Article

Language *and* Speech

Mechanism of Disyllabic Tonal Reduction in Taiwan Mandarin

Language *and* Speech
2015, Vol. 58(3) 281–314

DOI: 10.1177/0023830914543286
las.sagepub.com
SAGE

Chierh Cheng and Yi Xu
Department of Speech, Hearing and Phonetic Sciences, University College London, UK

Abstract

This study was designed to test the hypothesis that time pressure is a direct cause of tonal reduction in Taiwan Mandarin. Tonal reduction refers to the phenomenon of the tones of a disyllabic unit being contracted into a monosyllabic unit. An experiment was carried out in which six native Taiwan Mandarin male speakers produced sentences containing disyllabic compound words /ma/+/ma/ with varying tonal combinations at different speech rates. Analyses indicated that increasing time pressure led to severe tonal reductions. Articulatory effort, measured by the slope of F0 peak velocity of unidirectional movement over F0 movement amplitude, is insufficient to compensate for duration-dependent undershoot (in particular, when time pressure exceeds certain thresholds). Mechanisms of tonal reduction were further examined by comparing F0 velocity profiles against the Edge-in model, a rule-based phonological model. Results showed that the residual tonal variants in contracted syllables are gradient rather than categorical—as duration is shortened, the movement towards the desired targets is gradually curtailed.

Keywords

Tonal reduction, disyllabic contraction, Taiwan Mandarin, articulatory effort, time pressure, F0 velocity profile, Edge-in model, target undershoot, hyper- and hypo-articulation

Introduction

Sounds in connected speech are pronounced differently from their canonical forms. For example, the English word *yesterday* can be pronounced differently, ranging from a canonical form [jɛstə˞ɹeɪ] to a severely reduced form [jɛʃeɪ] (Ernestus & Warner, 2011). Pronunciation variations manifest gradual changes in duration and degree of target realisation. Similar examples are common in many 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 to *wo zhao* [wo˨ ʈʂau˥˩]. Figure 1 illustrates the

Corresponding author:
Chierh Cheng, Department of Speech, Hearing and Phonetic Sciences, University College London, Chandler House, 2 Wakefield Street, London, WC1N 1PF, UK.
Email: chierh.cheng@gmail.com

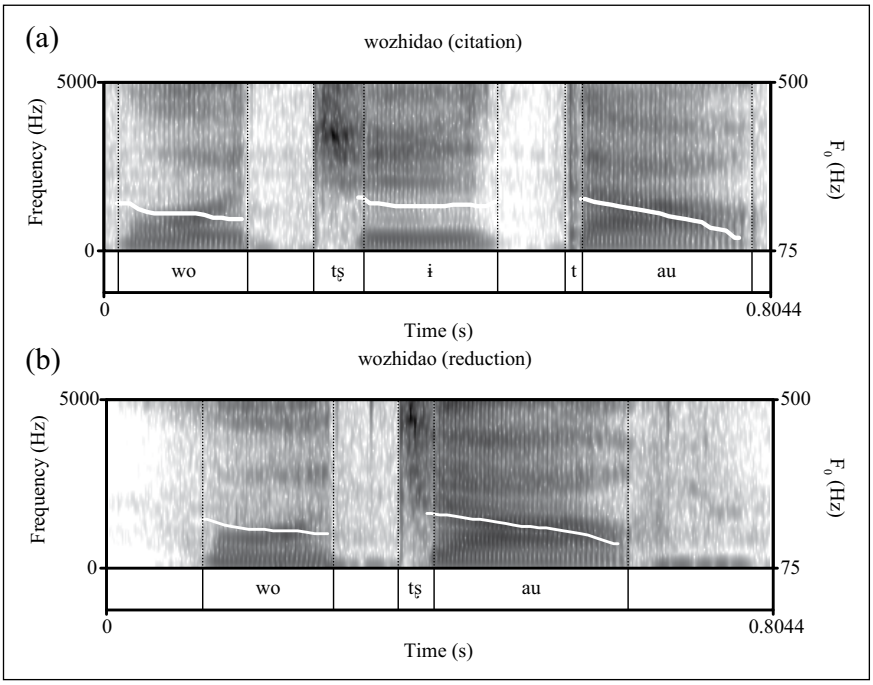


Figure 1. Spectrographic representation of [wo˩ tʂɨ˥tau˨˩] (a), 'I know' being reduced to [wo˩ tʂau˨˩] (b). Pitch values are shown as white contours overlaid on the spectrograms. Shorter duration and a reduced tonal range are also seen in the reduced token, [tʂau˨˩].

process, where the vowel /ɨ/ and the intervocalic consonant /t/ are absent from the spectrographic representation. The canonical tone shapes of H (˥, 55) in the syllable [tʂɨ] and F (˨˩, 51) in the syllable [tau] are realised as a slightly sloping contour. In more extreme cases, trisyllables can also be reduced to monosyllabic units, such as *wo bu zhi dao* [wo˩ pu˨˩ tʂɨ˥ tau˨˩], 'I don't know' becoming *wo bao* [wo˩pwau˨˩]. Several terms have been used to refer to such a severe form of phonetic reduction, including 'massive reduction' (Johnson, 2004), 'syllable fusion' (Wong, 2004), 'syllable merger' (Duanmu, 2000) and 'syllable contraction'[2] or 'contraction' for short (Tseng, 2005). Throughout this paper, the term 'contraction' is used because it has been most frequently used in research concerning severe forms of phonetic reduction in the Sinitic languages. 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 appeared to be absent from the spectrograms. A more technical definition of contraction is given in Section 2.

Investigating factors underlying the discrepancies between citational and reduced forms have been challenging in speech sciences. In a recent review of research on phonetic reduction, Warner (2011, p. 1881) suggests three types of sources jointly contributing to the degrees of reduction:

> 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.

Previous research has shown a correlation between factors such as lexical information or frequency and segmental reduction (Aylett & Turk, 2006; Pluymaekers, Ernestus, & Baayen, 2005) as well as suprasegmental reduction (Zhao & Jurafsky, 2007). More and more attention has also been given to sociophonetic variations and paralinguistic factors to account for idiosyncratic uses of reduction (Hawkins, 2010; Local, 2003). The present research, however, is to examine articulatory factors that contribute to tonal reduction, including time pressure, articulatory effort and articulatory constraint. Our focus is on investigating tonal reduction in Taiwan Mandarin with the goal to identify some of the basic mechanisms of phonetic reduction in general.

With regard to articulatory factors in phonetic realisation, Lindblom (1963) observed the interplay between duration and formant realisation in a CVC structure and proposed a *duration-dependent undershoot* model: when the 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 & Pols, 1990, 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 hypothesised that speakers can adapt to different speaking situations and choose appropriate production strategies (i.e., by changing kinematic parameters) to avoid or allow reduction. This is known as the *Hyper- and Hypo-articulation (H&H) theory*, which characterises the trade-off between articulatory economy and perceptual comprehension. Importantly, H&H theory hypothesises 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. Studies of reduction based on consistent communicative contexts often refer to H&H theory to account for temporal and spectral reduction found in high-frequency items (Aylett & Turk, 2006). It has also been suggested that information regarding language redundancy, either because of context or word frequency, can influence the amount of effort exerted in articulation (Pluymaekers 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, pp. 13–14) has suggested that unfamiliar or unknown lexical items may lead to a speaking style that is clearer than normal.

Several perceptual studies, however, have produced results that do not confirm the prediction that a clear speech style (which conceptually would be given 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) rates. However, the intelligibility advantage of clear speech was lost at fast speech rate; that is, clear speech no longer maintains 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 having speakers 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 the speech 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 reorganise 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 may indicate a physiological limit on the maximum speed of articulatory movement, as suggested by Lindblom's original model of phonetic undershoot (1963), and as empirically shown by Xu and Sun (2002) on the maximum speed of pitch change. It is also 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 Hypothesis and predictions

Studies regarding maximum speech 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, we hypothesised *a time-pressure account of tonal reduction*, according to which time pressure is a direct cause[3] of extreme reduction.

1.1.1 Prediction 1: Increasing time pressure leads to severe reductions. As noted previously, unfamiliar words may lead to a clearer speech style than familiar words due to a likely allocation of greater articulatory effort. If, however, time pressure is a direct cause of extreme reduction, when asked to increase articulation rate, speakers would still contract disyllabic compound words even if the words are unfamiliar.

