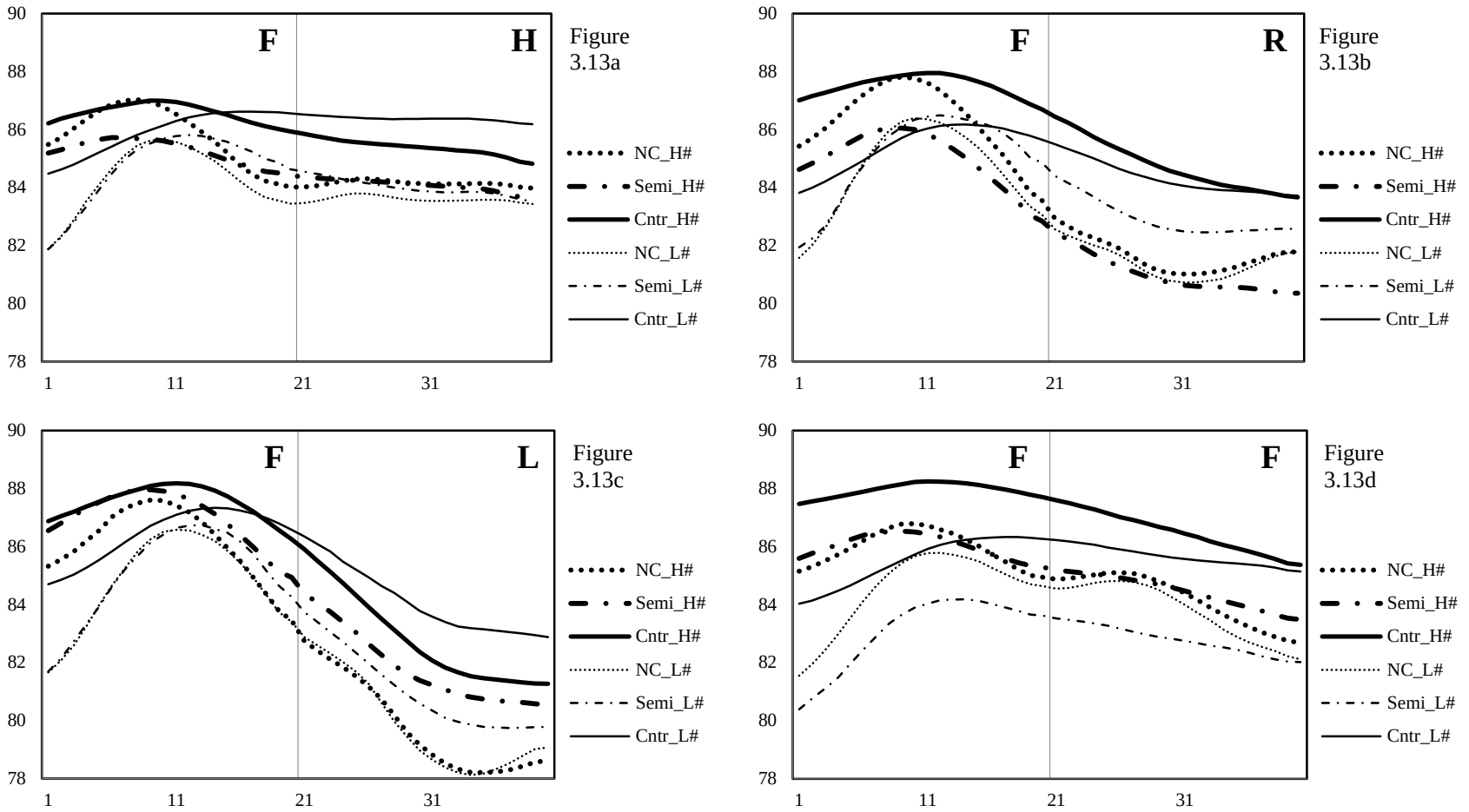


Figure 3.13: F_0 contours of tone dyads FH (a), FR(b), FL(c) and FF(d). Tones preceding the tone dyads are indicated by line thickness and contraction types line style, as shown in the legend. The x-axis is 40 evenly spaced measurement points and the y-axis is in semitones.

Three direct observations can be made from the mean F_0 contours of Figures 3.10-3.13. First, as expected, non-contracted contours have larger pitch ranges than semi-contracted and contracted contours (e.g. RF in Figure 3.11d, among others). Secondly, there is a robust carry-over effect in all sixteen tone dyads. This can be seen by comparing the onset F_0 values of contours with different preceding tones. Those with a preceding H tone were generally higher than those with a preceding L tone (e.g. Figure 3.11). Thirdly, contracted contours display a higher overall F_0 in comparison to non-contracted and semi-contracted contours (e.g. contracted (H)#HR and (L)#HR in Figure 3.10b, contracted (H)#FL and (L)#FL contours in Figure 3.13c). The only three exceptions were (H)#HF (Figure 3.10d), (H)#RR (Figure 3.11b) and (H)#LF (Figure 3.12d), where the semi-contracted F_0 contours were higher than the contracted contours.

In addition to the higher overall F_0, contracted contours are also flatter and more deviant than their non-contracted counterparts. In particular, for the dynamic tone R the critical rising patterns were often absent in contracted conditions. Note that also being a dynamic tone, the falling pattern in F was not as susceptible as the rising pattern in R to reduction. That is, most falling movements in F were still present across different contraction types and tonal environments. Take (H)#H*R* in Figure 3.10b as an example, the final rise in the R tone is missing from the contracted contours. Similarly, in Figure 3.11c little rising movement can be seen in the R tone in (H)#*R*L. In these cases, it is reasonable to ask whether the tonal targets are deleted or modified (as predicted by phonological theories such as the

Edge-in model), or the underlying targets remain unchanged but were not fully realized due to the time constraint. This issue will be examined in the next section.

B. Incompatibility with the predictions of Edge-in model

To investigate whether, during contraction, tonal targets of the corresponding non-contracted units are preserved or whether they are modified via a phonological process as predicted by the Edge-in Model (Yip, 1988), we further examine the F_0 velocity profiles. Such profiles can give a good indication of articulatory movements toward the underlying tonal targets (Gauthier et al., 2007). For simplicity, the analysis presented here was carried out on the tone dyad (H)#H*R* (see Figure 3.14) and (H)#F*F* (see Figure 3.15). Other tone dyad combinations were also analysed and the results were in line with those presented here.[6]

[6] The plots of F_0 velocity profiles of all tone dyads and contraction types are provided in Figures 3.16-3.19. Note that in these figures a consistent 'jerk' is seen towards the end of the second interval. This small sudden fluctuation is probably due to the fact the following syllable in the carrier sentence begins with a voiceless consonant (i.e. /tʂ^h/) that interrupts the continuous F_0 and affects the smoothness of the velocity profiles.

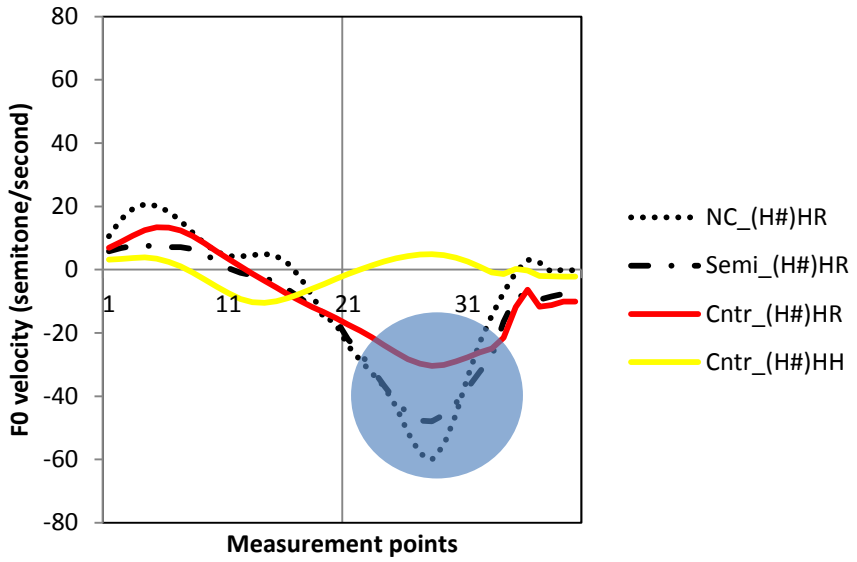


Figure 3.14: Mean F_0 velocity contours of (H#)HR of three contraction types and that of contracted (H#)HH (cf. Figure 3.10a&b, p. 75).

Figure 3.14 shows the aforementioned F_0 velocity contours along with a contracted (H)#HH F_0 velocity contour. As mentioned previously (see Figure 3.10b), during the execution of a canonical HR in the (H)#HR context, the F_0 velocity remains positive and only undergoes comparatively small variations during the production of the first syllable. In the second syllable the velocity then decreases and becomes negative prior to a final rise. The semi-contracted and contracted F_0 velocity contours shown in Figure 3.14 both exhibit this general behaviour but with slightly less variation in the first syllable and a shallower trough in the second. Importantly, the F_0 contour rise towards the end of the second syllable is present in all cases. This trend is interpreted as revealing the underlying intentions of the speakers to target the R tone by approaching the zero-velocity line.

It is also informative to compare a contracted (H)#HR to a contracted (H)#HH tone dyad. For both these tone dyads, the Edge-in model predicts similar surface forms when contraction occurs. That is, when a HR sequence is reduced to a single syllable, the Edge-in model predicts a resulting surface form of (5~~53~~5→55), which is the same as that of a contracted HH sequence (5~~55~~5→55). However, Figure 3.14 shows that a clear difference is present between the contracted forms of (H)#HR and (H)#HH. In contrast to the contracted (H)#HR F_0 contour, during the second interval the contracted (H)#HH contour does not exhibit the 'falling and rising' pattern, but instead displays a small oscillation around the zero velocity level. This again demonstrates that speaker's still attempt to achieve the underlying targets of the non-contracted forms even within a limited duration and that no targets are deleted as proposed by the Edge-in model.

Figure 3.15 shows another example of predictions made based on the Edge-in model being incompatible with the observed results for tone dyads (H)#FF and (H)#FL. Based on the Edge-in model, both contracted tone dyads should be realised as similar falling forms, that is (H)#FF (5~~15~~1→51) and (H)#FL (5~~12~~1→51), i.e. a falling movement from the top to the bottom of a speaker's pitch range. In Figure 3.15, three varying F_0 velocity contours of (H)#FF along with a contracted (H)#FL F_0 velocity contour are shown. In the initial period of the first interval, all F_0 contours further increase their velocity from the previous high-ending H tone and thus they form a preparatory rise for the target F in the first interval. This preparatory rise for a target F is seen again for non-contracted and semi-contracted (H)#FF at around the 21st measurement point. At a similar

point in time, the F_0 velocity of the contracted (H)#FF remains negative following the first F. The velocity hovers around -20 semitones/second while the contracted (H)#FL continues to decrease to a minimum of roughly -70 semitones/second. This stagnant velocity in the contracted (H)#FF may be explained by the fact there is no time for many 'meaningful' oscillations to occur. But it also suggests that the target of the second syllable did not change into that of a L tone. This can be seen more clearly in comparison with the true (H)#FL cases (yellow line), in which the velocity trajectory becomes very negative due to steep fall into the L tone target. This is contrary to the prediction of the Edge-in model that both contracted (H)#FF and (H)#FL should exhibit similar 'falling' F_0 trajectories.

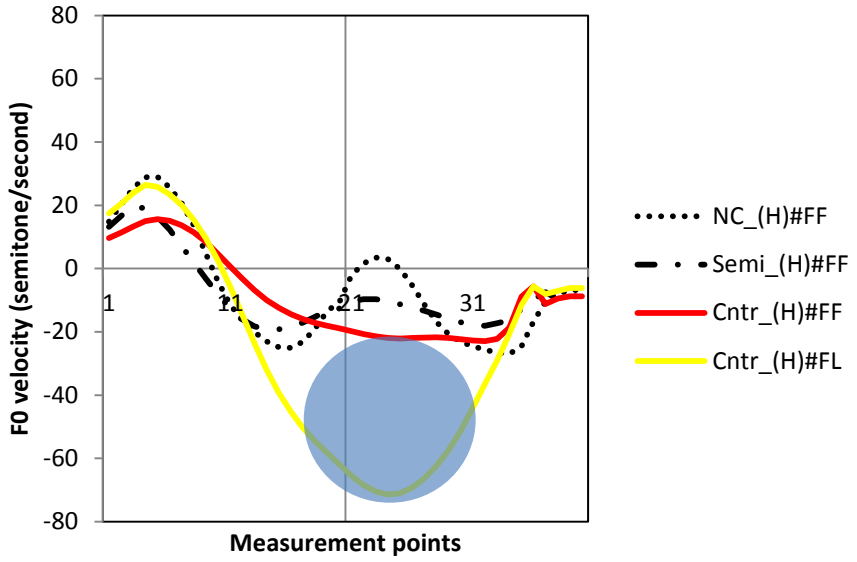


Figure 3.15: Mean F_0 velocity contours of (H#)FF of three contraction types and that of contracted (H#)FL (cf. Figure 3.13c&d, p. 78).

3.3 General discussion and conclusions

In this chapter we tested whether tonal reduction in contracted syllables can also be accounted for by time pressure. The three predictions were all confirmed, providing further evidence for the general hypothesis that *time pressure is the direct cause of extreme reduction*. Firstly, that *tonal contraction can occur in nonsense words if time pressure is sufficiently high* was supported. Ordinal logistic regression analysis (see Sec. 3.2.1.B) suggested that speech rate has a significant effect on the type of contraction that occurs (also see Figure 3.8). Further analyses also indicated a close relation between duration and F_0 excursion size (Tables 3.3 and 3.4).

Secondly, the slope of the regression line of F_0 peak velocity over F_0 movement amplitude supports the prediction that *articulatory effort is not decreased when contraction occurs*. Similarly to the segmental data reported in Studies 2a and 2b (Chapter 2), the tonal data labelled as semi-contracted or contracted often exhibited a decreased duration and excursion size but not a decreased articulatory effort (Table 3.4). Furthermore, when the maximum rate of pitch change was computed, it appears that speakers had already reached their physiological limit, particularly in cases when duration was comparable to that of a contracted syllable (Figure 3.9). That is, speakers could not change pitch at a rate faster than this physiological limit and thus inevitably had to undershoot the desired tonal targets.

Despite the high time pressure placed on the majority of tokens that were reduced, the third prediction that *speakers still attempt to approach each and every*

underlying tonal target was supported through examining F_0 velocity profiles across different contraction types. Take for example tone dyad (H)#H*R*, evidence displaying speakers' attempts to reach the R tone under varying time pressures are shown in Figure 3.14. Unlike the mechanism suggested by the Edge-in model, the absence of a final rise in contracted H*R* (Figure 3.10b) can be better explained by the time pressure account. That is, the shorter duration in contracted syllables prevents the velocity change from being translated into substantial changes in the overall F_0 counter as is also seen in contracted F*F* (Figure 3.15). This analysis of F_0 velocity profiles in contrast with the predictions made by the Edge-in model not only strengthens the validity of the time pressure account for extreme reduction, but also brings out the continuous nature of the effect of duration on target realisation.

3.3.1 The properties of tones in connected speech

The results presented in this chapter display two typical properties of tones in connected speech: (a) reduced pitch range as exhibited by a small F_0 excursion size and (b) simplified F_0 contours which are seen as general sloping contours shown in particular by the absence of a final rise for the R tones. Shrunken tonal space is similar to that observed for vowel space in unstressed syllables or at fast speech rate, which as noted above, can be comparably explained by time pressure. Regarding the tonal shapes of a contracted syllable, previous research has found that at fast speed, syllable duration can be so short that the dynamic tone R is realised with a virtually flat F_0 contour (Xu and Wang, 2005; Kuo et al., 2007). It

has also been argued above that even though the original tonal elements are attempted briefly in velocity profile, there is not enough time for the effort to result in large F_0 movements.

Tseng (2005b) analysed spontaneous speech in Taiwan Mandarin and reported that the most frequent tone combinations for tonal merger all contain an F tone, especially for disyllabic contractions with an F tone in the second syllable. She suspected that a falling movement may be relatively easier for speakers to execute when duration is as limited as in a contracted syllable and therefore being retained more frequently. An increased rate of contraction in tone dyads with a falling tone in the second syllable was not found in this experiment (Table 3.2), but it is shown in Figures 3.10-3.13 that an F tone is generally less susceptible to the loss of its dynamic features (i.e. a falling movement) than that of an R tone (i.e. a rising movement). This may help explain why a majority of (near or already) fossilised words often carry an F tone, as observed in Tseng (2005b). Furthermore, as a side note, this high dependency of duration on target realisation seems to further agree with the evidence found in Xu (1998) and Xu and Wang (2001) that R and F tones in Mandarin Chinese have dynamic phonetic targets.

Therefore, the residual tonal variants in contracted syllables are not necessarily generated by rule (i.e., retaining only the edge portions of tonal contours like the initial part of the first tone and final part of the second tone), but are in fact rather mechanical – as duration is shortened, the movement toward the desired targets is gradually curtailed. Moreover, observations of F_0 contours and their velocity

profiles allowed us to see even further evidence of articulatory movements toward the underlying targets of the four tones, as has previously been demonstrated by Gauthier et al. (2007).

3.3.2 Conclusion

In this chapter, an experiment was designed to evaluate the nature of tonal reduction. It involved the examination of tonal variations recited under varying timing pressures and in systematically varied tonal environments. To summarise, we have demonstrated that tonal reduction is largely dependent on duration and that speakers still attempt to reach each underlying tonal target within the limited duration. Moreover, there appears to be a physiological limit to the 'extra' effort a speaker can apply when producing a tone, and when under extreme time pressure this extra effort may not be sufficient to fully offset the effect of time pressure, thus resulting in reduction. That is, it is the speed limit of articulation together with the fast speaking rate that leads to contraction as well as reductions that are less severe. Having satisfied all three predictions, the hypothesis that *time pressure is the direct cause of tonal reduction* is largely supported.