1.1.2 Prediction 2:When contraction occurs, articulatory effort is not decreased. Consistent with Prediction 1, high articulatory effort is assumed to be exerted when producing unfamiliar words. This would mean that, if phonetic reduction did occur in unfamiliar words under high time pressure, it could not be due to reduced articulatory effort. We will verify Prediction 2 by measuring the articulatory effort of F0 movements under varying time pressures (see more details in Section 1.2.

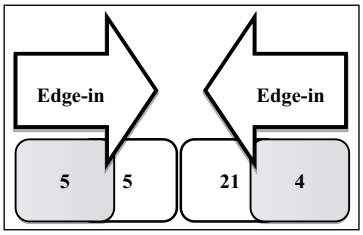


Figure 2. An Edge-in model for deriving the output tone 54 from two source syllables, [kən55] + [pən214] →[kəm54], meaning 'basically'. The bilabial plosive /p/ gives rise to a realisation of coda /m/.

1.1.3 Prediction 3: When contraction occurs, properties of the original tone can still be found. This final prediction is to further verify the time-pressure hypothesis by investigating how disyllabic tonal units are realised during contraction. As an additional observation, we also compare the contracted units against the prediction made by the Edge-in model (Yip, 1988; see more details on the model below). The added comparison is to see whether, during contraction, tonal targets of the corresponding non-contracted units are still present (assuming target approximations are curtailed by a limited time interval despite extra effort being applied, as per Predictions 1 and 2), or they are modified via a phonological process as predicted by the Edge-in model.

The Edge-in 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 the two targets on the outer edges intact. Figure 2 shows an example of the proposed Edge-in process.

1.2 The challenge of measuring articulatory effort

Currently, there is no accepted standard method of measuring articulatory effort. Malécot (1955, p. 36) described articulatory effort as 'a kinaesthetically felt degree of force of articulation'. This implies a psychological referent of articulatory effort that speakers can 'feel in their head'. However, this is not an objective measurement that directly corresponds to physiological reality (Parnell & Amerman, 1977; Tatham & Morton, 2006). Lindblom (1990) borrowed from Nelson (1983) the notion that 'peak velocity' (i.e., the highest absolute value in the continuous velocity profile of the movement) is an indicator of 'articulatory effort'. Nelson (1983) characterised skilled movements using basic dynamic principles, and proposed that the peak velocity of a unidirectional movement is equal to the impulse cost (time integral of the magnitude of the force per unit mass) when there is no friction, as expressed in the following equation:

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

where $u(t)$ is the applied force per unit mass (acceleration) and T is total movement time. Here I has the dimension of velocity, according to equation (B7) in Nelson (1983). In Equation (1), impulse cost is proportional to movement time (duration) given the unidirectionality of the movement. This means that the longer the movement time the greater the effort. Thus, movement time is not separated from articulatory force. This, however, is different from the notion of economy of effort

envisioned in the H&H theory, which gives the impression that effort is relatively independent of duration (see Kirchner, 1998, for a similar criticism).[4]

Empirically, it has been consistently found that peak velocity is quasi-linearly related to movement amplitude, whether the movement is measured from articulators directly (Kelso, Vatikiotis-Bateson, Saltzman, & Kay, 1985; Ostry, Keller, & Parush, 1983; Ostry & Munhall, 1985; Perkell, Zandipour, Matthies, & Lane, 2002) or from acoustics (Xu & Sun, 2002; Xu & Wang, 2009). This 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 of the different amplitudes may indicate greater muscle stiffness (Perkell et al., 2002), which would be a better indicator of articulatory effort than peak velocity alone (given the time confound indicated by the above equation). Therefore, if peak 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. Note also that because peak velocity is proportional to movement time as indicated by Equation (1), normalisation of peak velocity over movement amplitude also partially de-couples time (t) and force (u).

There might be questions as to whether it is valid to use acoustic measurements such as F0 to infer articulatory effort. The validity of F0 dynamics as an indicator of articulatory effort can be seen in the linear relations between F0 velocity and F0 movement amplitude as found in previous studies (Xu & Sun, 2002; Xu & Wang, 2009), which closely resemble the linear relations in articulatory movements (Hertrich & Ackermann, 1997; Kelso et al., 1985; Ostry & Munhall, 1985; Vatikiotis-Bateson & Kelso, 1993). Such quasi-linear relations were actually first reported for limb movements (Cooke, 1980), and its applicability to speech was justified on the basis of similar linear relations found in jaw and laryngeal movements (Ostry et al., 1983). Thus the similarity of F0 dynamics to both that of limbs and articulators warrants its use in assessing articulatory effort. Furthermore, it is worth noting that the H&H theory of economy of effort is developed based solely on examination of acoustic measurements, that is, formant frequencies (Lindblom, 1963, 1990; Moon & Lindblom, 1994).

In this study, therefore, the slope of regression of F0 peak velocity over F0 movement amplitude was used to test Prediction 2 regarding articulatory effort. As further justification for the method, we will show that a quasi-linear relationship between F0 peak movement velocity and movement amplitude is present.

2 Methodology

2.1 Testing materials

Disyllabic compound words /ma/+/ma/ with a total of 16 tone dyads (4 tones × 4 tones) embedded in two carrier sentences were constructed as testing materials. To observe continuous F0 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 sequence and for the purpose of getting the unidirectional movements required for proper measurement of peak velocity, two carrier sentences with an H or an L tone preceding the target sequence were composed (see Table 1). The tone following the target sequence is always H. The reason for not varying this following tone is because previous research has shown that contextual tonal variations are predominantly due to carryover rather than anticipatory effects (Gandour, Potisuk, & Dechongkit, 1994; Xu, 1997).

Table 1. Carrier sentences with a High/Low 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*/mai*L* _____ 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, don't you! Of course I won't eat/buy _____ salad that kinda stuff, because I dislike the sour sauce of _____ salad.

These 16 disyllabic compound words were assumed to be unfamiliar to the subjects because they were made of existing morphemes but with a likely low conditional probability (Pluymaekers et al., 2005), except for the T1T1 combination, which means 'mother' in Mandarin. Even so, these compound words were presented to the speakers as made-up names for novel kinds of salads, in order to make all items usable in the frame sentences.

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; Xu & Wang, 2009). This was carried out by devising a carrier sentence consisting of three phrases, each phrase having a slot for the same target sequence (see Table 1). The first phrase (i.e., the first position in a carrier sentence) contains 9 underlying syllables, the second 13 and the third 17, 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).[5]

2.2 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 (due to the trade-off between time and frequency resolution, cf. Fulop, 2011) than those of female speakers. Clear formant trajectories facilitate segmentation, which is critical for our evaluation of whether a token is contracted or not (see details on segmentation in the next section). The recordings were conducted in an anechoic chamber at University College London. Speech was recorded with a Shure SM10A microphone placed approximately 30 centimetres from the subject's mouth. The speech signals were recorded into 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 the 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.[6] 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 so that the examiner could not understand it without looking at the scripts, he was asked to repeat the entire trial (i.e., the carrier sentence displayed on the monitor from 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 4.5 (±0.15), 6.1 (±0.16) and 9.3 (±0.12) underlying syllables per second, respectively. Three repetitions of the

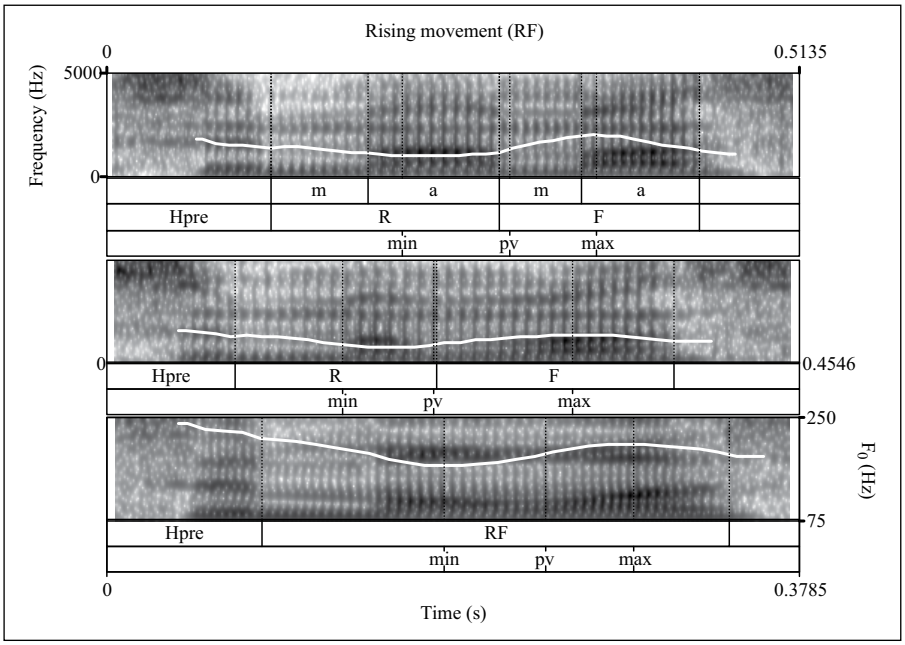


Figure 3. Rising movement and labelling examples in cases of H#RF. From top to bottom: non-contracted, semi-contracted and contracted. F0 values are shown as white contours overlaid on the spectrograms, scaling from 75 to 250 Hz. 'min' indicates the minimum pitch value and 'max' the maximum pitch value of the tested contour movement. 'pv' indicates the peak velocity within the domain between 'min' and 'max'.