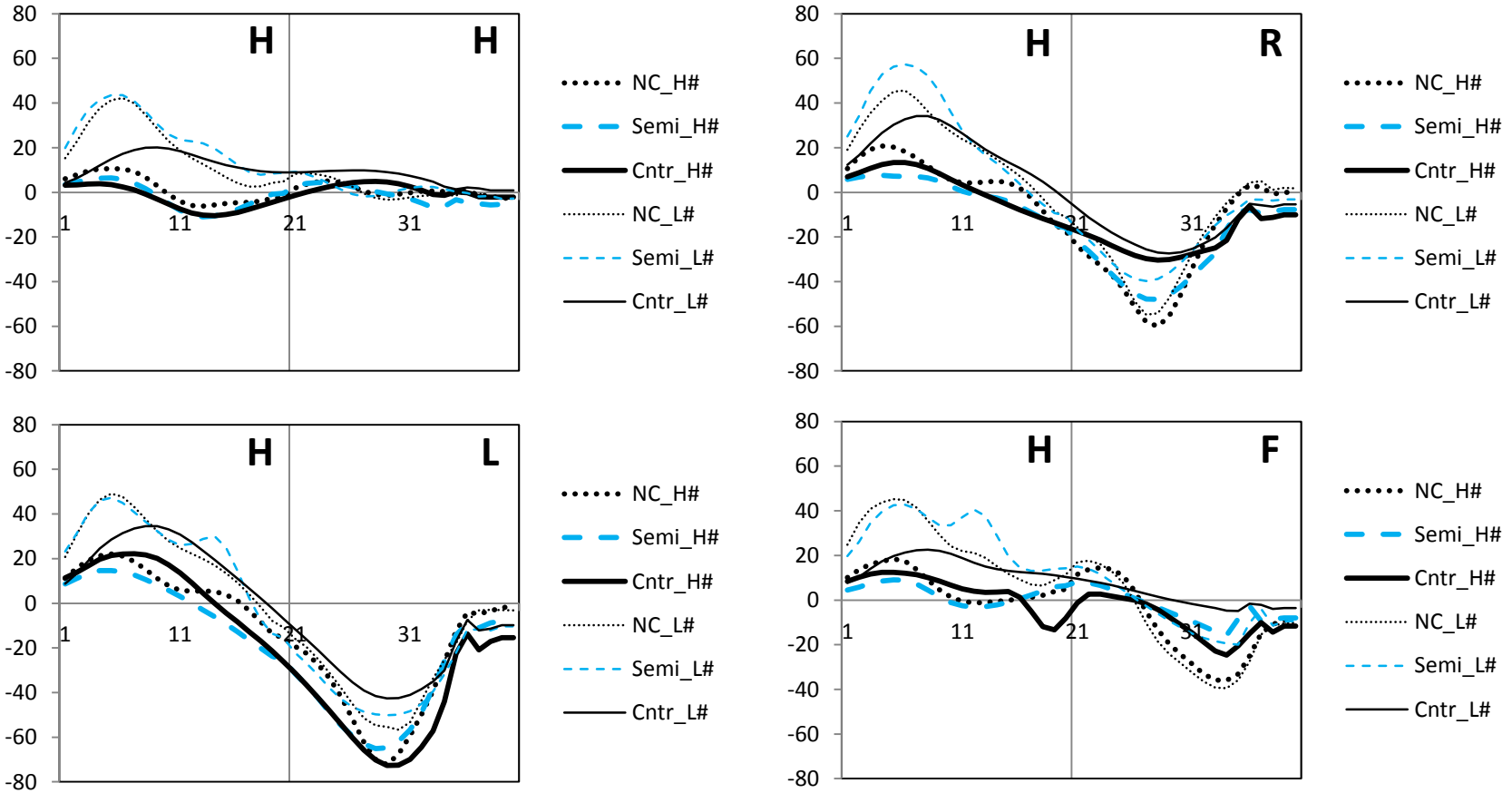


Figure 3.16: F_0 velocity profiles of tone dyad HH, HR, HL and HF. The x-axis is normalised 40 measurement points and the y-axis is in unit of semitone per second. Note that the F_0 velocity profiles were calculated before the time normalisation so the original velocity values were preserved.

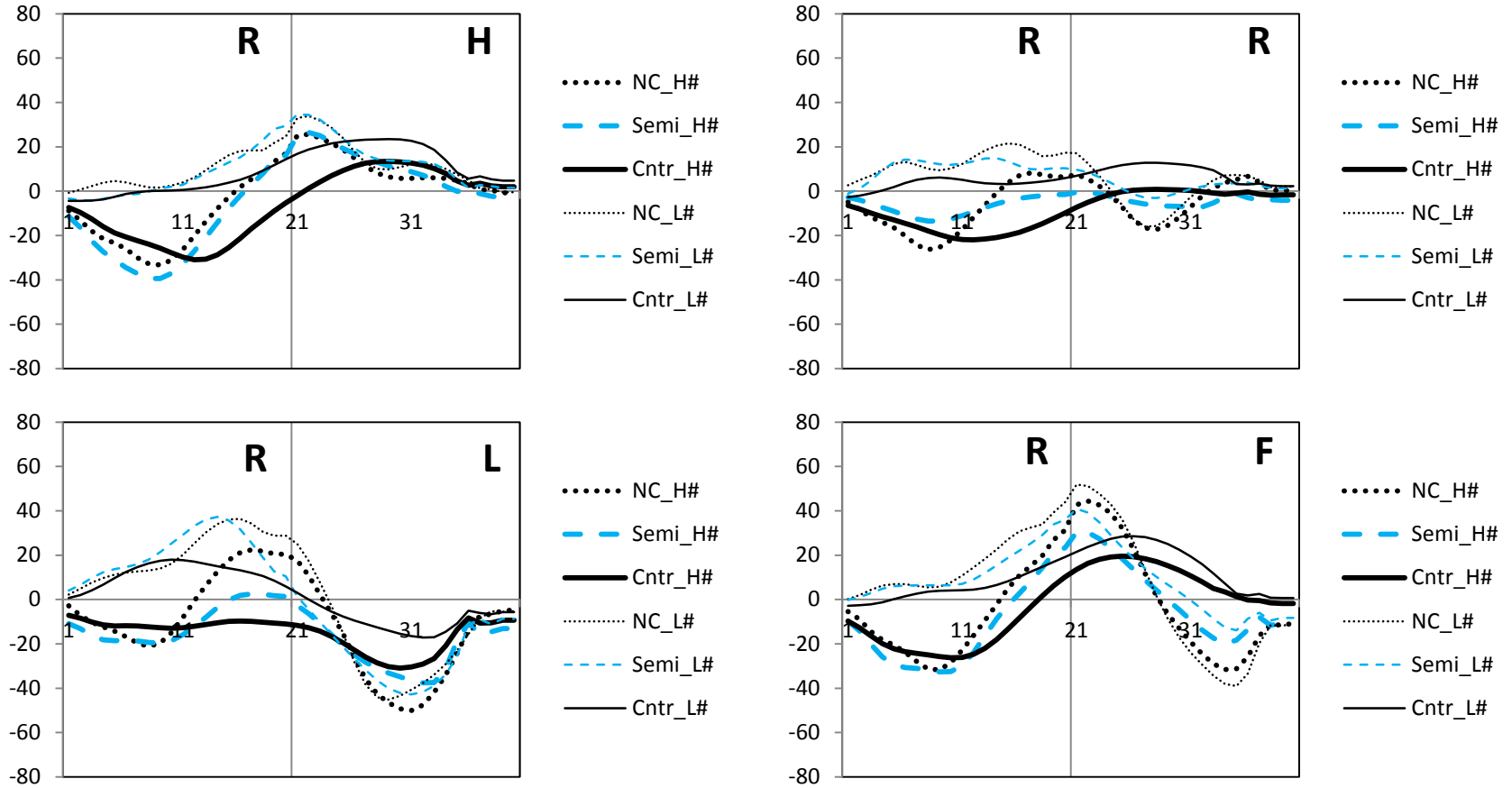


Figure 3.17: F_0 velocity profiles of tone dyad RH, RR, RL and RF. The x-axis is normalised 40 measurement points and the y-axis is in unit of semitone per second.

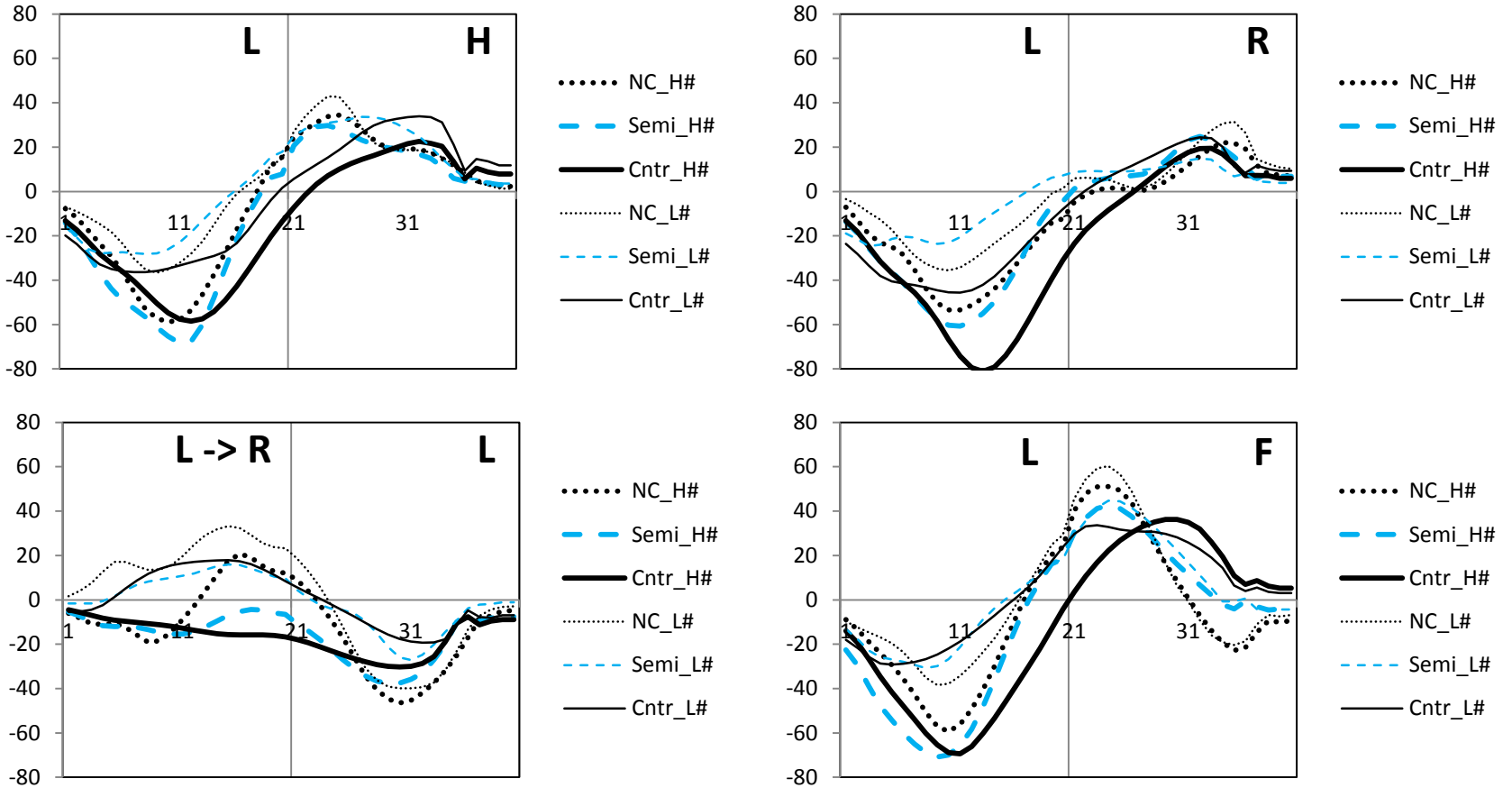


Figure 3.18: F_0 velocity profiles of tone dyad LH, LR, LL -> RL and LF. The x-axis is normalised 40 measurement points and the y-axis is in unit of semitone per second.

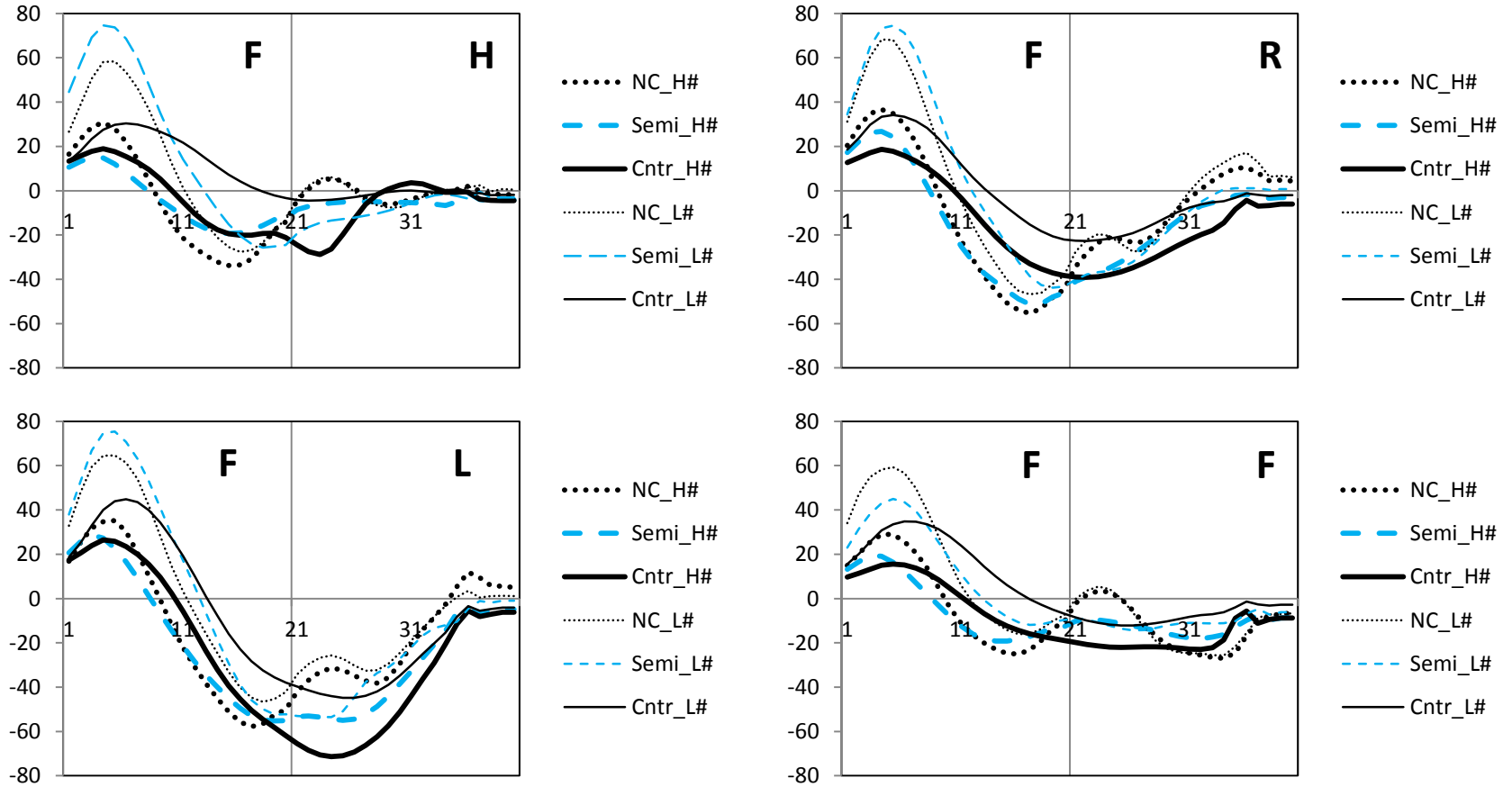


Figure 3.19: F_0 velocity profiles of tone dyad FH, FR, FL and FF. The x-axis is normalised 40 measurement points and the y-axis is in unit of semitone per second.

Chapter 4

Modelling tonal reduction

In Chapter 2, evidence that extreme segmental reduction is the direct result of time pressure was presented. It was also shown that during extreme reduction articulatory effort is likely to be increased rather than decreased. The same conclusions can also be drawn from the results regarding tonal reduction presented in Chapter 3. Additionally, evidence was shown that when two syllables are merged into one, within the resulting contracted syllable speakers still attempt to reach the original tones. In this chapter, the time pressure account will be further tested by making use of quantitative modelling, which attempts to simulate contracted F_0 contours given some model parameters extracted from non-contracted tones. The main model used for this purpose is the articulatory-based quantitative Target Approximation model (qTA; Prom-on et al., 2009) designed to simulate the articulatory dynamics of F_0 production. The supplementary modelling

method used is Functional Linear Model (FLM), based on Functional Data Analysis (FDA; Ramsay and Silverman, 2005; Ramsay et al., 2009).

The chapter will consist of two parts. Within each part, the specific modelling methods used will be introduced along with the procedures required to simulate or predict tonal contours with varying durations. Following this, results are evaluated to determine whether the corresponding synthesis (qTA) or prediction (FLM) provides a close match to recorded data.

4.1 The qTA model

The quantitative target approximation model (qTA; Prom-on et al., 2009) is an implementation of the theoretical target approximation model (TA; Xu and Wang, 2001). The qTA model specifies a continuous link between articulatory mechanisms of F_0 contour generation and the functional components of speech melody, which in our case is distinguishing words through lexical tones. The qTA model represents F_0 as a response of a pitch target approximation process (Prom-on et al., 2009; Xu, 2005; Xu and Wang, 2001). A schematic outlining the TA process is shown in Figure 4.1.