entire block were recorded, each with a different randomisation order. In total, the number of tokens produced in this experiment was 16 (tone dyads) × 2 (preceding tones H or L) × 3 (positions in the carrier) × 6 (subjects) × 3 (speeds) × 3 (repetitions) = 5184. Out of these, 14 (0.2 %) were discarded from further analysis due to inadequate voice quality, such as creaky voice or speaker errors.

2.3 Segmentation, measurements and statistics

2.3.1 Segmentation. One problem with research in this area is that, as of yet, there is no standard method for segmenting reduced articulation for the purpose of defining contraction. Some researchers used human annotators to mark up occurrences of syllable contraction by employing operational criteria, such as omission of syllables and omission of syllable boundary (Tseng & Liu, 2002). Some used 'trough depth', adapted from Mermelstein's (1975) algorithm, assuming that the loudness differences of the adjacent source syllables could serve to indicate whether or not contraction occurs, and if necessary the trough depth values could be used as a continuous variable in data analysis. In practice, however, it has been reported that in certain cases the values of trough depth are greater than the reference value (i.e., 2 dB according to Mermelstein, 1975) to be labelled as two separate syllables and thus non-contracted, and the onset of the second syllables are actually elided and should be in fact regarded as partially contracted (Kuo, 2010, Figure 3). For a similar reason, Myers and Li (2009) also acknowledged that the adapted algorithm sometimes had to be tweaked by hand, with reference to a spectrogram, so that it did not measure irrelevant troughs due

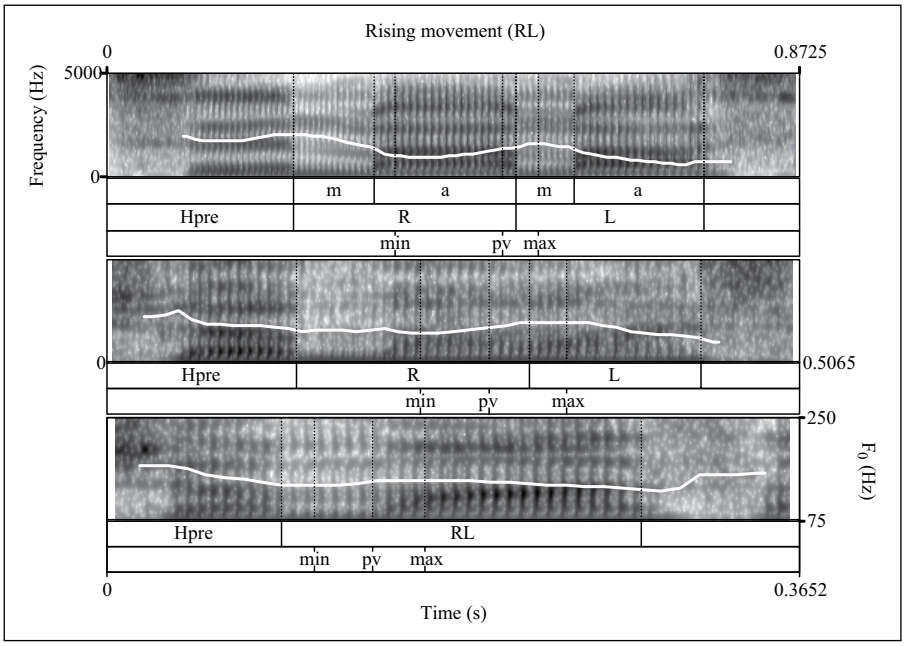


Figure 4. Rising movement and labelling examples in cases of H#RL. From top to bottom: non-contracted, semi-contracted and contracted. F0 values are shown as white contours overlaid on the spectrograms, scaling from 75 to 250 Hz. 'min' indicates the minimum pitch value and 'max' the maximum pitch value of the tested contour movement. 'pv' indicates the peak velocity within the domain between 'min' and 'max'.

to the peaks appearing on noise, fricatives or the release burst of stops at the edges of the testing materials.

As briefly mentioned in Section 1, in this research, a contracted syllable is defined as the absence of an intervocalic segment. Therefore we adapted a set of working definitions for labelling contraction. The labelling was guided by visually inspecting the spectrogram and waveform for segmental integrity, loudness and duration of the syllables, while listening to the speech for the auditory impression of juncture. *Non-contracted* tokens were labelled when a distinct N:V boundary was identified due to the presence of the nasal murmur. When a second nasal was absent (i.e., no nasal murmur, no obvious damped amplitude in the high-frequency spectrum or an undisrupted F1 or F2 trajectory from the previous /a/ vowel) occurred, no boundary was marked and the token was labelled as *contracted*. Since definitive classification is impossible, intermediate cases were labelled as *semi-contracted* with two intervals. Examples of this labelling are shown in Figures 3–6. Consistency of the labelling was double checked one month following the initial labelling. A small number 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.

2.3.2 Measurements. The extraction of the F0 contours was carried out using a modified version of ProsodyPro, a general purpose Praat script for large-scale F0 analysis (Xu, 2013). The script extracts F0 by displaying the vocal cycle markings generated by the Praat programme (Boersma & Weenink, 2010) and allowing users to perform manual rectifications by adding missing vocal

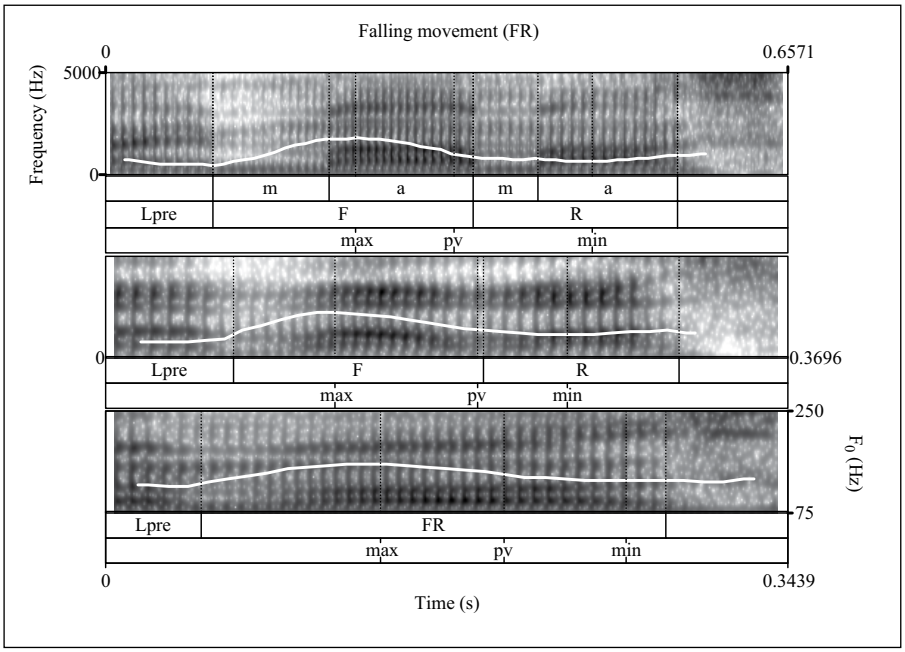


Figure 5. Falling movement and labelling examples in cases of L#*F*R. From top to bottom: non-contracted, semi-contracted and contracted. F0 values are shown as white contours overlaid on the spectrograms, scaling from 75 to 250 Hz. 'min' indicates the minimum pitch value and 'max' the maximum pitch value of the tested contour movement. 'pv' indicates the peak velocity within the domain between 'min' and 'max'.

pulse marks and removing excessive marks. The script then converts the vocal periods into F0 values and applies a trimming algorithm (Xu, 1999) to remove sudden jumps and dips. It generates various output data for further analysis. One output of the script is time-normalised F0 contours that were used only for graphic display purposes (see Figures 10–13, for example). In order to make direct comparisons of contracted and non-contracted tokens, time-normalisation was done based on the underlying rather than surface syllabification. Specifically, 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 (i.e., the whole word). Although after contraction, the original syllable boundary became uncertain, this was the best way to make sure the comparison was as direct as possible (i.e., via the alignment of the 40 measurement points between the contracted and non-contracted tokens).

To assess articulatory effort for Prediction 2, measurements were taken from two unidirectional F0 movements (rising and falling). This was done automatically with the modified ProsodyPro script. Because increased speech rate could lead to 'flattened' F0 contours, making the detection of F0 turning points impossible, measurements were taken from only tone dyads where a unidirectional rising or falling movement could be identified by the script. 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 unproblematically identified. In (H)#*R*F (see Figure 3), the trajectory of F0 needs to first move down from the preceding H ending tone (towards the

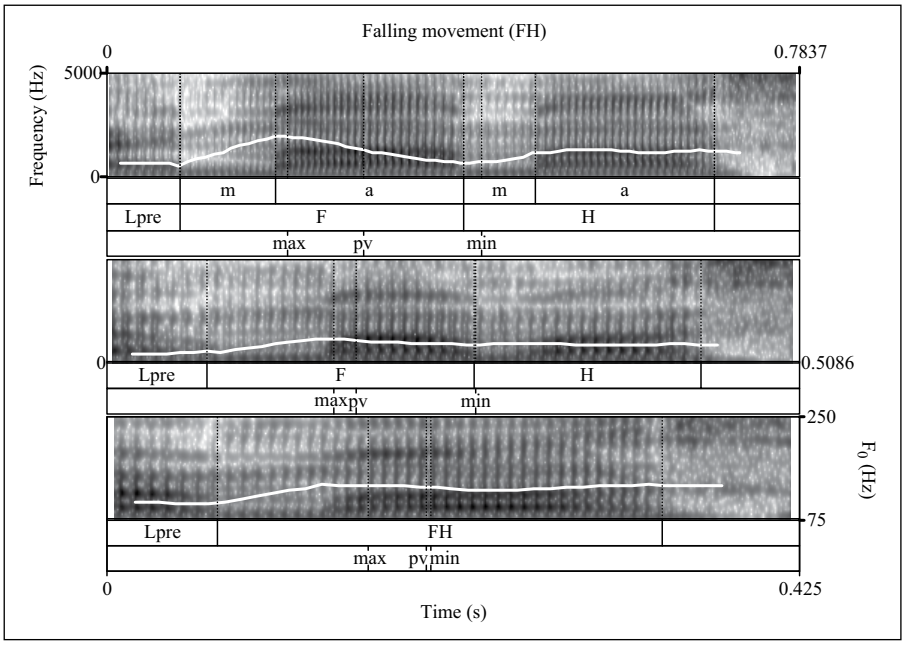


Figure 6. Falling movement and labelling examples in cases of L#*F*H. From top to bottom: non-contracted, semi-contracted and contracted. F0 values are shown as white contours overlaid on the spectrograms, scaling from 75 to 250 Hz. 'min' indicates the minimum pitch value and 'max' the maximum pitch value of the tested contour movement. 'pv' indicates the peak velocity within the domain between 'min' and 'max'.

minimum) in order to realise a rising movement for the R tone in the first syllable. The presence of a maximum was further guaranteed by the required falling movement of the F tone in the second syllable. The modified ProsodyPro script first located the F0 minimum ('min') in the early part of /mama/. It then searched for the F0 maximum ('max') between the 'min' point and the end of /mama/. In between 'min' and 'max', the location and value of the peak velocity ('pv') was obtained from the continuous F0 velocity trajectory (generated by ProsodyPro) corresponding to the unidirectional F0 rise. Figures 3 and 4 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 5 and 6). In total, 383 out of 647 tokens from the four selected tone sets were found to be usable for assessing articulatory effort (an inclusion rate of 59.2%). This low inclusion rate is an indication that the rate of articulation of the speakers had indeed been pushed to the limit, as will be seen more clearly in Section 3.2.[7]

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 and a minimum in the second F in an (L)#FF sequence occurred even less frequently than the (H)#*R*F and (H)#*R*L cases. This is owing to the high articulatory demand resulting from consecutive dynamic tones, that is, RR or FF, in an incompatible tonal environment even at a normal speech rate (Kuo, Xu, & Yip, 2007; Xu & Wang, 2009).

Three kinematic measurements for each unidirectional movement were taken to assess articulatory effort: (1) F0 movement duration, that is, time difference between adjacent maximum

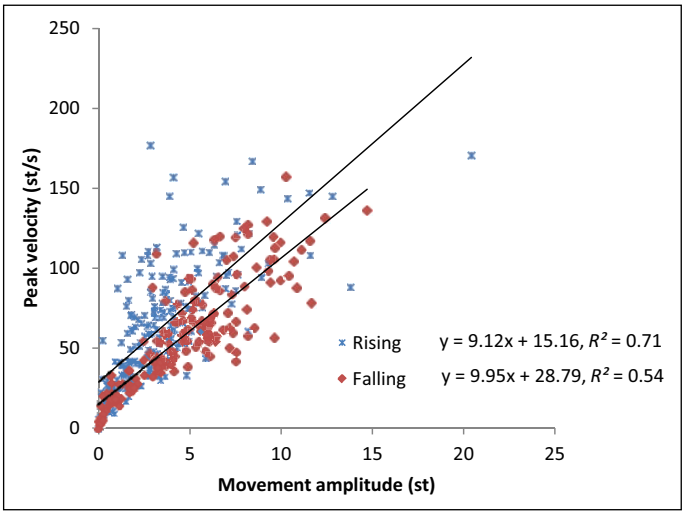


Figure 7. Linear regressions of F0 peak velocity (*y*-axis in semitones/second) over F0 movement amplitude (*x*-axis in semitones) for both rising and falling movements (in absolute values of peak velocity). In total, 216 data points were valid for a *rising* movement of the R tone in (H)#*R*F and (H)#*R*L and 167 data points for a *falling* movement of the F tone in (L)#*F*R and (L)#*F*H.

F0 and minimum F0 in seconds; (2) F0 movement amplitude, that is, F0 difference between adjacent maximum F0 and minimum F0 in semitones; and (3) F0 peak velocity, that is, maximum absolute value in the first derivative of a continuous unidirectional F0 movement, in semitones/second. Figure 7 displays an overall scatter plot of F0 peak velocity as a function of F0 movement amplitude for all selected rising and falling movements. The relationship between F0 peak velocity and F0 movement amplitude was highly linear ($r = 0.74$, $p < 0.001$), which is consistent with the movement analyses of previous studies (acoustic movements: Xu & Sun, 2002; Xu & Wang, 2009; articulatory movements: Hertich & Ackermann, 1997; Kelso et al., 1985; Ostry & Munhall, 1985; Vatikiotis-Bateson & Kelso, 1993). Given the highly linear relationship between F0 peak velocity and F0 movement amplitude, this experiment also used the slope (i.e., gradient) of their regression line to assess articulatory effort as did the other studies just mentioned.

2.3.3 Statistics. Two main statistical analyses were conducted to verify Predictions 1 and 2, respectively. Firstly, to see whether increasing time pressure led to severe reductions (Prediction 1), a multinomial logistic regression was performed with CONTRACTION TYPE (non-contracted, semi-contracted and contracted) as the ordinal dependent variable and SPEED, POSITION in the carrier sentence, and REPETITION as predictor variables.

Secondly, to compare articulatory efforts among the three contraction types (Prediction 2: when contraction occurs, articulatory effort is not decreased), three separate one-way repeated measures analyses of variance (ANOVAs) were performed with CONTRACTION TYPE as the independent variable, and DURATION, F0 excursion SIZE and SLOPE of the regression line of F0 peak velocity over F0 movement amplitude as dependent variables, respectively. In addition, the formula in calculating the maximum speed of pitch change (Xu & Sun, 2002, detailed below) were also applied to the data used for Prediction 2, followed by *t*-tests to see whether speakers have reached their physiological limits in the case of extreme reduction.