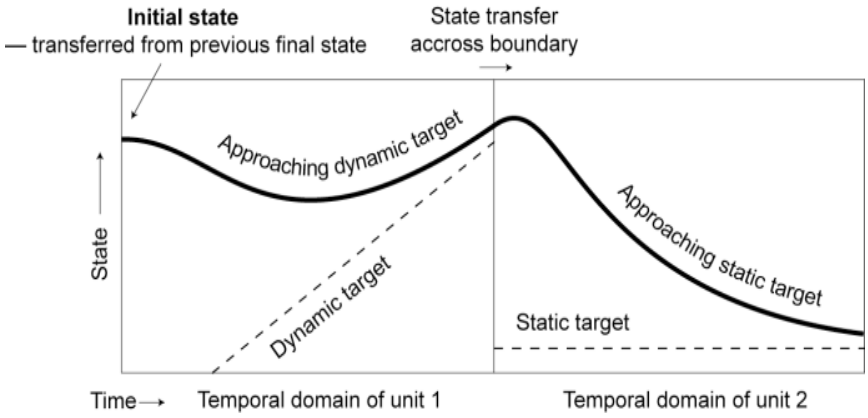


Figure 4.1: Target approximation model, adapted from Xu and Wang (2001).

A pitch target is defined as the underlying goal of the local prosodic event (Prom-on et al., 2009). It can be represented by a simple linear equation:

$$x(t) = mt + b, \tag{4.1}$$

where *x(t)* is the pitch target. The parameters *m* and *b* are the slope and height of the pitch target, respectively. Based on the assumption that target approximation is synchronized with the host syllable (Xu and Wang, 2001), the time *t* is relative to the onset of the syllable.

Now, let the F_0 response of the vocal fold tension control mechanism driven by the pitch target (Prom-on et al., 2009) be labelled *y*. The core mechanism of the model is represented as a third-order critically damped linear system, expressed mathematically as:

$$y(t) = x(t) + (c_1 + c_2 t + c_3 t^2)e^{-\lambda t}, \tag{4.2}$$

where *x(t)* is given in (4.1) and the constants c_1, c_2, c_3 and λ are described below. The polynomial and exponential multiplier represent the natural response of the tension control system, where λ is the rate of target approximation, which controls how fast a target is approximated. The transient coefficients c_1, c_2 and c_3 are determined by the initial conditions and other syllable dependent model parameters. The initial conditions of the articulatory process include initial F_0 level, $y(0)$, initial velocity, $y'(0)$, and initial acceleration, $y''(0)$. By solving the system of linear equations resulting from applying the initial conditions, the transient coefficients can be calculated using the following formulae:

$$c_1 = y(0) - b, \tag{4.3}$$

$$c_2 = y'(0) + c_1\lambda - \mathrm{m}, \tag{4.4}$$

$$c_3 = (y''(0) + 2c_2\lambda - \mathrm{c}_1\lambda^2)/2, \tag{4.5}$$

According to TA (target approximation), the degree to which a tonal target is realised depends on: (1) the distance between initial F_0 and the target, (2) the rate of target approximation and (3) the duration of the syllable. Thus, when (1) and (2) remain constant, shortening the syllable duration alone can lead to undershoot of the tonal target. As shown in Figure 4.2, qTA (quantitative target approximation) can simulate increased flattening of F_0 contours in two consecutive rising tones by simply shortening the duration of the syllables. Such flattening is similar to that seen in Xu and Wang (2009).

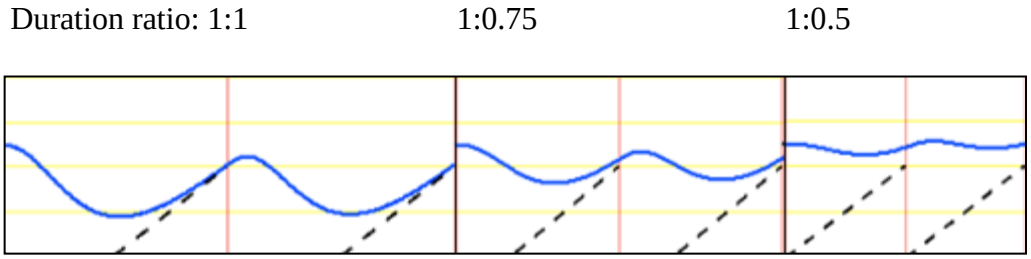


Figure 4.2: Effect of syllable shortening on two consecutive rising tones preceded by a high tone (not shown here) simulated by an interactive demonstration of qTA that can be found at http://www.phon.ucl.ac.uk/home/yi/qTA/.

4.1.1 Methodology

A. Corpus

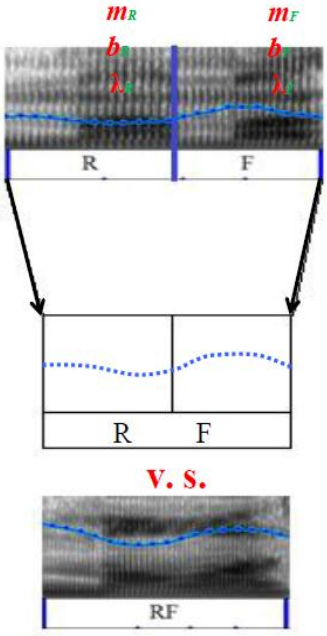
The corpus was taken from the speech materials recorded in Study 3 (Chapter 3). Considering individual variability and a carry-over effect resulting from tonal context, the corpus was divided into 12 subsets with respect to individual speakers (C, K, H, S, A and W) as well as tonal contexts (preceding tones H# and L#). Of the 12 subsets, six had a relatively balanced count of both non-contracted and contracted items in all 16 tone dyads. These were subsets C_H#, C_L#, K_H#, K_L#, H_L# and S_L# and were used for qTA modelling and simulation.

B. Modelling and simulation procedure

To investigate whether adjusting duration alone can simulate tonal reduction, our basic modelling strategies were as follows. First, to obtain raw target parameters we trained the qTA model on each of the *non-contracted* items in the aforementioned six subsets. These raw parameters were then averaged with

respect to tone dyad (i.e. HH, HR, …, FL and FF) to produce representative tonal target parameters (*m*, *b*, λ). Further details of this procedure are given in Sec. 4.1.2.

For each selected subset, three subsequent simulations were conducted. *Simulation 1* was to simulate F_0 contours of each of the individual non-contracted bi-tonal sequences using respective canonical target parameters (Tables 4.2-4.7) along with their own duration and initial F_0 (Sec. 4.1.3.A). *Simulation 2* was to simulate the reduced tones, again using the canonical parameters (Tables 4.2-4.7), but now with the *duration* and *initial* F_0 of the *contracted* items as input to the model (Sec. 4.1.3.B). A schematic of the simulation procedure is shown in Figure 4.3 below.



← Parameter values trained from *non-contracted* contours.

← Reduce the duration to the same length as the contracted units (steps to shorten duration are detailed in Sec. 4.1.3. B).

← Synthesised contracted contours (resulting from the application of the *non-contracted m, b* and λ with initial F_0 and duration of the *contracted* contours).

← Compared the results with the recorded contracted contours.

Figure 4.3: A schematic representing the procedure of Simulation 2.

In *Simulation 3*, the duration of the contracted tones was again used, but the target values applied were randomly selected from the canonical target sets. This was to test the possibility that reduced F_0 contours were simply flattened (Sec. 4.1.3.C). The performance of the simulations (i.e. the results of comparison between the synthesized and natural curves) was evaluated in terms of goodness-of-fit to the original F_0 contours, measured in Root Mean Squared Error (RMSE, defined in Eq. 4.6 below) and Pearson's correlation coefficient (defined in Eq. 4.7 below).

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y(t_i) - f_0(t_i)\right)^2}, \tag{4.6}$$

$$Correlation = \frac{N\sum_{i=1}^{N} y(t_i)f_0(t_i) - \sum_{i=1}^{N} y(t_i)\sum_{i=1}^{N} f_0(t_i)}{\sqrt{N\sum_{i=1}^{N} y(t_i)^2 - \left(\sum_{i=1}^{N} y(t_i)\right)^2}\sqrt{N\sum_{i=1}^{N} f_0(t_i)^2 - \left(\sum_{i=1}^{N} f_0(t_i)\right)^2}}, \tag{4.7}$$

where f_0 represents the value of the recorded pitch contours and y is the value of the predicted contour given by equation (4.2). N is the number of points used to approximate the contours.

4.1.2 Extracting qTA parameters from non-contracted bi-tonal sequences

Modelling was carried out using a Praat script that implements the qTA model (Xu and Prom-on, 2010). The script is a modified version of the publicly released PENTAtrainer (Xu and Prom-on, 2010-2011). It simulates the F_0 contours of an utterance by applying qTA through automatic analysis-by-synthesis. For each interval to be simulated, the script extracts the target approximation parameters

introduced in (4.2), *m*, *b* and λ. Parameter estimation was done automatically in the script by minimizing the sum of squared errors between the simulated and original F_0 contours. The Praat script was applied to all non-contracted bi-tonal sequences in order to extract the target parameters (*m*, *b* and λ) from each syllable in a disyllabic word. Averaged across all six subsets, a mean RMSE of 0.32 semitones was obtained in the first syllable and 0.47 semitones in the second.[7] The overall correlation is 0.97 and the average RMSE of the two syllables and correlation values are shown as the leftmost points in Figure 4.6. These low RMSE and high correlation values indicate that the qTA model, when using the extracted target parameters, accurately *resynthesized* each of the individual natural F_0 contours.

The parameters extracted from each individual contour were averaged with respect to tone type (H, R, L and F) to reflect the general properties of the pitch targets for each of the tones as shown in Figure 4.4, which displays boxplots of mean qTA parameters and duration of each tone type (obtained from both syllables). A logistic regression test also indicated that overall a significant relationship between the dependent (i.e. tone type, H, R, L and F) and independent variables (parameter triplet *m*, *b* and λ) existed ($p < .0001$). Tone sandhi effect was considered in the significance test for the logistic regression, i.e. in an LL

[7] The mean target parameters, duration and the RMSE and correlation values from comparison with the original F_0 contours are provided in Table 4.1, and the six subset's respective canonical parameters in Tables 4.2-4.7.

sequence the first L was assigned to the R category. Overall, to predict the tone type, 48 parameter triplets were trained for tone H, 54 for R, 42 for L and 48 for F, irrespective of whether the tone occurred in the first or second syllable.

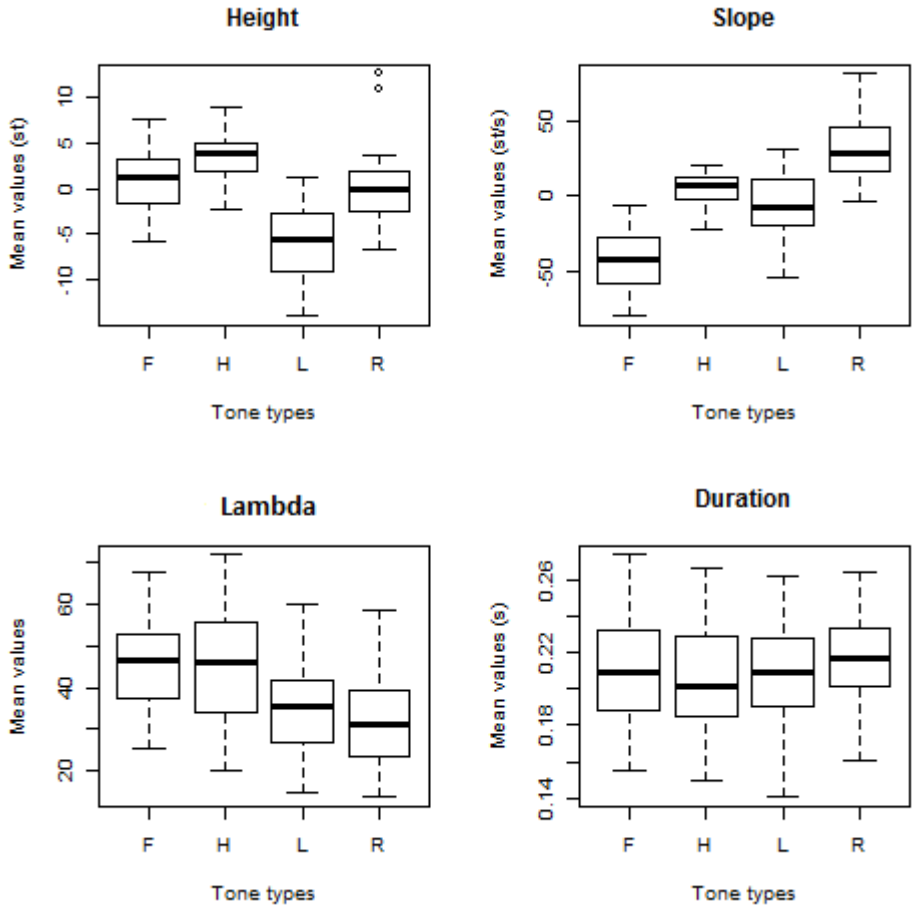


Figure 4.4: Boxplots of qTA parameters and duration for each tone type.

Figure 4.5 shows lexical tone functions in both syllables, similarly indicating that the four tones are distributed into four distinct clusters. It is interesting to note that the target values of the parameters in the second syllable (i.e. _H, _R, L and _F) cluster more toward the centre (indicated by the oval shaded region) while the first syllable's tones (i.e. H_, R_, L_ and F_) are distributed away from the centre. That is, in comparison to the first syllable, the second syllable exhibits a reduced pitch range. Additionally, the derived R (LL -> RL) clusters closely to the underlying R, indicating a virtual merger of the L and R tones under sandhi, which is consistent with previous studies (Peng, 2000 for Taiwan Mandarin, and Xu, 1997 for Beijing Mandarin).

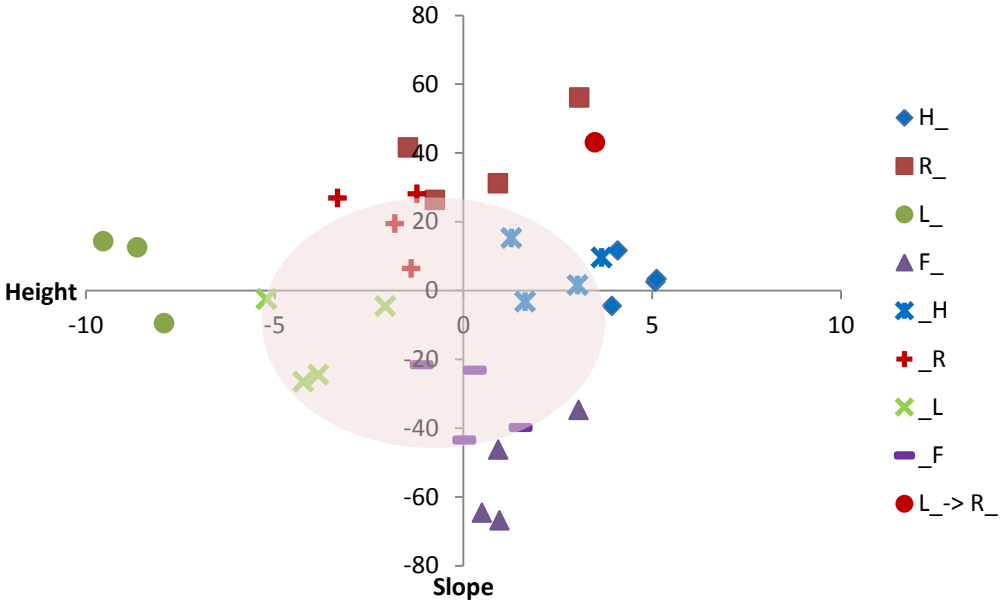


Figure 4.5: Distribution of the lexical tone function, with respect to parameters *b* (Height; the x-axis) and *m* (Slope; the y-axis), of both syllables. The oval shape signifies the centralised cluster of parameter values *m* and *b* in the second syllables (_H, _R, _L and _F) in comparison to the first syllable (H_, R_, L_ and F_).

After obtaining the canonical target parameters for each tone dyad, a series of simulations (i.e. Simulations 1-3, as described earlier in Sec. 4.1.1.B) testing the proposed time pressure account were conducted, the results of which are presented in the next section.

4.1.3 Simulation results

A. Simulation 1: Using canonical parameters to simulate non-contracted bi-tonal sequences

The canonical parameters of the six subsets (shown in Tables 4.2-4.7) were used to simulate F_0 curves of corresponding *individual* non-contracted bi-tonal sequences using a customised Praat script that performs qTA synthesis. This script resynthesizes non-contracted F_0 curves by applying the canonical parameters with each individual curve's initial F_0 and duration. Results indicate a high correlation of 0.86. RMSE values between predicted and actual F_0 contours are also fairly low (1^{st} syll.: 1.53, 2^{nd} syll.: 2.15). Mean RMSE of the two syllables and correlation values are shown as the second points from left in Figure 4.6. As this figure shows, in comparison to the parameter training reported in Sec. 4.1.2, the decreased goodness-of-fit is expected since here, all individual contours of a particular bi-tonal sequence of each subset were simulated with the same set of mean target values from the respective tone dyads.