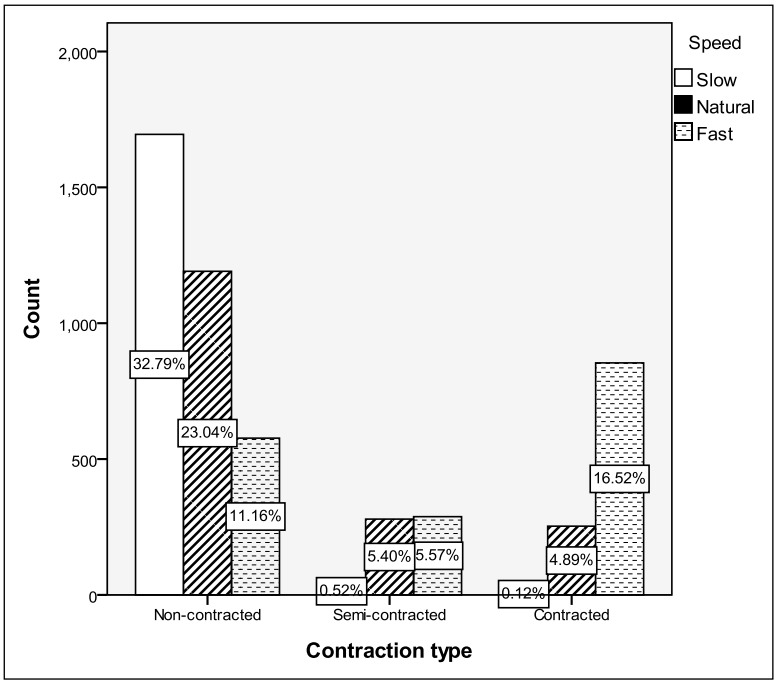


Figure 8. Contingency of contraction type at different speeds. The *x*-axis shows three different contraction types and the *y*-axis shows the frequency count. The relative frequency shown in percentage was calculated by dividing the number of tokens in a specific group by the total number of tokens (i.e., 5170).

3 Results

Three predictions were tested (and the results presented accordingly) to see whether tonal reduction can be explained by time pressure.

3.1 Prediction 1: Increasing time pressure leads to severe reductions

The multinomial logistic regression confirmed the first prediction. SPEED was found to be positively related to CONTRACTION TYPE (*Coef.* = 2.50, *S.E.* = 0.39, $p < 0.0001$). That is, for a unit increase in speed, the expected ordered log odds increased by 2.50 as one moved to the adjacent higher category of contraction type (i.e., from non- to semi- to contracted). No significance was found for POSITION (*Coef.* = 0.54, *S.E.* = 0.45, $p = 0.23$) or REPETITION (*Coef.* = 0.56, *S.E.* = 0.44, $p = 0.21$). No interactions among predictors were found (POSITION × SPEED: *Coef.* = −0.22, *S.E.* = 0.18, $p = 0.21$; POSITION × REPETITION: *Coef.* = −0.10, *S.E.* = 0.20, $p = 0.60$; SPEED × REPETITION: *Coef.* = −0.13, *S.E.* = 0.17, $p = 0.46$; POSITION × SPEED × REPETITION: *Coef.* = 0.04, *S.E.* = 0.08, $p = 0.61$).[8]

Figure 8 displays the relationship between contraction type and speed with the relative frequency shown in percentage. The *x*-axis shows three different contraction types and the *y*-axis shows the count of speech rate by contraction type. As shown in Figure 8, non-contracted units decreased as speaking rate increased.[9] Conversely, both semi-contracted and contracted units

Table 2. Three separate one-way repeated measures analyses of variance using CONTRACTION TYPE as the independent variable with respective dependent variables: DURATION (in seconds), F0 excursion SIZE (in semitones) and SLOPE (as an indicator of articulatory effort) of the regression line of F0 peak velocity over F0 movement amplitude of the three contraction types.

	DURATION	SIZE	SLOPE
Non-	0.175	4.67	13.36
Semi-	0.141	2.65	17.16
Contracted	0.054	0.86	22.99
F value	$F(2,8) = 68.86$	$F(2,8) = 16.01$	$F(2,8) = 9.49$
p value	$p < 0.0001$	$p = 0.002$	$p = 0.008$

increased with speaking rate. Within 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, indicating that under high time pressure, even unfamiliar words can be contracted. In general, therefore, Prediction 1 is supported.

3.2 Prediction 2: When contraction occurs, articulatory effort is not decreased

3.2.1 Articulatory effort. Table 2 displays results of three separate one-way repeated measures ANOVAs using CONTRACTION TYPE as the independent variable on respective dependent variables: (1) DURATION; (2) F0 excursion SIZE; and (3) SLOPE of the regression line of F0 peak velocity over F0 movement amplitude (as an indicator of articulatory effort). As Table 2 shows, SLOPE is not decreased and, in fact, it increased with the severity of phonetic reduction.

The ANOVAs showed large effects of CONTRACTION TYPE on all three dependent variables. A post-hoc (Tukey) analysis of DURATION revealed that all CONTRACTION TYPE were significantly different from each other with NC (non-contracted) being the longest and C (contracted) the shortest (DURATION: [NC > C], $p < 0.001$***; [Semi > C], $p < 0.001$***; [NC > Semi], $p < 0.001$***). Post-hoc (Tukey) analyses of SIZE also showed significant differences in all comparison, with NC being the largest and C the smallest (SIZE: [NC > C], $p < 0.001$***; [Semi > C], $p = 0.028$*; [NC > Semi], $p = 0.002$**). Post-hoc (Tukey) analyses of SLOPE found significant differences in (NC versus C) and in (Semi versus C), but not in (NC versus Semi) (SLOPE: [NC < C], $p < 0.001$***; [Semi < Cntr], $p = 0.024$*; [NC < Semi], $p = 0.121$).[10]

In summary, the degree of reduction, as reflected by contraction type, was negatively related to duration and F0 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. Similar results were also found in our investigation on extreme segmental reduction (Cheng & Xu, 2013).

3.2.2 Maximum speed of pitch change. In view of the insufficient compensation from an increased articulatory effort for items of limited duration (Table 2), 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 be helpful to assess whether speakers did indeed approach their maximum speed of pitch change. According to Xu and Sun

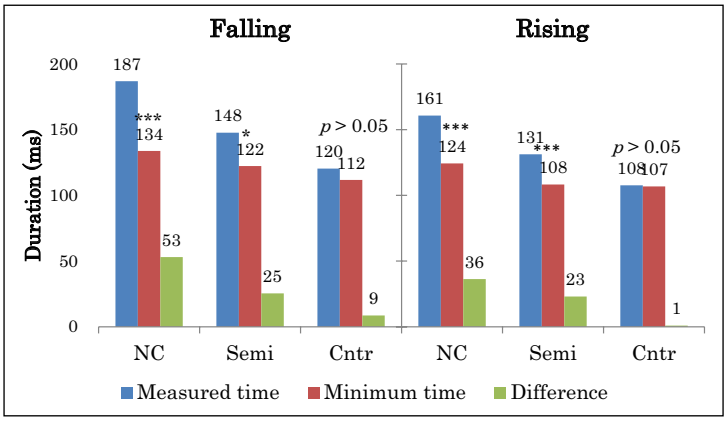


Figure 9. Measured time (blue) at different contraction types and movement directions compared to the minimum time (red) required for the same amount of F0 movement amplitude computed by Equations (2) and (3). The green bars indicate the differences between these two time intervals. The asterisks and *p* values indicate the statistical significance as described in the text ($p < 0.05$*; $p < 0.01$**; $p < 0.001$***). (Colour online only.)

(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 F0 movement, which can be approximated by the following two formulae:

$$T = 100.4 + 5.8d(\text{pitch lowering}), \tag{2}$$

$$T = 89.6 + 8.7d(\text{pitch raising}), \tag{3}$$

where T is the minimum movement time in milliseconds and d is the F0 movement amplitude in semitones. Equations (2) and (3) were applied to the current data. Figure 9 shows the observed duration and the theoretical minimum time needed to generate the same F0 movement amplitude together with their time differences.

As can be seen in Figure 9, the measured time of non-contracted and semi-contracted units were both significantly longer than the corresponding minimum duration according to Welch's two-sample *t*-test (Welch's *t*-test is an adaptation of Student's *t*-test intended for use with two samples having possibly unequal variances), indicating that the time interval was ample and that speakers were not required to reach their maximum speed (Falling NC: $t = 8.64$, $df = 136.6$, $p < 0.001$; Falling Semi: $t = 2.67$, $df = 23.1$, $p < 0.05$; Rising NC: $t = 8.39$, $df = 232.6$, $p < 0.001$; Rising Semi: $t = 3.93$, $df = 47.4$, $p < 0.05$). 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: $t = 0.92$, $df = 17.2$, $p > 0.05$; Rising Cntr: $t = 0.09$, $df = 24.3$, $p > 0.05$) and extreme reduction was therefore inevitable.

3.3 Prediction 3: When contraction occurs, properties of the original tone can still be found

Now, by putting into perspective high time pressure, extra articulatory effort and physiological limits observed in contracted units, we ask whether properties of the original tone can still be found. The answer is proved to be positive as will be shown in the descriptions of tonal contours below, followed by the comparative results of the Edge-in model.

3.3.1 Tonal contours of different contraction types. To see the tone shapes in a straightforward manner, mean F0 contours of the 16 tone dyads in different contraction regimes are first displayed in Figures 10–13. Since no effects of sentence position, repetition or their interactions were found (as shown in Section 3.1), F0 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 12(c)) was always realised as RL due to an obligatory tone sandhi rule in Mandarin Chinese that modifies the first L in an LL sequence to R (Chao, 1968).

Three direct observations can be made from the mean F0 contours of Figures 10–13. Firstly, as expected, non-contracted contours have larger pitch ranges than semi-contracted and contracted contours (e.g., RF in Figure 11(d), among others). Secondly, there is a robust carryover effect in all 16 tone dyads. This can be seen by comparing the onset F0 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 11). Thirdly, contracted contours display a higher overall F0 in comparison to non-contracted and semi-contracted contours (e.g., contracted (H)#HR and (L)#HR in Figure 10(b), contracted (H)#FL and (L)#FL contours in Figure 13(c)). The only three exceptions were (H)#HF (Figure 10(d)), (H)#RR (Figure 11(b)) and (H)#LF (Figure 12(d)), where the semi-contracted F0 contours were higher than the contracted contours.

In addition to the higher overall F0, which may be due to an increase of overall effort under high time pressure, 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, though also a dynamic tone, F was not as susceptible as R to reduction. That is, most falling movements in F were still present across different contraction types and tonal environments. Taking (H)#H*R* in Figure 10(b) as an example, the final rise in the R tone is missing from the contracted contours. Similarly, in Figure 11(c) 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 realised due to the time constraint. This issue will be examined in the next section.

3.3.2 Incompatibility with 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 F0 velocity profiles. Such profiles can give a good indication of articulatory movements towards the underlying tonal targets (Gauthier, Shi, & Xu, 2007). For simplicity, the analysis presented here was carried out on the tone dyad (H)#H*R* (see Figure 14) and (H)#F*F* (see Figure 15). Other tone dyad combinations were also analysed and the results were in line with those presented here.[11]

Figure 14 shows the aforementioned F0 velocity contours along with a contracted (H)#HH F0 velocity contour. As mentioned previously (see Figure 10(b)), during the execution of a canonical HR in the (H)#HR context, the F0 velocity remains positive and only undergoes comparatively

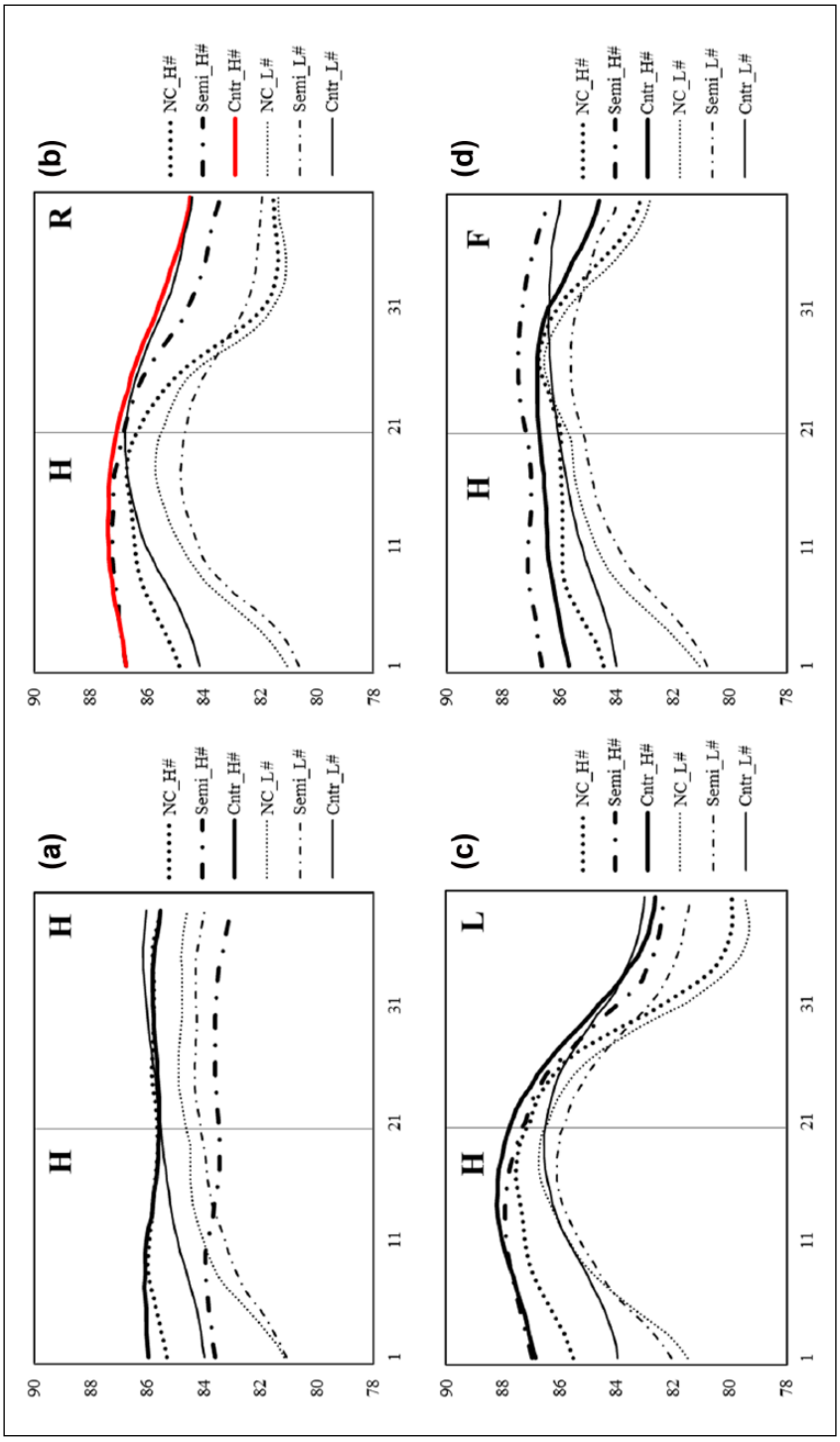


Figure 10. F0 contours of tone dyads HH (a), HR (b), HL (c) and HF (d). Tones preceding the tone dyads are indicated by line thickness and contraction types by line style, as shown in the legend. The *x*-axis is 40 evenly spaced measurement points and the *y*-axis is in semitones.