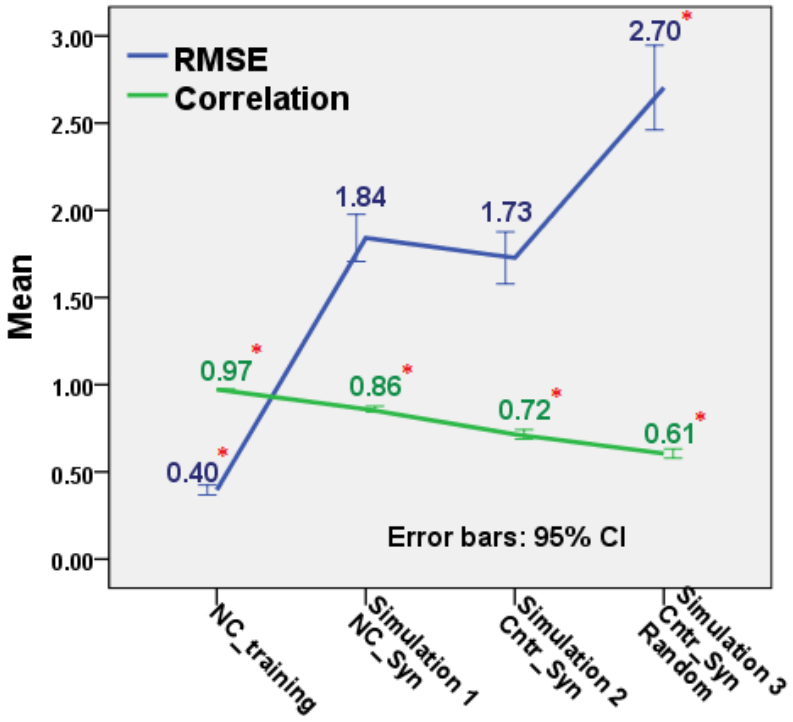


Figure 4.6: RMSE and correlation values from parameter training and those from the three subsequent simulations. The blue line indicates mean RMSE values and the green line indicates correlation values. On the x-axis, **NC_training** (Sec. 4.1.2): Parameter training on non-contracted items (the extracted values were averaged based on the 16 types of tone dyads and used as *canonical target parameters*); **Simulation 1**: Non-contracted tonal simulation (synthesized using *canonical target parameters* with *original*, i.e. non-contracted duration, 1153 items in total); **Simulation 2**: Contracted tonal simulation (synthesized using *canonical target parameters* with duration of the *contracted* items, 1010 items in total); **Simulation 3**: Contracted tonal simulation with random target assignment (synthesized with duration of the *contracted* items but *random assignment of* canonical target parameters, 1010 items in total). The red asterisks indicate the statistical significance of tests conducted in Sec. 4.1.3.C.

B. Simulation 2: Simulating F_0 contours of contracted tonal sequences with canonical parameters

In this simulation the proposed time pressure account is tested. We applied the canonical target parameters obtained in Sec. 4.1.2 to the contracted tonal sequences. This required several steps. First, a mean ratio of the relative duration of the two syllables in a bi-tonal sequence was computed from the non-contracted tokens, which were averaged with respect to each subset and its 16 tone dyads. Secondly, each contracted bi-tonal sequence, which consisted of only a single interval due to the loss of the intervocalic consonant, was divided into two intervals, each having the same relative duration as the mean relative duration of the corresponding canonical sequence. In the third step, each subset's respective canonical parameters from non-contracted tone dyads were applied to each of the individual contracted tonal sequences, interval by interval. Figure 4.7 shows an example simulation.

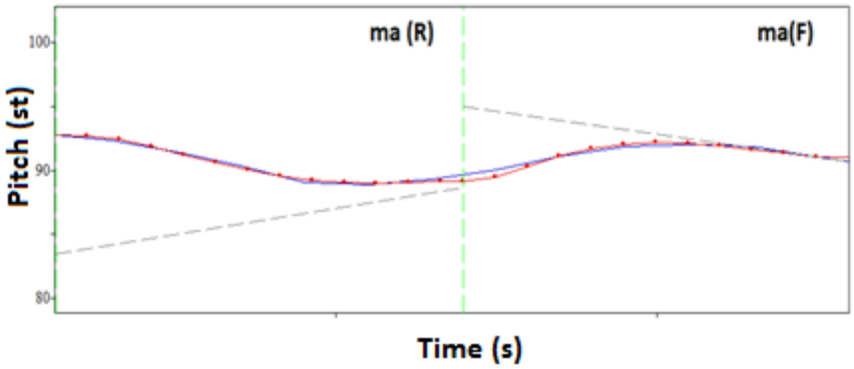


Figure 4.7: Simulation of a contracted bi-tonal sequence using the canonical parameters of the tone sequence RF (preceded by a High tone, not shown here) and contracted duration. The figure indicates the pitch targets (grey dashed lines), synthesized F_0 (red dotted curve) against the original F_0 (blue curve).

The mean RMSE and correlation results are shown as the third points from left in Figure 4.6. The evaluation of goodness-of-fit gives a correlation of 0.72, a RMSE of 1.23 (interval 1) and 2.23 (interval 2). The mean RMSE value is actually slightly lower than that of Simulation 1 (Point 2 in Figure 4.6), which is likely due to a smaller pitch range of contracted curves resulting in smaller errors. The correlation value is not as good as that of Simulation 1, but this could be due to the intrinsic uncertainty introduced in the simulation process. Unlike in Simulation 1, where syllable duration values were those of the original individual tokens, duration values in the current simulation were the estimated averages of the non-contracted tokens. Thus the proportional durations of syllable 1 and syllable 2 and variations of individual tokens were estimated because neither was recoverable. Taking this factor into consideration, a correlation of 0.72 is still satisfactory, and is only slightly lower than the 0.74 obtained in Prom-on et al. (2009, p. 418, Table X) for tone-only simulations.

C. Simulation 3: Simply flattened? Random target application

The results of Simulation 2, despite seemingly agreeing well with our hypothesis, could be due to the fact that tones are simply flattened such that all tone dyads become similar to each other. To examine this possibility, the canonical targets were randomly assigned to contracted tone sequences in another simulation. The tone types for the random pairings are shown in the last column of Table 4.1 (p. 119). If the reduced tone sequences are no longer related to the canonical sequences, results from this simulation should be little different from those of Simulation 2.

As shown in Figure 4.6, in comparison to Simulation 2, the correlation value decreased to 0.61, and RMSE increased to 2.17 for interval 1 and 3.24 for interval 2. A multivariate analysis of variance (MANOVA) on correlation and RMSE values showed significant differences across the four sections (Correlation: $F_{(3,764)} = 233.2$, $p < .000$; RMSE: $F_{(3,764)} = 140.9$, $p < .000$). Post hoc analysis (Tukey HSD) on correlation further indicated that results from all sections differed significantly from each other. Another post hoc analysis (Tukey HSD) on RMSE showed no significant difference between Simulations 1 and 2 ($p = .745$). Thus, the result of simulation 3 (i.e. random target application with shortened duration) is significantly worse than that of Simulation 2 (i.e. matching the correct target parameters with the shortened duration). This means that the correlations seen in Simulation 2 are unlikely due to simple F_0 flattening.

D. Summary of qTA modelling

So far in this chapter further evidence in support of the time pressure account has been presented through the use of qTA modelling. Detailed F_0 contours of contracted tokens were constructed by assuming tonal target approximation under time pressure, revealing an explicit link between duration and F_0 realisation (Simulations 1 and 2). Additionally, it is shown in Simulation 3 that the same underlying tonal targets are attempted by the speakers even when severe tonal reduction occurs. It is therefore concluded that an articulatory-based model such as the qTA model adequately reflects the articulatory mechanism of speakers when generating F_0 contours under varying degrees of time pressure.

4.2 Functional linear modelling

The second examination of the effect of varying duration on tonal realisation is through Functional Linear Modelling (FLM), which is one of the techniques available in the family of advanced statistical techniques called Functional Data Analysis (FDA), introduced in the late 90's by J. Ramsay and colleagues (Ramsay and Silverman, 2005; Ramsay et al., 2009). FDA techniques allow one to conduct statistical analysis on a set of contours (F_0 in our case) using only the information contained in their shape. Quantitative analysis of F_0 contours and other time-varying quantities (e.g. formants) is usually carried out by selecting a few shape features such as minimum and maximum coordinates, slopes etc., and then using standard statistical analysis on the derived fixed-length feature vectors (e.g. Morén and Zsiga, 2011). This approach forces one to choose in advance which shape features are relevant and which are not. In the majority of cases, feature extraction is carried out by hand. Another approach is to use a model such as the Fujisaki model (Fujisaki et al., 2005) or the aforementioned qTA model (Prom-on et al., 2009). These models attempt to take into account the physiology of phonation. Their performance depends on how faithfully the larynx or the vocal tract is modelled and how well the parameter tuning is carried out. In contrast, FDA is a flexible platform that allows one to: (1) use sampled contour values directly as input to the statistical analysis; and (2) refrain from introducing hypotheses on the nature of the analysed signal. The output of FDA is based solely on the regularities found within the set of input contour shapes. In view of FDA's flexibility in handling analysis of curves, we will investigate the time

pressure account by using duration alone to predict tone shapes of varying degrees of realisation.

4.2.1 Methodology

Functional Linear Models (FLMs) extend ordinary linear models to accept functions (of time) as input and/or output. The FLM model applied here takes a real number as an independent variable (predictor) and predicts a whole contour shape expressed as a function of time. In our case, the predictor *d* is a convenient transformation of the duration *D* of a disyllabic unit, while the output *y(t)* is the predicted F_0 contour shape defined on a *fixed* time interval. In this way, the shape and duration *d* are decoupled. Formally we have:

$$y(t) = \beta_0(t) + \beta_1(t) \cdot d, \qquad (4.8)$$

where *d* is the logarithm of the normalised duration *D*, that is, $d = \log(D / \bar{D})$ where $\bar{D}$ is the average duration measured across all tokens in the model. $\beta_0(t)$ and $\beta_1(t)$ are the functional parameters to be estimated and are analogous to the scalar parameters estimated in ordinary linear regression. The training of equation (4.8) is assembled in a similar manner to that in which ordinary linear regression models are trained (see Ramsay et al., 2009 for details). The main difference for the user is that each training element is a (d, $F_0(t)$) pair, while the original F_0 contours are sampled. Hence the sampled F_0 contours must be represented by continuous functions in time before training can take place. Moreover, they have

to be modified such that they span the same time interval as the functions $y(t)$, $\beta_0(t)$ and $\beta_1(t)$ must be defined on a common interval (Gubian et al., 2011).

A. Corpus

The acoustic data used to construct FLM was from subsets C_H#, C_L#, K_H#, K_L#. The other two subsets, H_L# and S_L#, used in the qTA modelling were not included in view of a need for balanced datasets, i.e. same speakers with compatible tonal contexts. Extraction of F_0 contours was first carried out with the vocal cycle marking of the Praat program and then with manual repair of octave jumps and other distinct irregularities using a Praat script (Xu, 2005-2011). For each target curve, 20 measurement points were generated, 10 equidistant points per syllable for *non-contracted* tokens (hence 20 in total) and 20 equidistant points for *contracted* tokens. F_0 values were converted to semitones and the average of the 20 samples was subtracted from all contours. This helped to reduce the variability owing to individual differences and made the estimation of functional linear modelling more straightforward.

B. Data preparation

The problem of choosing a function $y(t)$ that best fits a set of samples is solved by applying standard smoothing techniques. The user must choose a basis function, which for non-periodic signals is typically a B-spline. The internal parameters of the B-spline basis and the degree of smoothing imposed on the curve fitting were empirically determined by generalized cross-validation (Ramsay and Silverman,

2005; Ramsay et al., 2009). Examples of the quality of the smoothing process can be seen in Figures 4.11 and 4.12: Dots indicate the original F_0 samples and dashed lines indicate their respective functional representation. To obtain an (apparent) constant duration for each curve, a fictitious [0, 1] normalised time interval is simply divided into 20 constant intervals (i.e. one per sample point). In this way, the functions *y(t)* will be scaled so that the half curves spanning each syllable (in the *non-contracted* case) will be aligned with the centre of the interval. This improves the analysis quality in that it takes away variability due to random misalignment of syllables. All FDA operations were carried out using the freely available `fda` R package (R Development Core Team, 2011).

In order to test the effect of duration on tonal reduction, a model like equation (4.8) was built for each tone combination. The rationale is that once the tone sequence is known, (4.8) is a simple yet adequate description of a gradual shape adjustment rule controlled by the amount of time available for production. In practice, models had to be further specialised by building separate models for different preceding tones (L or H) and also for the two selected speakers (C and K) since the simple structure of (4.8) cannot accommodate the influence of such factors. Regarding speaker dependency, recall that equation (4.8) is trained on the surface realisation of a number of F_0 contours, while no physiological parameter estimation takes place. In total, 16 (tone dyads) x 2 (preceding tones) x 2 (speakers) = 64 FLMs were produced. Each model was trained using approximately 25 (log duration, smoothed F_0 curve) pairs with a nearly even proportion of *non-contracted*, *semi-contracted, contracted* samples.

The 64 models were evaluated in terms of goodness-of-fit on their training set. For each model we computed the root mean squared error (RMSE) and the R^2 coefficient of determination averaged on their training set. R^2 is defined as:

$$R^2 = 1 - SSE_{FIT}/SSY_{HORIZ}, \quad (4.9)$$

where SSE is the sum of squared errors resulting from approximating the sampled F_0 values with *y(t)* in (4.8), SSY the sum of squared errors from fitting a horizontal line at a height corresponding to the average F_0 value (i.e. fitting a horizontal line is taken as a baseline goodness-of-fit) (Motulsky and Christopoulos, 2003), as schematized in Figure 4.8 below. Note that R^2 is not the square of anything and takes a negative value when the predicted curve *y(t)* makes a squared error larger than that of fitting a horizontal line (i.e. when SSE > SSY).

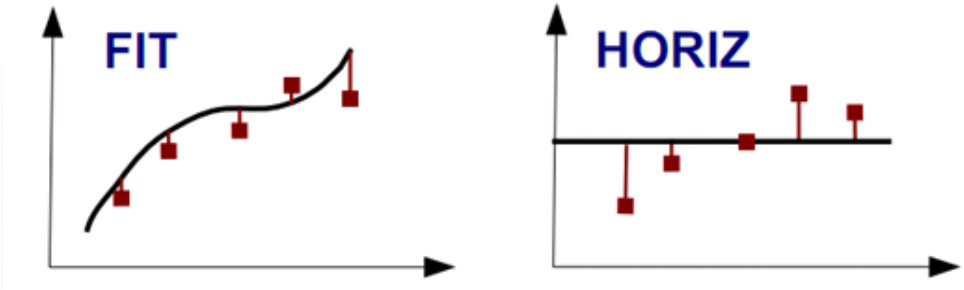


Figure 4.8: A schematic representing SSE_{FIT} and SSY_{HORIZ} in generating coefficient of determination (R^2) as one measurement of goodness-of-fit in Sec. 4.2.

4.2.2 Results

A. Evaluation

Figures 4.9 and 4.10 contain boxplots displaying the R^2 and RMSE value distributions across the 64 models. The leftmost columns show R^2 and RMSE values measured when predicting all the contours belonging to a specific tone-speaker combination. The two middle columns display the same data, but separately for non-contracted and contracted contours, respectively (results for semi-contracted contours are not shown separately but are included in the first column). At least half of the models exhibit an acceptable goodness-of-fit, which is remarkable if we consider the simplicity of (4.8) and the small number of training contours per model. The small differences between goodness-of-fit for non-contracted and contracted contours can be explained by the fact that contracted curves are flatter than non-contracted ones, thus the gain of fitting (4.8) relative to fitting a horizontal line (i.e. the baseline error considered in R^2) tends to be smaller for contracted contours, hence a smaller R^2 value in contracted cases. On the other hand, RMSE tends to be smaller (i.e. better) in the contracted case because contracted curves have smaller amplitude oscillations and therefore errors tend to have smaller absolute values.

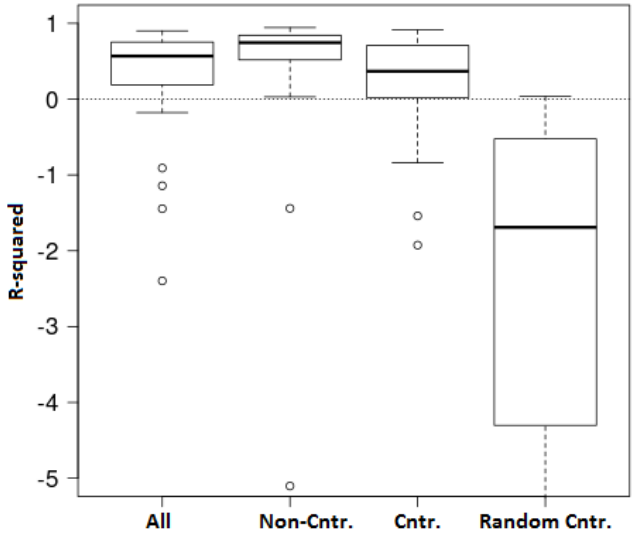


Figure 4.9: R^2 values from comparing observation and prediction. From left to right, the data are displayed for all contours, non-contracted contours, contracted contours and contracted contours with mismatched models.

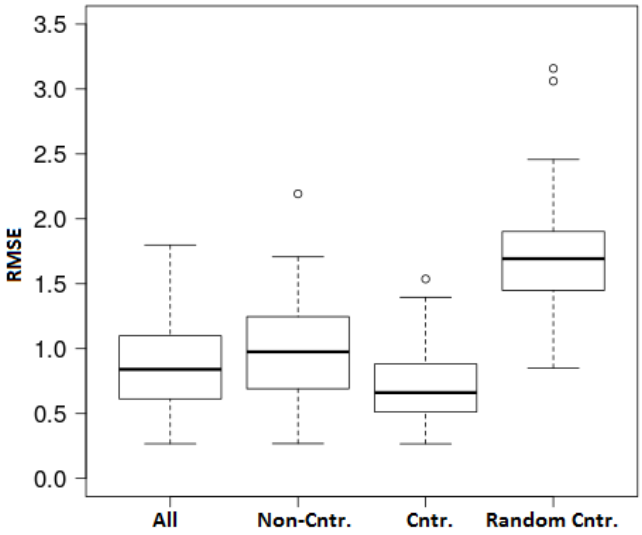


Figure 4.10: RMSE values from comparing observation and prediction. From left to right, the data is displayed for all contours, non-contracted contours, contracted contours and contracted contours with mismatched models.

Similarly to Simulation 3 which used the qTA model, at this point we wish to test whether it is in fact easy to predict the shape of a contracted tone combination because tones may simply be flattened and all tone combinations become similar to each other. To verify this we tried to predict contracted contours using mismatched models. The rightmost columns in Figures 4.9 and 4.10 display goodness-of-fit when predicting contracted contours using a model randomly picked from among the 31 models trained on the same speaker but on another tone combination/context. The large absolute and relative performance deterioration shows that contracted F_0 contours still preserve information in their shape and they are not simply flat. This confirms that the performance of the linear model (4.8) in predicting contracted contours is *not* due to F_0 flattening occurring uniformly across the board. In particular, comparing the 3rd and the 4th columns of Figure 4.9, we see that fitting a matched model (4.8) to predict a contracted contour is better than a flat line ($R^2 > 0$) in 75% of the cases (i.e. from the lowest line of the box to the top), while the converse is true when fitting a mismatched model (4.8).