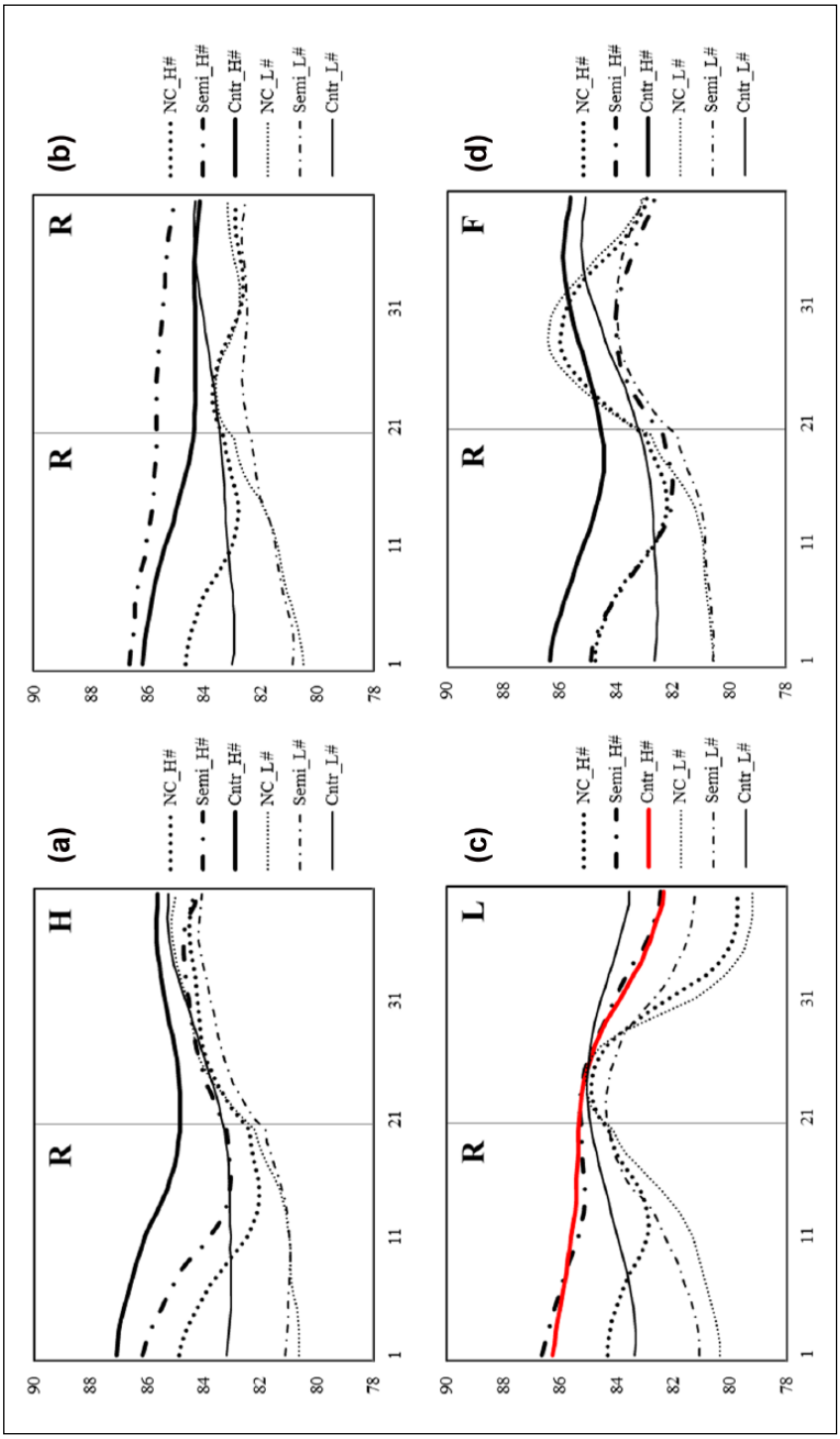


Figure 11. F0 contours of tone dyads RH (a), RR (b), RL (c) and RF (d). Tones preceding the tone dyads are indicated by line thickness and contraction types by line style, as shown in the legend. The *x*-axis is 40 evenly spaced measurement points and the *y*-axis is in semitones.

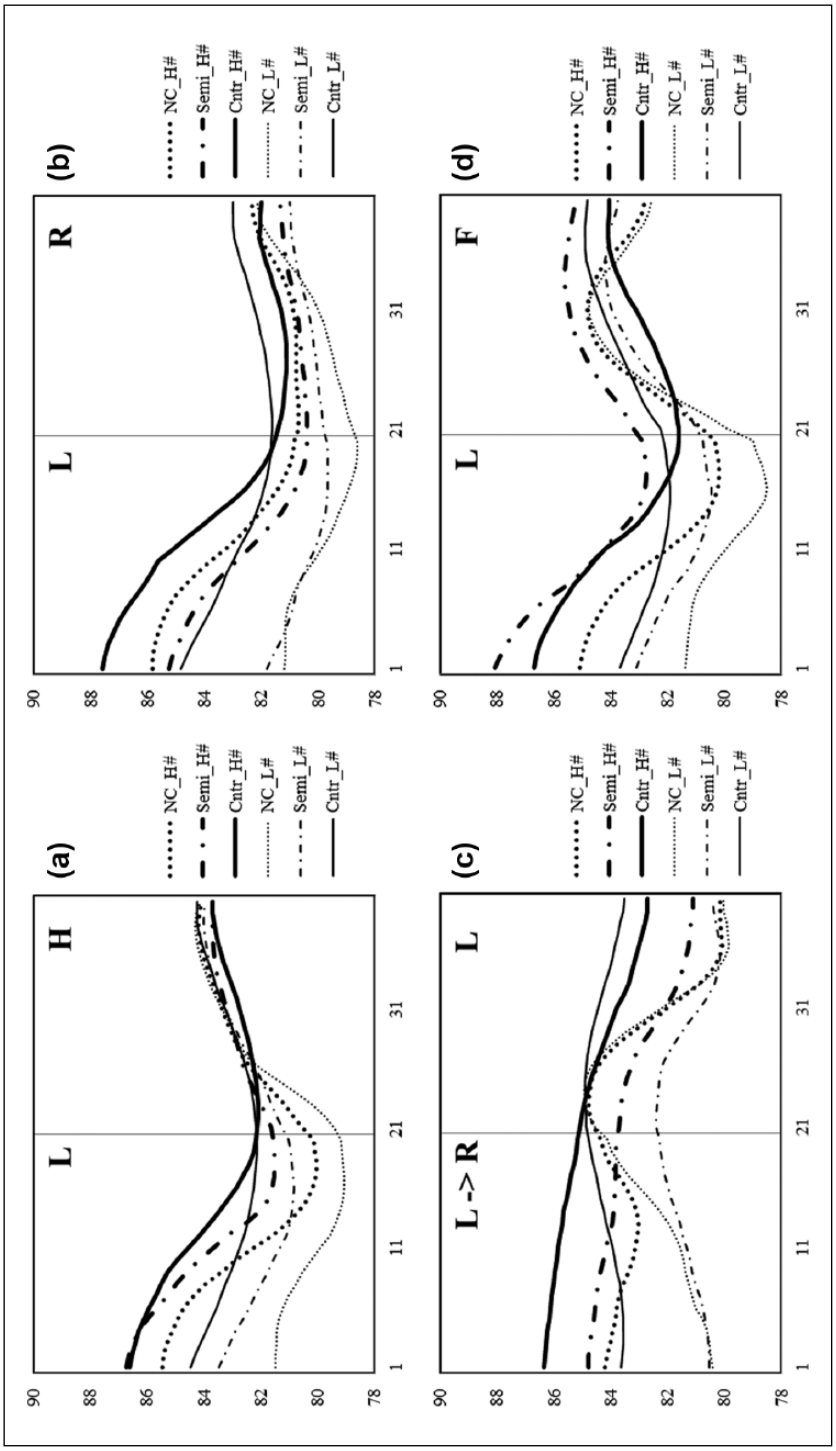


Figure 12. F0 contours of tone dyads LH (a), LR (b), LL -> RL (c) and LF (d). Tones preceding the tone dyads are indicated by line thickness and contraction types by line style, as shown in the legend. The *x*-axis is 40 evenly spaced measurement points and the *y*-axis is in semitones.