B. Specific fitting examples

We now show some specific examples in order to gain further insight into the feasibility of using a simple model as (4.8). Figures 4.11 and 4.12 show some selected contours from the model constructed for speaker K in the tonal context L#RH, which globally scored RMSE = 0.35 and R^2 = 0.80. Figure 4.11 shows cases where a good fit was seen whereas Figure 4.12 shows cases where a poor fit was seen.

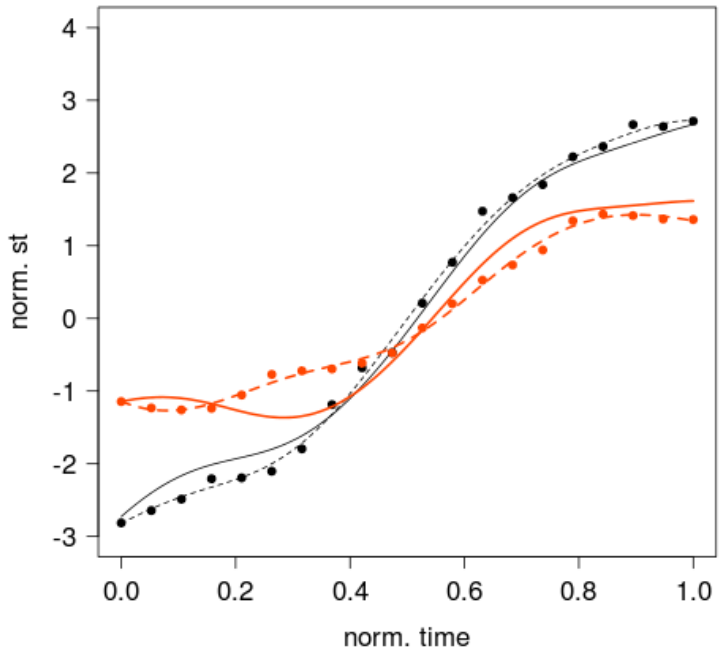


Figure 4.11: Cases where the model exhibited a good fit for Subject K and tone set L#RH. Measurement points are shown as dots and the smoothed *y(t)* contours shown as dashed curves. Solid curves are the respective predictions: Thin black line represents non-contracted token and thick orange line contracted token. The x-axis represents the normalized time from 0 to 1 (so it is not in seconds) and the y-axis is measured semitones (note the mean F_0 has been removed from each curve).

In Figure 4.11 (cases where a good fit was observed), we see an adequate approximation when the predictor of duration changes from long (D = 0.402s, d = 0.109, black solid) to short (D = 0.212s, d = -0.169, orange solid). This provides evidence for the continuous nature of durational effects on phonetic realisation since the contracted F_0 contour seems to preserve some traits of the non-contracted one. This is in support of the time pressure account on the nature of extreme reduction and this in turn is nicely captured by the FLM (4.8).

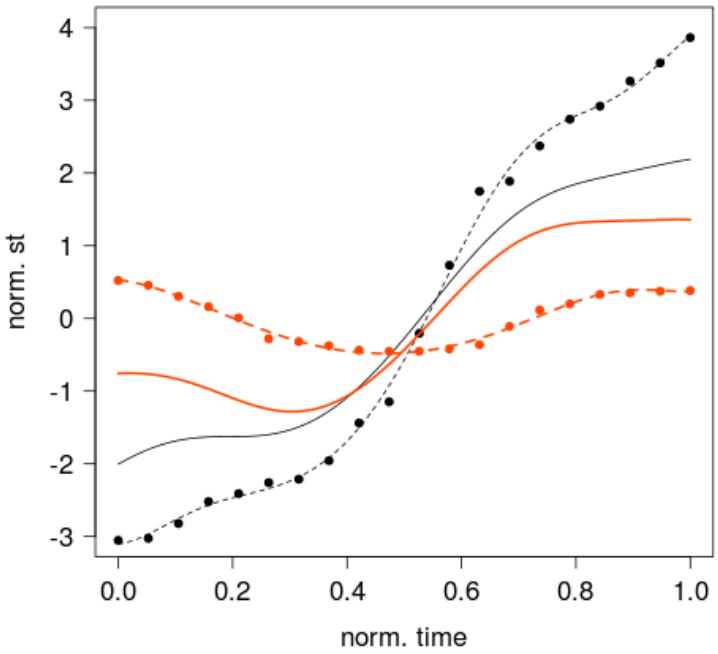


Figure 4.12: Cases of the model exhibited a poor fit for Subject K and tone set L#RH. Measurement points are shown as dots and the smoothed *y(t)* contours shown as dashed curves. Solid curves are the respective predictions: Thin black line represents non-contracted token and thick orange line contracted token. The x-axis represents the normalized time from 0 to 1 (so it is not in seconds) and the y-axis is measured semitones (note the mean F_0 has been removed from each curve).

In Figure 4.12 (cases where a poor fit was observed), we see that the chosen non-contracted curve (black dashed) has a greater amplitude than its prediction (D = 0.300s, d = -0.017, black solid) but is still of a similar shape (i.e. rising from low to high). On the other hand, the predicted contracted curve (orange solid line) shows a general mismatch in terms of shape and range. We attribute this to the limited predicting power of (4.8), which uses only one predictor, *d*. When time pressure is as high as in our cases of contracted syllables (D = 0.131s in the example), speakers simply cannot realise each target fully in time (though they

still attempted to), and thus produce a seemingly 'flattened' contour. By visually inspecting several other cases, patterns similar to those exemplified above were seen.

C. Summary of functional linear modelling

A data-driven approach was adopted to investigate the close link between duration and tonal reduction. With the support of functional linear modelling applied to an experimentally-controlled acoustic database, we found evidence in favour of the continuous nature of the durational effects on phonetic reduction. It should be noted that in this supplementary modelling method, both contracted and non-contracted tokens were used in the training and in the evaluation, which is different from the qTA modelling in Sec. 4.1. In the qTA modelling, after extracting the target parameters of the non-contracted tokens, detailed F_0 contours of the contracted tokens were generated by applying respective canonical targets and shortened durations. Owing to such training discrepancies as well as their different evaluation methods, the performance of both modelling methods cannot be directly compared. However, it is recognised that a model as simple as the one used in the functional linear model demonstrates that duration alone is capable of capturing varying degrees of F_0 realisation. This further supports the time pressure account regarding phonetic reduction. On the other hand, cases in which a poor fit is seen also indicate that the speakers' inherent physiological limit needs to be considered when modelling extreme reduction, as shown in the examples presented here.

4.3 Conclusion

Computational modelling is an empirical method to test the validity of a particular hypothesis with respect to empirical data/observations. In this chapter, two different models were adopted to further test the experimental findings reported in Chapter 3: Time pressure is the direct cause of tonal reduction in Taiwan Mandarin and that speakers still attempt to realise the same underlying tonal target within a reduced duration. The success of both modelling approaches strengthens the account regarding the effect of time pressure on phonetic reduction, providing more evidence for Lindblom's (1963) model of durational undershoot. Moreover, the qTA modelling carried out further demonstrates that detailed F_0 contours of contracted tokens can be generated by an articulatory-based computational model. This provides a strong predicative tool for modelling contraction and will help aid future investigations into the mechanisms governed by articulatory process.

Table 4.1: Canonical qTA target parameters *m*, *b* and *λ* (*m* and *b* define the slope and height of a linear target, and *λ*, the rate of target approximation), extracted from all non-contracted bi-tonal sequences, shown together with mean duration in seconds and local RMSE in semitones of each syllable, and their overall correlation values. Note that each subset has its own canonical parameters (see Tables 4.2 – 4.7) and Table 4.1 presents the average. The last column is the random tone dyads for Simulation 3 detailed in Sec. 4.1.3.C.

1st syllable	**m**	**b**	**λ**	**Duration (s)**	**RMSE (st)**	**2nd syllable**	**m**	**b**	**λ**	**Duration (s)**	**RMSE (st)**	**Correlation**	***Random***
H	-4	4	41	0.185	0.21	**H**	2	3	30	0.187	0.30	0.94	*RF*
H	2	5	41	0.228	0.28	**R**	19	-2	37	0.214	0.45	0.98	*FH*
H	3	5	42	0.226	0.30	**L**	-3	-5	35	0.201	0.72	0.97	*LR*
H	12	4	47	0.213	0.27	**F**	-44	0	45	0.200	0.30	0.98	*RL*
R	26	-1	24	0.223	0.27	**H**	10	4	47	0.199	0.36	0.98	*LF*
R	31	1	24	0.211	0.28	**R**	28	-1	33	0.202	0.39	0.96	*FL*
R	56	3	23	0.231	0.31	**L**	-27	-4	42	0.205	0.57	0.98	*FR*
R	42	-1	25	0.224	0.30	**F**	-40	2	56	0.209	0.39	0.98	*HH*
L	12	-9	25	0.223	0.42	**H**	15	1	64	0.212	0.47	0.98	*LL*
L	-10	-8	27	0.224	0.42	**R**	27	-3	50	0.208	0.57	0.96	*HL*
L	43	3	31	0.210	0.35	**L**	-25	-4	42	0.197	0.55	0.97	*FF*
L	14	-10	27	0.207	0.50	**F**	-22	-1	55	0.210	0.59	0.96	*HR*
F	-46	1	41	0.224	0.29	**H**	-3	2	50	0.202	0.35	0.98	*RR*
F	-65	0	38	0.230	0.31	**R**	6	-1	42	0.214	0.48	0.99	*LH*
F	-67	1	36	0.223	0.35	**L**	-5	-2	41	0.193	0.67	0.98	*RH*
F	-35	3	46	0.191	0.29	**F**	-23	0	52	0.193	0.35	0.97	*HF*

Table 4.2: Subset C_H# (speaker C, preceding tone H) with qTA parameters (*m*, *b* and *λ*) and mean evaluation values (RMSE and Correlation) from parameter training. Duration_ratio is used for Simulation 2 in shortening duration of the non-contracted to that of the contracted. The last column is the random tone dyads for Simulation 3 detailed in Sec. 4.1.3.C.

1st syllable	**m**	**b**	**λ**	**RMSE (st)**	**Duration_ratio**	**2nd syllable**	**m**	**b**	**λ**	**RMSE (st)**	**Correlation**	***Random***
H	-23	5	22	0.29	0.5024	**H**	4	2	28	0.43	0.91	*RF*
H	-15	5	31	0.27	0.5042	**R**	28	-5	31	0.42	0.99	*FH*
H	-5	5	36	0.26	0.5481	**L**	-12	-10	27	0.88	0.98	*LR*
H	15	2	38	0.29	0.4959	**F**	-65	-6	25	0.48	0.98	*RL*
R	46	-3	21	0.34	0.5289	**H**	7	2	44	0.41	0.98	*LF*
R	63	-1	19	0.45	0.4953	**R**	37	-2	30	0.70	0.94	*FL*
R	73	1	18	0.41	0.5373	**L**	-30	-7	35	0.64	0.95	*FR*
R	54	-5	18	0.34	0.5084	**F**	-53	-1	53	0.58	0.98	*HH*
L	12	-12	15	0.49	0.5298	**H**	21	1	68	0.57	0.98	*LL*
L	-25	-9	22	0.45	0.5216	**R**	24	-3	55	0.58	0.98	*HL*
L	65	2	26	0.45	0.5276	**L**	-31	-9	37	0.82	0.98	*FF*
L	12	-12	23	0.56	0.5221	**F**	-47	-1	48	0.90	0.98	*HR*
F	-67	-4	36	0.44	0.5071	**H**	-9	-1	57	0.51	0.98	*RR*
F	-80	-4	35	0.43	0.5055	**R**	1	-5	38	0.56	0.99	*LH*
F	-75	-1	30	0.43	0.5551	**L**	-6	-7	33	0.74	0.99	*RH*
F	-51	0	40	0.30	0.4941	**F**	-37	-5	57	0.38	0.99	*HF*

Table 4.3: Subset C_L# (speaker C, preceding tone L) with qTA parameters (*m*, *b* and *λ*) and mean evaluation values (RMSE and Correlation) from parameter training. Duration_ratio is used for Simulation 2 in shortening duration of the non-contracted to that of the contracted. The last column is the random tone dyads for Simulation 3 detailed in Sec. 4.1.3.C.

1st syllable	**m**	**b**	**λ**	**RMSE (st)**	**Duration_ratio**	**2nd syllable**	**m**	**b**	**λ**	**RMSE (st)**	**Correlation**	***Random***
H	12	7	48	0.23	0.4987	**H**	14	6	30	0.26	0.99	*RF*
H	8	6	41	0.36	0.4874	**R**	4	-1	44	0.42	0.98	*FH*
H	17	9	55	0.37	0.4902	**L**	-10	-3	37	0.81	0.98	*LR*
H	21	7	51	0.37	0.4926	**F**	-46	-1	35	0.37	0.97	*RL*
R	48	0	20	0.47	0.5044	**H**	9	9	55	0.48	0.98	*LF*
R	42	4	19	0.42	0.5041	**R**	53	0	19	0.55	0.98	*FL*
R	63	11	14	0.53	0.5084	**L**	-51	-4	38	0.81	0.99	*FR*
R	52	4	26	0.52	0.5005	**F**	-62	4	45	0.60	0.98	*HH*
L	12	-9	21	0.46	0.4936	**H**	17	4	49	0.77	0.97	*LL*
L	-7	-10	23	0.63	0.5094	**R**	40	-3	46	1.19	0.91	*HL*
L	66	13	15	0.75	0.5050	**L**	-54	-1	45	0.81	0.98	*FF*
L	-10	-9	20	0.89	0.4857	**F**	-32	-4	47	1.03	0.93	*HR*
F	-56	0	38	0.48	0.5118	**H**	0	2	56	0.42	0.97	*RR*
F	-67	4	34	0.46	0.5002	**R**	17	0	32	0.61	0.99	*LH*
F	-76	2	35	0.62	0.5227	**L**	14	0	36	1.05	0.98	*RH*
F	-27	8	49	0.54	0.4949	**F**	-26	3	53	0.32	0.98	*HF*

Table 4.4: Subset K_H# ((speaker K, preceding tone H) with qTA parameters (*m*, *b* and *λ*) and mean evaluation values (RMSE and Correlation) from parameter training. Duration_ratio is used for Simulation 2 in shortening duration of the non-contracted to that of the contracted. The last column is the random tone dyads for Simulation 3 detailed in Sec. 4.1.3.C.

1st syllable	**m**	**b**	**λ**	**RMSE (st)**	**Duration_ratio**	**2nd syllable**	**m**	**b**	**λ**	**RMSE (st)**	**Correlation**	***Random***
H	-20	2	20	0.14	0.4813	**H**	13	-1	28	0.23	0.81	*RF*
H	4	3	27	0.25	0.5194	**R**	28	-6	36	0.46	0.99	*FH*
H	16	1	25	0.26	0.5416	**L**	-16	-8	42	0.59	0.99	*LR*
H	5	1	29	0.38	0.5268	**F**	-42	-2	54	0.31	0.94	*RL*
R	25	-7	27	0.36	0.5078	**H**	18	-1	52	0.35	0.97	*LF*
R	22	-2	26	0.31	0.4878	**R**	4	-5	33	0.32	0.93	*FL*
R	55	-1	28	0.30	0.5210	**L**	-26	-8	44	0.55	0.98	*FR*
R	37	-6	25	0.37	0.4840	**F**	-32	-2	66	0.40	0.98	*HH*
L	2	-10	25	0.56	0.4871	**H**	18	-2	72	0.41	0.98	*LL*
L	-6	-9	28	0.46	0.5187	**R**	11	-6	51	0.55	0.98	*HL*
L	24	-1	37	0.29	0.4919	**L**	-7	-7	40	0.48	0.94	*FF*
L	29	-14	17	0.58	0.4747	**F**	-12	-3	61	0.60	0.95	*HR*
F	-46	-2	46	0.23	0.5037	**H**	-2	-1	55	0.33	0.98	*RR*
F	-69	-3	40	0.18	0.4991	**R**	5	-6	40	0.48	0.99	*LH*
F	-51	-3	37	0.32	0.5057	**L**	-7	-5	41	0.69	0.99	*RH*
F	-44	0	50	0.21	0.4800	**F**	-25	-3	44	0.33	0.99	*HF*

Table 4.5: Subset K_L# (speaker K, preceding tone L) with qTA parameters (*m*, *b* and *λ*) and mean evaluation values (RMSE and Correlation) from parameter training. Duration_ratio is used for Simulation 2 in shortening duration of the non-contracted to that of the contracted. The last column is the random tone dyads for Simulation 3 detailed in Sec. 4.1.3.C.

1st syllable	**m**	**b**	**λ**	**RMSE (st)**	**Duration_ratio**	**2nd syllable**	**m**	**b**	**λ**	**RMSE (st)**	**Correlation**	***Random***
H	0	3	46	0.20	0.4658	**H**	-10	5	33	0.24	0.98	*RF*
H	13	7	43	0.31	0.5120	**R**	17	1	39	0.58	0.98	*FH*
H	-9	6	27	0.42	0.5196	**L**	28	-4	30	1.29	0.93	*LR*
H	16	5	53	0.26	0.5129	**F**	-42	2	55	0.23	0.99	*RL*
R	12	2	33	0.19	0.5204	**H**	7	5	56	0.25	0.99	*LF*
R	9	2	21	0.20	0.5011	**R**	34	0	26	0.28	0.96	*FL*
R	82	3	19	0.24	0.5194	**L**	-15	-3	36	0.64	0.98	*FR*
R	58	-2	21	0.21	0.4924	**F**	-39	2	56	0.26	0.99	*HH*
L	32	-11	21	0.46	0.5080	**H**	11	-1	67	0.38	0.98	*LL*
L	-9	-10	29	0.46	0.4978	**R**	30	-6	54	0.49	0.98	*HL*
L	54	2	29	0.28	0.4998	**L**	2	-3	34	0.55	0.97	*FF*
L	23	-10	28	0.44	0.4825	**F**	-20	-2	61	0.39	0.98	*HR*
F	-43	2	41	0.21	0.5081	**H**	-4	3	46	0.32	0.98	*RR*
F	-68	2	35	0.20	0.5082	**R**	5	1	42	0.46	0.99	*LH*
F	-79	1	33	0.26	0.5122	**L**	6	-1	42	0.79	0.96	*RH*
F	-49	4	37	0.23	0.4839	**F**	-7	1	48	0.46	0.96	*HF*

Table 4.6: Subset H_L# ((speaker H, preceding tone L) with qTA parameters (*m*, *b* and *λ*) and mean evaluation values (RMSE and Correlation) from parameter training. Duration_ratio is used for Simulation 2 in shortening duration of the non-contracted to that of the contracted. The last column is the random tone dyads for Simulation 3 detailed in Sec. 4.1.3.C.

1st syllable	**m**	**b**	**λ**	**RMSE (st)**	**Duration_ratio**	**2nd syllable**	**m**	**b**	**λ**	**RMSE (st)**	**Correlation**	***Random***
H	-3	4	52	0.20	0.4965	**H**	-13	4	35	0.24	0.98	*RF*
H	-2	5	44	0.31	0.5098	**R**	17	0	37	0.43	0.98	*FH*
H	-9	5	51	0.26	0.5118	**L**	14	-2	39	0.36	0.99	*LR*
H	7	4	53	0.16	0.5118	**F**	-35	5	38	0.19	0.99	*RL*
R	20	1	17	0.16	0.5192	**H**	6	5	39	0.30	0.98	*LF*
R	29	4	32	0.21	0.5071	**R**	-4	2	56	0.22	0.98	*FL*
R	39	2	24	0.15	0.5259	**L**	-11	0	52	0.30	0.98	*FR*
R	31	0	26	0.25	0.5159	**F**	-29	3	52	0.24	0.98	*HH*
L	5	-4	44	0.28	0.4887	**H**	12	2	59	0.30	0.98	*LL*
L	-3	-5	31	0.30	0.4995	**R**	36	-2	39	0.34	0.97	*HL*
L	38	2	41	0.15	0.5300	**L**	-28	-1	60	0.21	0.98	*FF*
L	2	-4	47	0.26	0.4930	**F**	-6	0	47	0.22	0.92	*HR*
F	-29	5	40	0.20	0.5031	**H**	-5	4	40	0.17	0.97	*RR*
F	-46	0	43	0.35	0.5079	**R**	8	0	55	0.45	0.98	*LH*
F	-62	3	37	0.23	0.5322	**L**	-14	0	50	0.37	0.99	*RH*
F	-9	3	47	0.22	0.4946	**F**	-20	3	47	0.29	0.94	*HF*

Table 4.7: Subset S_L# (speaker S, preceding tone L) with qTA parameters (*m*, *b* and *λ*) and mean evaluation values (RMSE and Correlation) from parameter training. Duration_ratio is used for Simulation 2 in shortening duration of the non-contracted to that of the contracted. The last column is the random tone dyads for Simulation 3 detailed in Sec. 4.1.3.C.

1st syllable	**m**	**b**	**λ**	**RMSE (st)**	**Duration_ratio**	**2nd syllable**	**m**	**b**	**λ**	**RMSE (st)**	**Correlation**	***Random***
H	7	3	60	0.18	0.4715	**H**	0	2	25	0.39	0.96	*RF*
H	7	5	60	0.21	0.4957	**R**	22	0	33	0.38	0.98	*FH*
H	10	5	57	0.22	0.4809	**L**	-19	-3	35	0.37	0.99	*LR*
H	7	5	59	0.17	0.4951	**F**	-30	3	63	0.24	0.99	*RL*
R	7	2	28	0.13	0.4785	**H**	11	2	36	0.38	0.98	*LF*
R	22	0	26	0.12	0.4918	**R**	45	-2	31	0.30	0.97	*FL*
R	24	3	39	0.20	0.4984	**L**	-27	-3	48	0.49	0.98	*FR*
R	18	1	34	0.15	0.4982	**F**	-24	3	62	0.26	0.98	*HH*
L	13	-6	27	0.26	0.4786	**H**	13	3	68	0.37	0.98	*LL*
L	-7	-4	30	0.21	0.4722	**R**	21	0	59	0.29	0.97	*HL*
L	12	2	40	0.18	0.4741	**L**	-29	-3	38	0.41	0.98	*FF*
L	30	-8	26	0.29	0.4709	**F**	-13	2	68	0.38	0.98	*HR*
F	-37	4	47	0.20	0.4964	**H**	-2	4	46	0.36	0.98	*RR*
F	-59	4	41	0.20	0.5069	**R**	3	2	45	0.35	0.98	*LH*
F	-58	4	44	0.27	0.5222	**L**	-21	1	46	0.40	0.99	*RH*
F	-29	4	53	0.25	0.4701	**F**	-24	2	63	0.30	0.97	*HF*

Chapter 5

General conclusion

Phonetic variability is known to be ubiquitous in natural speech, and much of this variability is related to phonetic reduction. Without understanding the mechanisms underlying phonetic reduction, the fundamental issue of invariance and variability cannot be resolved. In this thesis, experimental and corpus data sets were analysed and computation modelling was performed to explore the underlying mechanisms of one form of extreme reduction, known as contraction, in Taiwan Mandarin. Analysis was conducted first on segmental reduction and then on tonal reduction. The focus of the analysis was to evaluate the general hypothesis that *time pressure is the direct cause of extreme reduction.* Based on this hypothesis a number of predictions regarding the effect of time pressure on contraction were made and then tested in Chapters 2 and 3. Many of the predictions were confirmed, thus providing support for the main hypothesis stated

above. In Chapter 4 the effects of durational changes on the degree of tonal reduction were further demonstrated through computational modelling. Results of this thesis help elucidate certain issues faced by theories of phonetic variation, including popular proposals such as 1) exemplar-based models and 2) the H&H theory.

5.1 Existing accounts

5.1.1 Exemplar-based models

To explain the discrepancy between canonical forms and their varied forms of phonetic realisation, the simplest explanation can be offered by an exemplar-based account (Bybee, 2002; Goldinger, 1998, 2007; Hawkins, 2003; Johnson, 1997a, b and 2007; Pierrehumbert, 2001). In an exemplar-based model of speech production/perception, all variant forms are stored in the lexicon, and variations are an integral part of lexical representation. Therefore, in principle, there is no issue of variability in these models. However, as suggested by Plug (2005), there is still a need to find a level of phonetics at which perceptual representations can be translated into motor commands. Regarding phonetic implementation, in a comprehensive review of the roles of abstractions and exemplars in speech processing, Ernestus (in press) advocates a hybrid model and suggests that duration variants are also stored as exemplars and speakers make choices in terms of the type of articulatory gestures depending on the time available for the articulation of the word. Such a suggestion not only lays the burden of maintaining the highly gradient articulatory variations to the memory, but also

leaves no room for possible direct mechanical relation between duration and phonetic reduction.

5.1.2 H&H theory

Alternatively, there could exist a more causal relationship between invariance and variability as suggested by the H&H theory (Lindblom, 1990), which is based on the principle of economy of effort and has been applied by various frequency accounts and social and speaking style accounts. Before discussing the findings reported in this thesis, first recall Lindblom's key characterization of the principle (1990, p. 417): *"……within limits speakers appear to have a choice whether to undershoot or not to undershoot. We also noted that avoiding undershoot at short segment durations entails a higher biomechanical cost".* To a certain degree, this thesis is in accordance with this statement, that is, speakers aim to achieve targets within a limited duration by increasing effort so as to avoid undershoot. However, two things should be noted. First, what the present thesis has observed is a form of reduction that is more severe than that reported in Lindblom's studies, and it has been found that when duration is extremely short, there appears to be a barrier beyond which speakers simply have no choices but to undershoot, sometimes severely, the desired targets. This constraint may be related to what Lindblom meant by "*within limits*", but what exactly what those *limits* are requires clarification.

Secondly, it should be noted that H&H is based on Moon and Lindblom (1994), which did not calculate peak velocity over movement amplitude as a measure of articulatory effort and there was confounded assessment of articulatory effort (as mentioned in Sec. 1.3). Therefore, it is uncertain whether an increased effort was actually observed when speakers produced reduced items in Moon and Lindblom (1994). Instead, they only observed that the relation between duration and phonetic reduction is less linear in clear than in normal speech, which could actually be the result of reduced effort when speech rate is lowered in clear speech. On the other hand, the finding presented in the present thesis that articulatory effort cannot offset the effect of duration-dependency suggests that a non-dominant role is played by articulatory effort when it comes to selecting phonetic target or determining the range of variations, as implied by the predictions of the exemplar-based models (Plug, 2005). That is to say, speech communication is conducted in a goal-orientated manner within a social framework rather than by an intentional motor behaviour (Rischel, 1991).

5.1.3 Are we really concerned with saving effort while talking?

Furthermore, one potential difficulty of the effort account of phonetic reduction is that it seems to be inconsistent with the ease with which we keep talking continuously in our everyday life. More often than not, the reason we talk is not for the sake of conveying a message, but often we simply enjoy talking or feel comforted by the process of conversing with someone. As Levelt (1989, p. xiii) observed, "*Talking is one of our dearest occupations. We spend hours a day*

conversing, telling stories, teaching, quarrelling… and, of course, speaking to ourselves". It is probably true that some physical activities, such as running or climbing hills, are constrained by 'economy of effort', but this is because they are high energy-consuming activities. That speech production is also a strenuous activity for human beings just like running has never been clearly demonstrated.

There have been, however, some rather indirect investigations. For example, Parnell and Amerman (1977) consulted subjects' subjective judgements of articulatory effort in producing pairs of CV syllables and found that voicing and fricatives were judged to require greater articulatory effort. Also, with the assumption that less strenuous articulations are induced when speakers are tired or impaired, Kaplan (2010) compared the speech of subjects who were intoxicated with the speech of the same subjects when they were sober. His results suggested a compression of the articulatory space in the intoxicated condition, thus linking weakened articulatory effort (under intoxication) to phonological patterns observed in an Optimality Theory framework. These efforts, being indirect and non-quantitative, also have not provided clear answers to the question of whether articulatory effort is the key to understanding phonetic reduction.

5.1.4 Duration and articulatory effort

The key, in our view, probably lies in the recognition that Nelson's definition of impulse cost (Eq. 1.1, p. 10), which has been used as an indicator of articulatory effort, actually consists of two components, namely, force and duration. Adopting

this notion as the basis of judging articulatory effort is actually incompatible with Lindblom's own proposal regarding articulatory effort in H&H theory. That is, duration and articulatory effort are two separate and mutually independent variables. Also, most effort-oriented accounts do not separate the contribution of duration and articulatory effort. To exacerbate the confusion, duration is sometimes used as a measure of speech 'effort'. For example, van Son and Pols (1999) measured duration as an indicator of speech effort in the production of consonants in a VCV structure in recited and spontaneous speech. They concluded that 'speech effort' (which is dependent on duration along with other factors) of the intervocalic consonant was weaker in spontaneous speech than in recited speech. But if force and duration are independent factors, their findings actually provide evidence only for a close correlation between duration and phonetic reduction, because they had obtained no independent measurements of articulatory force. The close correlation between duration and phonetic reduction is also seen in many other studies (e.g. vowel reduction: de Jong et al., 1993; Lindblom, 1963, 1990; Moon and Lindblom, 1994; consonant reduction: de Jong et al., 1993; Fougeron and Keating, 1997; van Son and Pols, 1999; Warner and Tucker, 2007; tonal reduction: Myers and Li, 2009; Berry, 2009).

In a recent attempt to improve the computational model of Articulatory Phonology (Browman and Goldstein, 1990 and 1992), Simko and Cummins (2009) report that a cost function with only two cost components, one representing production cost and the other parsing cost (i.e. to replicate the trade-off between articulatory effort minimization and perceptual parsing cost suggested by the

H&H theory), is not sufficient to simulate variations along a hyper-hypo continuum. They found that a third cost component – duration, is an imperative element that needs to be included in the model. Additionally, in Simko and Cummins (2011), a premium is placed on the duration of an utterance in sequencing articulatory gestures due to the fact that duration cost "*represents a global constraint imposed on the manner of speech production reflecting an intentional choice of the speaker with respect to speaking rate*" (p. 547).

Furthermore, our measurements of articulatory effort suggest that the role of effort is not as clear as duration in realizing varying degrees of phonetic realisation. This result somewhat echoes the suspicion of Bauer (2008) and Simpson (2001) regarding the articulatory explanation of phonetic reduction. Research on phonetic reduction often reports a lenition process such as plosive > affricative > fricative > approximant > elision, where it is assumed that a segment in connected speech is degraded on a strength hierarchy from stronger to weaker. However, it is known that, in comparison to a plosive, a fricative is actually a more demanding muscular movement for a speaker owing to the subtle aerodynamic effects required (Ladefoged and Maddieson, 1996, p. 137; Perkell, 1997, p. 352). This 'degradation process' of a strength hierarchy, to some extent, also challenges the account of articulatory economy in phonetic reduction. Bauer (2008) thus re-defines lenition as the failure to reach a phonetically specified target (i.e. articulatory undershoot or underachievement). This suggests that the reduced forms, e.g. a realisation of fricatives or approximants, are not intended as such but are rather byproducts in connected speech. Interestingly, Bauer (2008, p. 165) also

proposes that "*changes of duration should still continue to be counted as matters involving lenition/fortition*".

5.2 Current results

The current results are from experimental, corpus and modelling data. In Chapter 2, three studies (i.e. Study 1, Study 2a and Study 2b) were conducted to investigate the time pressure account of phonetic variation in reduced speech. In Study 1, native Taiwan Mandarin speakers produced sentences containing nonsense disyllabic words with varying phonetic structures at differing speech rates. Spectral analysis showed that extreme reduction occurred frequently in nonsense words produced under high time pressure. In Study 2a, further examination of formant peak velocity as a function of formant movement amplitude in experimental data suggested that articulatory effort was not decreased during reduction, but in fact likely to be increased. Study 2b examined high frequency words from three spontaneous speech corpora for reduction variations. Results demonstrate that patterns of reduction in high frequency words in spontaneous speech (Study 2b) were similar to those in nonsense words spoken under experimental conditions (Study 2a). That is, duration is the direct cause of extreme phonetic reduction, which can be observed in both nonsense and high-frequency words.

In Chapter 3, Study 3 investigated tonal reduction with varying tonal contexts and found that tonal reduction can also be explained in terms of time pressure.

Analysis of F_0 trajectories demonstrates that speakers attempt to reach the original underlying tonal targets even in cases of extreme reduction and that there was no weakening of articulatory effort despite the severe reduction. To further test the time pressure account, in Chapter 4 two computational modelling experiments were presented. The first applied the quantitative Target Approximation model (qTA) for tone and intonation and the second applied the Functional Linear Model (FLM). Results showed that severely reduced F_0 trajectories in tone dyads can be simulated to a high accuracy by qTA using generalized canonical tonal targets with only modification of syllable duration. Additionally, it was shown using FLM, adjusting duration alone can give a fairly good representation of contracted F_0 trajectory shapes.

Overall, current results suggest that target undershoot under time pressure is likely to be the direct cause of extreme reduction, whereas factors that have been commonly associated with reduction in previous research are very likely to have impacts on duration, which in turn determines the degree of target attainment through the time pressure mechanism.

5.3 Evidence of direct duration control for encoding information and its consequence for target attainment

As it happens, variability in duration is not only due to factors often discussed in the reduction literature, such as word frequency, speaking style and social factors. Rather, duration is also extensively used, according to the findings of many

empirical studies, to encode different kinds of linguistic information, as will be briefly reviewed below.

5.3.1 Syllable grouping and final lengthening

Studies of prosody and intonation have shown that speakers adjust the local duration of each element within a basic prosodic unit so as to demarcate different levels of prosodic components as well as to encode inter-constituent affinity. For example, from an aspect of lower-level prosody, the duration of a syllable varies according to the syllable's position in a prosodic unit and relative location to a prosodic boundary. Research has indicated that phrase-medial segments are shorter than those in phrase-initial and phrase-final positions (e.g. Lindblom and Rapp, 1973). In an analysis of French vowel reduction in conversational speech, Meunier and Espesser (2011) reported a strong correlation between durational and spectral reduction. Their results show that vowels in the final syllables of words were less often reduced while the preceding ones show reduced durations and centralized formant values. Research by Xu and Wang (2009) on tonal reduction as a function of syllable grouping shows that in comparison to other acoustic correlates such as F_0 displacement, v_p/d ratio (ratio of peak velocity to F_0 displacement) and the parameter C (shape of F_0 velocity profile), syllable duration exhibits the most consistent grouping-related patterns. That is, in a short phrase of 1-4 syllables, duration is shortest in the medial positions, longest in the final position and the second longest in the initial position, which exactly parallels the pattern of tonal reduction. The finding of Xu and Wang (2009) is in line with

Klatt (1975) who proposed that increasing the duration of terminal segments enables a more effective manifestation of F_0 movements signalling terminal or non-terminal breaks.

5.3.2 Focus, new topic and second mention

It is well known that many languages use F_0 as a prominent cue in signalling a focus function in an utterance (English: Cooper et al., 1985; Mandarin: Xu and Xu, 2005; Arabic: Yeou, 2004; Dutch: Ladd et al., 2000). It has also been found that focus encoding involves duration expansions of the focused component. In fact, some research has disentangled the role of F_0 and duration as phonetic correlates of focus prominence (Kügler, 2008) indicating that duration alone can be a functionally relevant prosodic cue. Furthermore, the understanding that adjusting duration can jointly signal certain pragmatic functions is also seen in the study of Wang and Xu (in press) on prosodic encoding of topic and focus in Mandarin. In addition to the main findings that focus and topic can be encoded simultaneously (i.e. focus is encoded with an *expanded* pitch range and new topic a *raised* pitch range), they found that a newly raised topic involves a slightly longer duration in comparison to an established one. Similar cases in which duration is used to encode linguistic information can also be found in research regarding word probability (i.e. frequency and number of mention). For example, in addition to the expected word frequency effect that words of higher frequency are significantly shorter than those of lower frequency words, Baker and Bradlow (2009) showed that a speaker tends to reduce the second mention of a word in

both clear and plain speech styles. They conclude that this is indication of a direct link between duration and probability, rather than a relationship solely mediated by prosodic prominence.

Given the apparently heavy information load of duration, it is unlikely that speakers have much room for making free choices in duration just for the sake of controlling the amount of phonetic reduction. It is more likely that local durations are mostly determined by the informational factors mentioned above, and speakers at most have the choice of globally slowing down if it is necessary to reduce the amount of phonetic reduction, as in the case of clear speech. The validity of this assumption, of course, awaits evidence from future research.

5.4 Concluding remarks

In summary, this thesis has provided evidence in support of the hypothesis that time pressure is the direct cause of extreme phonetic reduction. It has also provided evidence that an increased articulatory effort (measured as slope of peak velocity of unidirectional movement against movement amplitude) is insufficient to compensate for duration-dependent undershoot (in particular, when time pressure exceeds certain thresholds). We conclude that these results support the idea that time pressure is the most critical factor determining the occurrence of extreme phonetic reduction such as contraction. The explanation that less articulatory effort is involved in reduced items was not compatible with the current results.

Recently, more attention has been drawn to socially-constrained variations with particular attention being given to recognizing that speakers use phonetic variation as a resource to achieve a range of social goals (Byrd, 1994; Hawkins, 2003 and 2010, Labov, 2006; Local, 2003, Foulkes et al., 2010). That is, in addition to accounting for systematic variations from a 'purely linguistic' point of view (Foulkes and Docherty, 2006), a more complete model would also need to consider variations arising from a sociolinguistic perspective since the crucial factors determining intelligibility are the quality of the linguistic model driving the system (Ogden et al., 2000). Warner (2011) adeptly points out that there is no definitive explanation regarding the driving force behind reduction but suggests that: "*It seems very likely that articulatory factors (e.g. task dynamic stiffness, articulatory movement rate), information structure (greater reduction where information is less important), and intentional use of reduction as a feature that conveys information in itself all contribute to how much reduction a given utterance contains"* (p. 1881). Investigating a problem with such a large parameter space is difficult and poses many challenges, but it is hoped that the research presented here will help elucidate the relative importance of certain factors controlling phonetic reduction from a mechanical perspective. As mentioned in parts of this chapter, a great deal of research originating from differing perspectives has pointed towards the importance of duration in phonetic reduction. Continued effort and research in this subject area will help improve our understanding of this perennial problem in speech science.

Appendix

Validity of acoustic measurements to infer articulatory dynamics

The research by Lindblom (1963 and 1990; Moon and Lindblom, 1994), based on which his theories regarding articulatory effort were developed, mostly involved acoustic measurements, such as formant frequencies. Gay (1978) and van Son and Pols (1990 and 1992), who countered Lindblom's findings, also used acoustic evidence. However, most other studies attempting to examine articulatory effort have been based primarily on kinematic measurements of individual articulators (e.g. tongue dorsum/body: Ostry and Munhall, 1985; Perkell et al., 2002, Perkell and Zandipour, 2002; tongue blade: Perkell et al., 2002, Perkell and Zandipour, 2002; tongue tip: Adams et al., 1993; lips/jaw: Kelso et al., 1985; Adams et al., 1993; Harrington et al., 1995; Hertrich and Ackermann, 1997; Perkell et al., 2002, Perkell and Zandipour, 2002). By now, there seems to be a consensus that it is inappropriate to link formant movements to articulatory movements owing to a lack of a one-to-one relation between articulation and acoustics. Moreover, it seems that a tacit assumption behind this consensus is that articulatory measurements *do* have a one-to-one relation with phonetically relevant articulation because they directly measure movements of specific articulators. This assumption requires careful scrutiny, especially in light of what speech production is about.

There is little doubt that a major goal of speech production is to generate acoustic patterns that can be recognized as phonetic categories such as vowels and consonants. It is well established that vowel identity can be adequately represented by formant patterns (Chiba and Kajiyama, 1941; Fant, 1960; Hillenbrand et al., 1995; Ladefoged et al., 1987; Peterson and Barney, 1952; Stevens, 1998), which is further attested by the success (albeit partial) of formant synthesizers (Klatt, 1987). Furthermore, it is also known that vowel formants are determined by the shape of the entire vocal tract rather than by the shape of only a particular location of the vocal tract. For example, according to the perturbation theory (Chiba and Kajiyama 1941; Stevens, 1998), for the vowel [i] the narrow constriction between the front of the tongue and the hard palate, where there is an antinode for F2, must also be accompanied by the widening of the pharynx where there is a node for F2. Otherwise F2 would not have been as high as it is usually observed for [i] (Hillenbrand et al., 1995; Peterson and Barney, 1952). In fact, it has been shown that F2 is more sensitive to pharyngeal width than to constriction at the tongue blade (Fant and Pauli, 1974; Wood, 1986). Thus the high F2 of [i] is the result of at least two articulatory manoeuvres: tongue-blade raising and tongue-root fronting. It has also been shown that even the vertical position of the larynx differs across vowels in a manner that would enhance their formant differences (Demolin et al., 2000; Hoole and Kroos, 1998; Wood, 1986). For consonants, the phonetically relevant articulation should also take aerodynamics into consideration. To produce a [t] for example, not only the tip of the tongue should be raised against the alveolar ridge, but also the sides of the tongue need to be elevated to guarantee an airtight closure.

Thus to capture the phonetically relevant articulatory configurations and movements, it is necessary to have measurements that can indicate the shape of the entire vocal tract. In this sense, measurements of individual articulators, such as tongue tip, tongue blade, tongue dorsum, the lips and the jaw, etc., do not really have a one-to-one relation to the phonetically relevant articulation as a whole. Instead, the movement of any particular articulator is not for its own sake, but to serve as part of a collective functional manoeuvre that can be described as a task-specific 'coordinative structure' (Saltzman and Kelso, 1987). As a whole, such a functional unit achieves overall aerodynamic and acoustic effects (Mattingly, 1990; Hanson and Stevens, 2002) which constitute the phonetic category jointly. As a result, specific articulatory kinematic measurements can provide only a partial approximation of the ensemble underlying goal-oriented articulatory movements. That is, they may not fully reflect the dynamic constraint required in achieving a functional articulatory goal.

Similarly, acoustic measurements such as formant trajectories also provide only a partial approximation of the underlying goal-oriented movements. However, any phonetically relevant articulatory movements necessarily have to be reflected in the acoustics, as otherwise they would not have been audible. More importantly, the perturbation theory (Fant, 1960; Stevens, 1998) would predict that only the lower formants (up to F3) are individually controllable, since direct control of the higher formants would require separate manoeuvres of too many parts of the vocal tract simultaneously. As a result, little critical information is missing if only the first few formants are measured. In general, the dynamics of the first three

formants do reflect a significant portion of the phonetically relevant articulatory movement. Interestingly, Hertrich and Ackermann (1997) and Perkell et al. (2002), after careful examinations of articulatory dynamics, both suggested that the phonetically most relevant information may be found in the acoustic signal.

The lack of a one-to-one relation between any measurements and actual articulation can also be seen in the fact that measurements are often necessarily sparse. When measuring tongue movement with the magnetometer system such as EMMA and x-ray microbeam technologies, only a limited number of sensors or pellets can be placed on the tongue surface (Byrd et al., 1995). Yet the assumption is that, unless the sensors are too far apart, it is safe to assume that no sudden deformation would occur in between. The same is true with formant movements. Unless there is a sudden shift of the resonant cavity, as occurs during the oral to nasal shift, or sudden shift of formant affiliation with a particular resonator (Stevens, 1998), or an interruption due to the closure and frication of obstruent consonants, formants movements are largely smooth because the corresponding articulatory movement is largely smooth.

The comparability of articulatory and acoustic measurements can be empirically attested by examining whether similar dynamic patterns can be seen in acoustic and articulatory movements. At least for fundamental frequency, highly linear relations between F_0 velocity and F_0 movement amplitude have been found (Xu and Sun, 2002; Xu and Wang, 2009), which resemble the linear relations in articulatory or limb movement (Hertrich and Ackermann, 1997; Kelso *et al.*, 1985;

Ostry and Munhall, 1985; Vatikiotis-Bateson and Kelso, 1993). This is despite the fact that F_0 is the output of a highly complex laryngeal system (Honda, 1995; Zemlin, 1988). It will therefore also be an empirical question as to whether formant kinematics also exhibit similar linear relations to warrant dynamic analyses that have been applied to limb and F_0 movements, which is addressed in Chapter 2 of this thesis.

Bibliography

Adams, S. G., Weismer, G., & Kent, R. D. (1993). Speaking rate and speech movement velocity profiles. *Journal of Speech and Hearing Research, 36,* 41-54.

Adank, P. & Janse, E. (2009). Perceptual learning of time-compressed and natural fast speech. *Journal of the Acoustical Society of America*, *126*(5), 2649-2659.

Aylett, M. & Turk, A. (2006). Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *Journal of the Acoustical Society of America*, *119*(5), 3048-3058.

Bahill, A. T., Kallman, J. S., & Lieberman, J. E. (1982). Frequency limitations of the two-point central difference differentiation algorithm. Biological Cybernetics, 45, 1-4.

Baker, R. E. & Bradlow, A. R. (2009). Variability in word duration as a function of probability, speech style, and prosody. *Language and Speech*, *52*(4), 391-413.

Bauer, L. (2008). Lenition revisited. *Journal of Linguistics*, *44*, 605-624.

Barry, W. & Andreeva, B. (2001). Cross-language similarities and differences in spontaneous speech patterns. *Journal of the International Phonetic Association*, *31*(1), 51-66.

Berry, J. (2009). *Tone space reduction in Mandarin Chinese.* Unpublished manuscript, University of Arizona.

Berkovits, R. (1994). Durational effects in final lengthening, gapping, and contrastive stress. *Language and Speech*, *37*(3), 237-250.

Boersma, P. & Weenink, D. (2010). Praat: Doing phonetics by computer [Computer program]. Version 5.1.44, retrieved from http://www.praat.org/.

Browman, C. P. & Goldstein, L. (1990). Tiers in articulatory phonology, with some implications for casual speech. In J. Kingston & M. E. Beckman (Eds.), *Between the Grammar and Physics of Speech: Papers in*

Laboratory Phonology I (pp. 341-376). Cambridge: Cambridge University Press.

Browman, C. P. & Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica*, *49*, 155-180.

Bybee, J. L. (2002). Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change*, *14*, 261-290.

Byrd, D. (1994). Relations of sex and dialect to reduction. *Speech Communication*, *15*, 39-54.

Byrd D., Browman C. P., Goldstein, L. and Honorof, D. (1995). EMMA and x-ray microbeam comparison. *Journal of the Acoustical Society of America*, *97*(5), 3365-3365.

Chao, Y. R. (1930). A system of tone-letters. *Le Maître Phonétique*, *45*, 24-27.

Chen, Y. (2006). Durational adjustment under corrective focus in Standard Chinese. *Journal of Phonetics*, *34*, 176-201.

Cheng, C. E. (2004). *An acoustic phonetic analysis of tone contraction in Taiwan Mandarin.* MA, National Cheng Chi University, Taipei.

Cheng, C. & Xu. Y. (2008a). (When and) How are disyllables contracted into monosyllables in Taiwan Mandarin. The Second ASA-EAA Joint Conference – Acoustics ’08 Paris, France.

Cheng, C. & Xu, Y. (2008b). When (and How) are disyllables contracted into monosyllables in Taiwan Mandarin. British Association of Academic Phonetician ’08 Colloquium, University of Sheffield, U.K.

Cheng, C. & Xu. Y. (2009). Extreme reductions: Contraction of disyllables into monosyllables in Taiwan Mandarin. In *Proc.INTERSPEECH-2009*, 456-459.

Cheng, C., Xu, Yi., & Gubian, M. (2010). Exploring the mechanism of tonal contraction in Taiwan Mandarin. In *Proc. INTERSPEECH-2010*, 2010-2013.

Cheng, C., Xu, Y., & Prom-on, S. (2011). Modelling extreme tonal reduction in Taiwan Mandarin based on target approximation. In *Proc. ICPhS-XVII*.

Cheng, C. & Gubian, M. (2011). Predicting Taiwan Mandarin tone shapes from their duration. In *Proc. INTERSPEECH-2011*.

Chiba, T. & Kajiyama, M. (1941). *The Vowel, Its Nature and Structure.* Tokyo: Kaiseikan.

Chung, K. S. (2006). Contraction and backgrounding in Taiwan Mandarin. *Concentric: Studies in Linguistics*, *32*(1), 69-88.

Cooper, W. E., Eady, S. J., & Mueller, P. R. (1985). Acoustical aspects of contrastive stress in question-answer contexts. *Journal of the Acoustical Society of America*, *77*(6), 2142-2156.

Dankovičová, J. (1997). The domain of articulation rate variation in Czech. *Journal of Phonetics*, *25*(3), 287-312.

Dankovičová, J. & Nolan, F. (1999). Some acoustic effects of speaking style on utterances for automatic speaker verification. *Journal of the International Phonetic Association*, *29*(2), 115-128.

de Jong, K. J., Beckman, M. E., & Edwards, J. R. (1993). The interplay between prosody and coarticulation. *Language and Speech*, *36*(2-3), 197-212.

Demolin, D., Metens, T., & Soquet, A. (2000). Real time MRI and articulatory coordinations in vowels. In *Proc. 5th Speech Production Seminar*, 86-93.

Duanmu, S. (2000). *The Phonology of Standard Chinese*. New York: Oxford University Press.

Engstrand, O. (1988). Articulatory correlates of stress and speaking rate in Swedish VCV utterances. *Journal of the Acoustical Society of America*, *83*(5), 1863-1875.

Engstrand, O. & Krull, D. (2001). Segment and syllable reduction: Preliminary observations. *Lund University, Department of Linguistics, Working Papers, 49*, 26-29.

Ernestus, M. (in press). Acoustic reduction and the roles of abstractions and exemplars in speech processing. *Lingua*.

Fant, G. (1960). *Acoustic Theory of Speech Production*. The Hague: Mouton & Co.

Fant, G. & Pauli, S. (1974). Spatial characteristics of vocal tract resonance modes. In *Proc. Speech Communication Seminar-74*, 121-132.

Fougeron, C. & Keating P. A. (1997). Articulatory strengthening at edges of prosodic domains. *Journal of the Acoustical Society of America, 101*(6), 3728-3740.

Fourakis, M. (1991). Tempo, stress and vowel reduction in American English. *Journal of the Acoustical Society of America, 90*(4), 1816-1827.

Fosler-Lussier, E. & Morgan, N. (1999). Effects of speaking rate and word frequency on pronunciations in conversational speech. *Speech Communication, 29,* 137-158.

Foulkes, P. & Docherty, G. (2006). The social life of phonetics and phonology. *Journal of Phonetics, 34*, 409-438.

Foulkes, P., Scobbie, J. M., & Watt, D. (2010). Sociophonetics. In W. J. Hardcastle, J. Laver., & F. E. Gibbon (Eds.), *The Handbook of Phonetic Sciences* (pp. 703-754). Wiley-Blackwell.

Fowler, C. A. & Housum, J. (1987). Talkers' signaling of "new" and "old" words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language, 26,* 489-504.

Fujisaki, H., Wang, C., Ohno, S., & Gu, W. (2005). Analysis and synthesis of fundamental frequency contours of Standard Chinese using the command-response model. *Speech Communication, 47*, 59-70.

Gahl, S. & Garnsey, S. M. (2004). Knowledge of grammar, knowledge of usage: Syntactic probabilities affect pronunciation variation. *Language, 80*(4), 748-775.

Gandour, J., Potisuk, S., & Dechongkit, S. (1994). Tonal Coarticulation in Thai. *Journal of Phonetics, 22*(4), 477-492.

Gauthier, B. Shi, R., Xu, Y. (2007). Learning phonetic categories by tracking movements. *Cognition, 103*(1), 80-106.

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

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. Psychological Review, *105,* 251-279.

Goldinger, S. D. (2007). A complementary-systems approach to abstract and episodic speech perception. In *Proceedings of the 16th International Congress of Phonetic Sciences*, 49-54.

Gregory, M., Raymond, W. D., Bell, A., Fosler-Lussier, E., & Jurafsky, D. (1999).The effects of collocational strength and contextual predictability in lexical production. *Chicago Linguistic Society, 35*, 151-166.

Gubian, M., Boves, L., & Cangemi, F. (2011). Joint analysis of F_0 and speech rate with Functional Data Analysis. In *Proc. International Conference of Acoustics, Speech and Signal Processing-2011 (ICASSP).*

Hall, N. (2010). Articulatory phonology. *Language and Linguistics Compass*, *4*(9), 818-830.

Hawkins, S. (2003). Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics, 31*, 373-405.

Hawkins, S. (2010). Phonetic variation as communicative system: Perception of the particular and the abstract. In C. Fougeron, B. Kühnert, M. D'Imperio, & N. Vallée (Eds.), *Laboratory phonology 10.4* (pp. 479-510). Berlin: Mouton de Gruyter.

Harrington, J., Fletcher, J., & Roberts, C. (1995). Coarticulation and the accented/unaccented distinction: Evidence from jaw movement data. *Journal of Phonetics, 23*, 305-322.

Hanson, H. M. & Stevens, K. N. (2002). A quasiarticulatory approach to controlling acoustic source parameters in a Klatt-type formant synthesizer using HLsyn. *Journal of the Acoustical Society of America*, *112*(3), 1158-1182.

Hertrich, I. & Ackermann, H. (1997). Articulatory control of phonological vowel length contrasts: Kinematic analysis of labial gestures. *Journal of the Acoustical Society of America*, *102*(1), 523-536.

Hillenbrand, J., Getty, L., Clark, M., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, *97*(5), 3099-3111.

Honda, K. (1995). Laryngeal and extra-laryngeal mechanisms of F0 control. In F. Bell-Berti, & Raphael, L. (Eds), *Producing Speech: Contemporary Issues: For Katherine Safford Harris* (pp. 215-245). New York: AIP Press.

Hoole, P. & Kroos, C. (1998). Control of larynx height in vowel production. In *Proc. ICSLP-1998*, 531-534.

Hsiao, C. (1986). A preliminary investigation into phonetic contraction in Mandarin. *The World of Chinese Language, 40*, 46-50.

Hsiao, Y. C. (2002). Tone contraction. In *Chinese Languages and Linguistics VIII* (pp. 1-16). Taipei: Academia Sinica.

Hsu, H. C. (2003). A sonority model of syllable contraction in Taiwanese Southern Min. *Journal of East Asian Linguistics, 12*, 349-377.

Johnson, K. (1997a). The auditory/perceptual basis for speech segmentation. *OSU Working Papers in Linguistics, 50*, 101-113.

Johnson, K., (1997b). Speech perception without speaker normalization: An exemplar model. In K. Johnson & J. W. Mullennix (Eds.), *Talker Variability in Speech Processing* (pp. 145-165), Academic Press, San Diego

Johnson, K. (2004). Massive reduction in conversational American English. In K. Yoneyama & K. Maekawa (Eds.), *Proceedings of the 10th international symposium: Spontaneous speech: Data and analysis* (pp. 29-54), Tokyo, Japan.

Johnson, K. (2007). Decisions and Mechanisms in Exemplar-based Phonology. In M.J. Sole, P. Beddor, and M. Ohala. (Eds.) *Experimental Approaches to Phonology. In Honor of John Ohala* (pp. 25-40), Oxford University Press.

Jurafsky, D., Bell, A., Gregory, M., & Raymond, W. D. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. In J. Bybee & P. Hopper (Eds.), *Frequency and the Emergence of Linguistic Structure* (pp. 229-254). Amsterdam: Benjamins.

Kaplan, A. (2010). *Phonology shaped by phonetics: The case of intervocalic lenition*. PhD. University of California, Santa Cruz.

Karlgren, H. (1962). Speech rate and information theory. In A. Sovijärvi & P. Aalto (Eds.), *Proceedings of the Fourth International Congress of Phonetic Sciences* (pp. 671-677). The Hague: Mouton & Co.

Keating, P. A. (1997). Word-level phonetic variation in large speech corpora. In Alexiadou, A., Fuhrop, N., Kleinhenz, U., & Law, P. (Eds) (1998), *ZAS Papers in Linguistics*, *11*, 35-50.

Kelso, J. A., Vatikiotis-Bateson, E., Saltzman, E. L., & Kay, B. (1985). A qualitative dynamic analysis of reiterant speech production: Phase portraits, kinematics, and dynamic modeling. *Journal of the Acoustical Society of America*, *77*(1), 266-280.

Kirchner, R. M. (1998). *An effort-based approach to consonant deletion*. PhD. University of California.

Klatt, D. H. (1973). *Interaction between two factors that influence vowel duration. Journal of the Acoustical Society of America*, *54*, 1102-1104.

Klatt, D. H. (1975). Vowel lengthening is syntactically determined in a connected discourse. *Journal of Phonetics*, *3*(3), 129-140.

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America*, *59*(5), 1208-1221.

Klatt, D. H. (1987). Review of text-to-speech conversion for English. *Journal of the Acoustical Society of America*, *82*(3), 737-793.

Kohler, K. J. (1990). Segmental reduction in connected speech in German: Phonological facts and phonetic explanations. In W. J. Hardcastle & A. Marchal (Eds.), *Speech Production and Speech Modelling* (pp. 69-92). Dordrecht: Kluwer Academic Publisher.

Kohler, K. J. (1998). The disappearance of words in connected speech. *ZAS Working Papers in Linguistics, 11*, 21-34.

Kohler, K. J. (2000). Investigating unscripted speech: Implications for phonetics and phonology. *Phonetica, 57*, 85-94.

Krause J. C. & Braida, L. D. (2002). Investigating alternative forms of clear speech: The effects of speaking rate and speaking mode on intelligibility. *Journal of the Acoustical Society of America*, *112*(5), 2165-2172.

Kuo, G. (2010). Production and perception of Taiwan Mandarin syllable contraction. *UCLA Working Papers in Phonetics*, *108*, 1-34.

Kuo, Y. C., Xu, Y., & Yip, M. (2007). The phonetics and phonology of apparent cases of iterative tonal change in Standard Chinese. In Gussenhoven, C. & Riad, T. (Eds.), *Tones and Tunes Vol 2: Experimental Studies in Word and Sentence Prosody* (pp. 211-237). Berlin: Mouton de Gruyter.

Kügler, F. (2008). The role of duration as a phonetic correlate of focus. In *Proc. Speech Prosody-2008*, 591-594.

Labov, W. (2006). A sociolinguistic perspective on sociophonetic research. *Journal of Phonetics*, *34*(4), 500-515.

Ladd, D. R., Mennen I., Schepman, A. (2000). Phonological conditioning of peak alignment in rising pitch accents in Dutch. *Journal of the Acoustical Society of America*, *107*, 2685-2696.

Ladefoged, P. & Maddieson, I. (1996). *The Sounds of the World's Languages*. Oxford: Blackwell.

Ladefoged, P., Maddieson, I., Jackson, M., & Huffman, M. K. (1987). Characteristics of the voice source. In *Proc. ECST*-1987, 2226-2229.

Levelt, C. C., Schiller, N. O., & Levelt, W. J. (1999). The acquisition of syllable types. *Language Acquisition*, *8*(3), 237-264.

Levelt, J. M. (1989). *Speaking: From intention to articulation*. MIT press.

Lindblom, B. (1963). Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America*, *35*(11), 1773-1781.

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

Lindblom, B. (2000). Developmental origins of adult phonology: The interplay between phonetic emergents and the evolutionary adaptations of sound patterns. *Phonetica*, *57*, 297-314.

Lindblom, B. & Rapp, K. (1973). Some temporal regularities of spoken Swedish. *Paper of the Linguistic University of Stockholm*, *21*, 1-59.

Local, J. (2003). Variable domains and variable relevance: Interpreting phonetic exponents. *Journal of Phonetics*, *31*, 321-339.

Mattingly, I. G. (1990). *The global character of phonetic gestures. Journal of Phonetics, 18*(3), 445-452.

Malécot, A. (1955). An experimental study of force of articulation. *Studia Linguistica, 9*, 35-44.

McCarthy, J. J. (2007). Slouching towards optimality: Coda reduction in OT-CC. In Phonological Society of Japan (Ed.), *Phonological Studies 10* (pp.89-104). Tokyo: Kaitakusah.

Meunier, C. & Espesser, R. (2011). Vowel reduction in conversational speech in French: The role of lexical factors. *Journal of Phonetics, 39*(3), 271-278.

Moon, S. J. & Lindblom, B. (1994). Interaction between duration, context, and speaking style in English stressed vowels. *Journal of the Acoustical Society of America, 96*(1), 40-55.

Morén, B. & Zsiga, E. (2006). The lexical and post-lexical phonology of Thai tones. *Natural Language & Linguistic Theory, 24*(1), 113-178.

Motulsky, H. & Christopoulos, A. (2003). Fitting models to biological data using linear and nonlinear regression. A practical guide to curve fitting. GraphPad Software Inc., San Diego CA, www.graphpad.com.

Myers, J. & Li, Y. (2009). Lexical frequency effects in Taiwan Southern Min syllable contraction. *Journal of Phonetics, 37*, 212-230.

Nelson, W. L. (1983). Physical Principles for Economies of Skilled Movements. *Biological Cybernetics, 46*, 135-147.

Niebuhr, O. & Kohler, K. J. (2011). Perception of phonetic detail in the identification of highly reduced words. *Journal of Phonetics, 39*(3), 319-329.

Ogden, R., Hawkins, S., House, J., Huckvale, M., Local, J., Carter, P., Dankovičová, J., & Heid, S. (2000). ProSynth : An integrated prosodic approach to device-independent, natural-sounding speech synthesis. *Computer Speech and Language, 14*, 177-210.

Ostry, D. J, Keller, E., & Parush, A. (1983). Similarities in the control of the speech articulators and the limbs: Kinematics of tongue dorsum movement in speech. *Journal of Experimental Psychology. Human Perception and Performance, 9*(4), 622-636.

Ostry, D. J. & Munhall, K. G. (1985). Control of rate and duration of speech movements. *Journal of the Acoustical Society of America*, *77*(2), 640-648.

Parnell, M. & Amerman, J. D. (1977). Subjective evaluation of articulatory effort. *Journal of Speech and Hearing Research*, *20*, 644-652.

Peng, S. H. (2000). Lexical versus 'phonological' representations of Mandarin Sandhi tones. In M. B. Broe & J. B. Pierrehumbert (Eds.), *Papers in Laboratory Phonology V: Acquisition and the Lexicon* (pp. 152-167). Cambridge: Cambridge University Press.

Perkekll, J. S. & Klatt, D. H. (1986). *Invariance and Variability in Speech Processes*. Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.

Perkell, J. S. (1997). Articulatory processes. In Hardcastle, W. J. & Laver, J. (Eds.) *The Handbook of Phonetic Sciences* (pp. 333-370). Oxford: Blackwell.

Perkell, J. S., Zandipour, M., Matthies, M. L., & Lane, H. (2002). Economy of effort in different speaking conditions. I. A preliminary study of intersubject differences and modeling issues. *Journal of the Acoustical Society of America*, *112*(4), 1627-1641.

Perkell, J. & Zandipour, M. (2002). Economy of effort in different speaking conditions. II. Kinematic performance spaces for cyclical and speech movements. *Journal of the Acoustical Society of America*, *112*(4), 1642-1651.

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.

Pierrehumbert, J. (2001). Exemplar dynamics: Word frequency, lenition and contrast. In J. Bybee & P. Hopper (Eds.), *Frequency Effects and the Emergence of Linguistic Structure* (pp. 137-157). Amsterdam: John Benjamins.

Plug, L. (2005). Phonetic reduction and categorisation in exemplar-based representation: Observations on a Dutch discourse marker. In *Proc. ConSOLE XIII*, 287-311.

Pluymaekers, M., Ernestus, M., & Baayen, R. H. (2005). Lexical frequency and acoustic reduction in spoken Dutch. *Journal of the Acoustical Society of America*, *118*(4), 2561-2569.

Prom-on, S., Xu, Y., & Thipakorn, B. (2009). Modeling tone and intonation in Mandarin and English as a process of target approximation. *Journal of the Acoustical Society of America*, *125*(1), 405-424.

R Development Core Team. (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, URL: http://www.R-project.org.

Ramsay, J. O. & Silverman, B. W. (2005). *Functional Data Analysis*, *Second Edition*. Springer.

Ramsay, J. O., Hookers, G., & Graves, S. (2009). *Functional Data Analysis with R and MATLAB.* Springer.

Rischel, J. (1991). The relevance of phonetics for phonology: A commentary. *Phonetica*, *48*, 233-262.

Saltzman, E. & Kelso, J. A. S. (1987). Skilled actions: A task dynamic approach. *Psychological Review*, *94*(1), 84-106.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Systems Technical Journal*, 27, 379-423, 623-656.

Sigurd, B. (1973). Maximum rate and minimal duration of repeated syllables. *Language and Speech*, *16*(4), 373-395.

Simpson, A. P. (2001). Does articulatory reduction miss more patterns than it accounts for? *Journal of the International Phonetic Association*, *31*(1), 29-39.

Simko, J. & Cummins, F. (2009). Sequencing of articulatory gestures using cost optimization. In *Proc. INTERSPEECH-2009*, 60-63.

Simko, J. & Cummins, F. (2011). Sequencing and optimization within an embodied task dynamic model. *Cognitive Science*, *35*, 527-562.

Stampe, D. (1973). *A Dissertation on Natural Phonology*. PhD. University of Chicago.

Stevens, K. N. (1998). *Acoustic Phonetics*. MIT Press.

Suihkonen, P. (2005). On the two-level model in description of phonological and morphophonological processes in Finnish Dialects. *Nordic Journal of African Studies*, *14*(4), 464-478.

Tatham, M. & Morton, K. (2006). *Speech Production and Perception*. New York: Palgrave Macmillan.

Tiffany, W. R. (1980). The effects of syllable structure on diadochokinetic and reading rates. *Journal of Speech and Hearing Research*, *23*, 894-908.

Tseng, S. C. (2005a). Contracted syllables in Mandarin: Evidence from spontaneous conversations. *Language and Linguistics*, *6*(1), 153-180.

Tseng, S. C. (2005b). Syllable contractions in a Mandarin Conversational Dialogue Corpus. *International Journal of Corpus Linguistics*, *10*(1), 63-83.

Tseng, S. C. (2008). Spoken corpora and analysis of natural speech. *Taiwan Journal of Linguistics*, *6*(2), 1-26.

Turk, B. V. (2007). *Spoken word recognition of the reduced American English flap.* PhD. University of Arizona.

Vance, T. J. (2008). *The Sounds of Japanese.* Cambridge: Cambridge University Press.

van Son, R. J. J. H. & Pols, L. C. W. (1990). Formant frequencies of Dutch vowels in a text, read at normal and fast rate. *Journal of the Acoustical Society of America*, *88*(4), 1683-1693.

van Son, R. J. J. H. & Pols, L. C. W. (1992). Formant movements of Dutch vowels in a text, read at normal and fast rate. *Journal of the Acoustical Society of America*, *92*(1), 121-127.

van Son, R. J. J. H. (1993). Spectro-temporal features of vowel segments. *Studies in Language and Language Use 3.* PhD. University of Amsterdam.

van Son, R. J. J. H. & Pols, L. C. W. (1999). An acoustic description of consonant reduction. *Speech Communication*, *28*, 125-140.

van Santen, J. (1994). Assignment of segmental duration in text-to-speech synthesis. *Computer Speech and Language*, *8*(2), 95-128.

van Santen, J. & Shih, C. (2000). Suprasegmental and segmental timing models in Mandarin Chinese and American English. *Journal of the Acoustical Society of America*, *107*(2), 1012-1026.

Vatikiotis-Bateson, E. & Kelso, J. A. S. (1993). Rhythm type and articulatory dynamics in English, French and Japanese. *Journal of Phonetics, 21*(3), 231-265.

Wang, B. & Xu, Y. (in press). Differential prosodic encoding of topic and focus at sentence initial position in Mandarin Chinese. *Journal of Phonetics.*

Warner, N. (2011). Reduction. Invited chapter. In M. van Oostendorp, C. Ewen, E. Hume, & K. Rice (Eds.), *The Blackwell companion to phonology* (pp. 1866-1891). Malden, MA & Oxford: Wiley-Blackwell.

Warner, N. & B. V. Tucker. (2007). Categorical and gradient variability in intervocalic stops. Paper presented at the 81st Annual Meeting of the Linguistic Society of America, Anaheim.

Wong, W. Y. P. (2004). Syllable Fusion and Speech Rate in Hong Kong Cantonese. In *Proc. Speech Prosody-2004*, 255-258.

Wong, W. Y. P. (2006). *Syllable Fusion in Hong Kong Cantonese Connected Speech.* PhD. The Ohio State University.

Wood, S. (1986). The acoustical significance of tongue, lip, and larynx maneuvers in rounded palatal vowels. *The Journal of the Acoustical Society of America*, *80*(2), 391-401.

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

Xu, Y. (1998). Consistency of tone-syllable alignment across different syllable structures and speaking rates. *Phonetica*, *55*, 179-203.

Xu, Y. (1999). Effects of tone and focus on the formation and alignment of F0 contours. *Journal of Phonetics, 27*, 55-105.

Xu, Y. (2001). Sources of tonal variations in connected speech. *Journal of Chinese Linguistics, 17*, 1-31.

Xu, Y. (2005). Speech melody as articulatorily implemented communicative functions. *Speech Communication, 46*, 220-251.

Xu, Y. (2009). Timing and coordination in tone and intonation—An articulatory-functional perspective. *Lingua, 119*, 906-927.

Xu, Y. (2005-2011). ProsodyPro.praat [Computer program]. Version 3.4, available from: http://www.phon.ucl.ac.uk/home/yi/ProsodyPro/.

Xu, Y. & Liu, F. (2007). Determining the temporal interval of segments with the help of F_0 contours. *Journal of Phonetics, 35*, 398-420.

Xu, Y. & Prom-on, S. (2010). Articulatory-functional modelling of speech prosody: A review. In *Proc. INTERSPEECH-2010*, 46-49.

Xu, Y. & Prom-on, S. (2010-2011). PENTAtrainer.praat [Computer program]. Version, 1.7, available from: http://www.phon.ucl.ac.uk/home/yi/PENTAtrainer/.

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, M. (2005). Tonal and durational variations as phonetic coding for syllable grouping. *Journal of the Acoustical Society of America*, *117*(Pt. 2), 2573.

Xu, Y. & Wang, M. (2009). Organizing syllables into groups—Evidence from F_0 and duration patterns in Mandarin. *Journal of Phonetics, 37*, 502-520.

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

Xu, Y. & Xu, C. X. (2005). Phonetic realization of focus in English declarative intonation. *Journal of Phonetics*, *33*, 159-197.

Yeou, M. (2004). Effects of focus, position and syllable structure on F0 alignment patterns in Arabic. In B. Bel & I. Marlien (Eds.), *Actes des XXVes Journées d'Etude sur la Parole*, *Arabic language processing* (pp. 369-374). Fez, Morocco.

Yip, M. (1988). Template morphology and the direction of association. *Natural Language & Linguistic Theory*, *6*(4), 551-577.

Zemlin, W. R. (1988). *Speech and Hearing Science: Anatomy and Physiology*. Englewood Cliffs, New Jersey: Prentice Hall.

Zhang, J. & Lai, Y. (2010). Testing the role of phonetic knowledge in Mandarin tone sandhi. *Phonology*, *27*, 153-201.