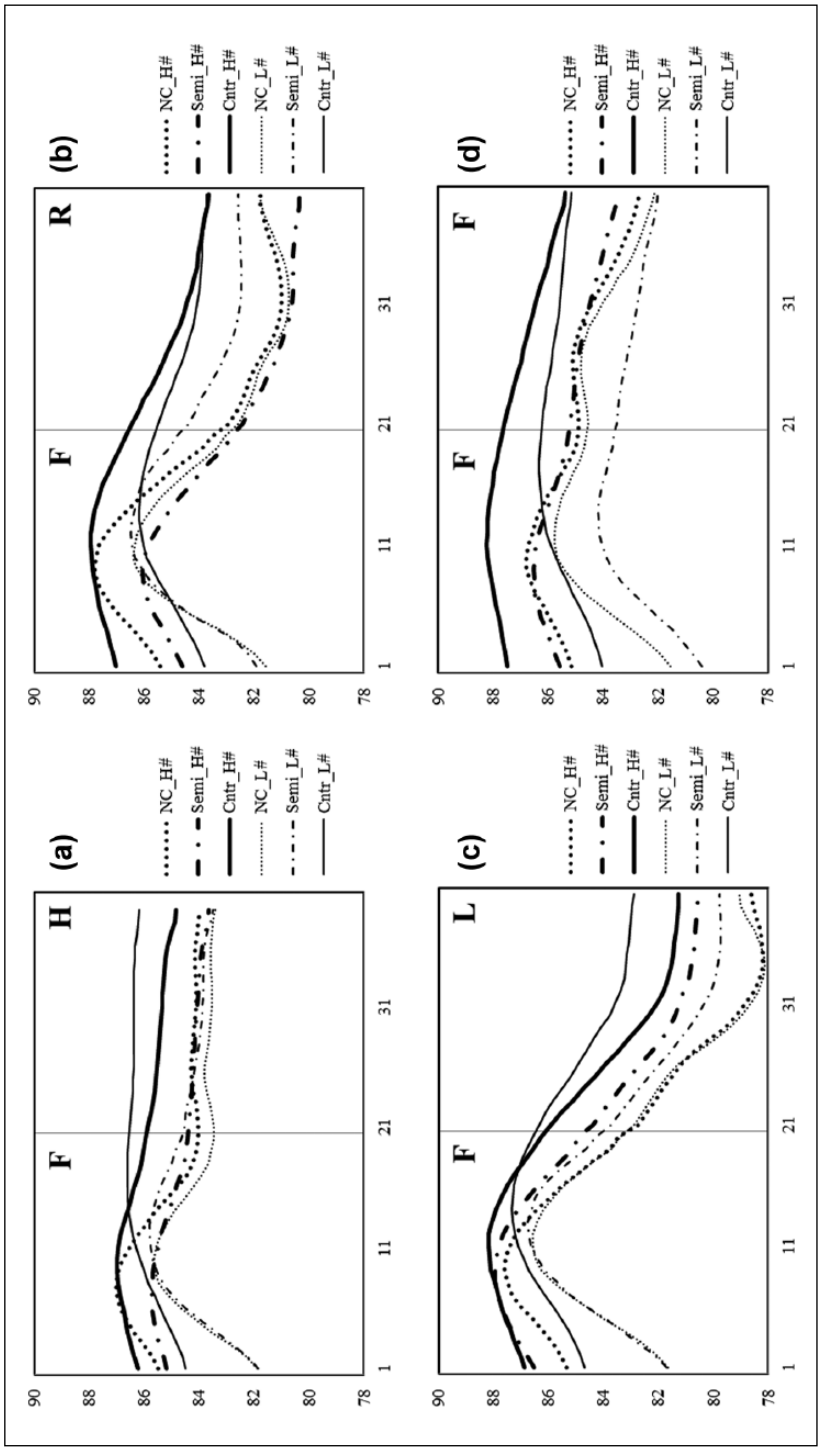


Figure 13. F0 contours of tone dyads FH (a), FR (b), FL (c) and FF (d). Tones preceding the tone dyads are indicated by line thickness and contraction types by line style, as shown in the legend. The *x*-axis is 40 evenly spaced measurement points and the *y*-axis is in semitones.

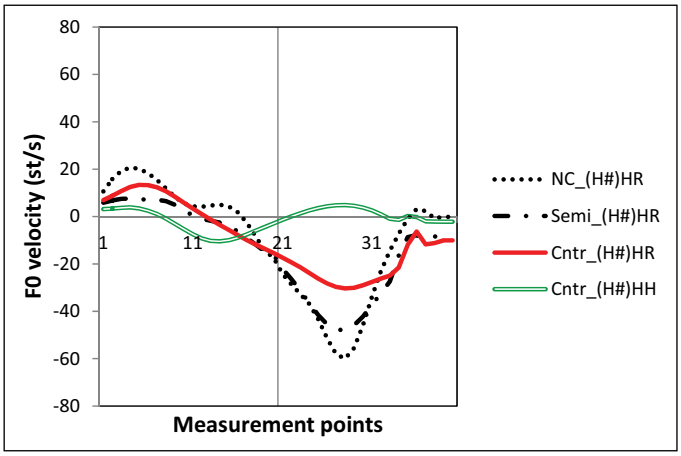


Figure 14. Mean F0 velocity contours of (H#)HR of three contraction types and that of contracted (H#)HH (cf. Figures 10(a) and (b)). The vertical line marks the 20th point of measurements.

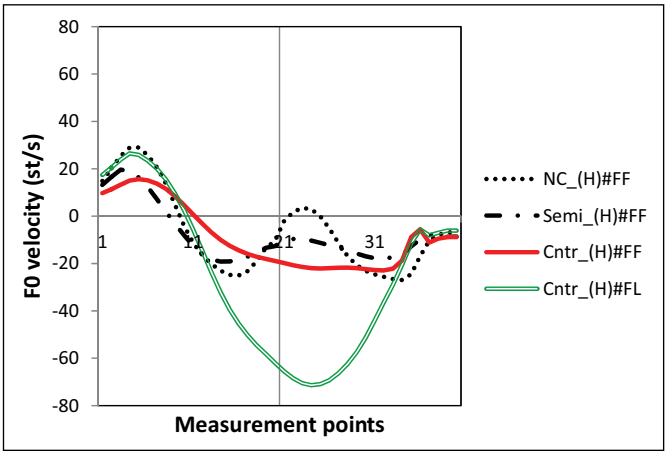


Figure 15. Mean F0 velocity contours of (H#)FF of three contraction types and that of contracted (H#)FL (cf. Figures 13(c) and (d)). The vertical line marks the 20th point of measurements. (Colour online only.)

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 F0 velocity contours shown in Figure 14 both exhibit this general behaviour but with slightly less variation in the first syllable and a shallower trough in the second syllable. Importantly, the F0 velocity approaches the zero-velocity line towards the end of the second syllable in all cases. This trend is an indicator of the presence of the rising target of the underlying R tone.

It is also informative to compare a contracted (H)#HR to a contracted (H)#HH tone dyad. For both of these dyads, the Edge-in model predicts similar surface forms when contraction occurs. That is, when an HR sequence is reduced to a single syllable, the Edge-in model predicts a resulting change of ~~5535~~→55, rendering it the same as a contracted HH sequence: ~~5555~~→55. However, Figure 14 shows that a clear difference is present between the contracted (H)#HR and contracted

(H)#HH. In contrast to the contracted (H)#HR F0 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. Thus the underlying targets of the contracted tones can still be found even with a much reduced duration, and no complete target deletion is evident, as suggested by the Edge-in model.

Figure 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 are realised as similar falling forms, that is, 5~~151~~→51 for (H)#FF, and 5~~121~~→51 for (H)#FL, since the tonal components near the dyad internal boundary are fully deleted. In Figure 15, three varying F0 velocity contours of (H)#FF along with a contracted (H)#FL F0 velocity contour are shown. In the initial period of the first interval, all F0 contours further increase their velocity from the previous high-ending H tone and 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 F0 velocity of the contracted (H)#FF remains negative following the first F. The velocity hovers around −20 semitones/second (i.e., –2 semitones per 100-ms syllable). However, the velocity of the contracted (H)#FL is decreased to a minimum of over −70 semitones/second. The stagnant velocity in the contracted (H)#FF may be explained by the fact there is no time for many 'meaningful' oscillations to occur. However, it also suggests that the target of the second syllable did not change into that of an L tone, as it contrasts sharply with the true (H)#FL cases (green line). This is contrary to the prediction of the Edge-in model that both contracted (H)#FF and (H)#FL should exhibit similar 'falling' F0 trajectories.

4 Discussion and conclusion

In this research we tested whether tonal reduction in contracted syllables can be accounted for by time pressure. The three predictions were all confirmed, providing support for the general hypothesis that *time pressure is a direct cause of extreme reduction*. Firstly, *increasing time pressure leads to severe reductions*. Ordinal logistic regression analysis (Section 3.1) suggested that speech rate has a significant effect on the type of contraction that occurs (also see Figure 8). Further analyses also indicated a close relation between duration and F0 excursion size (Table 2).

Secondly, the slope of the regression line of F0 peak velocity over F0 movement amplitude supports the prediction that *when contraction occurs articulatory effort is not decreased*. The tonal data labelled as semi-contracted or contracted exhibited decreased duration and excursion size but not decreased articulatory effort (Table 2). Furthermore, compared to the maximum rate of pitch change established previously (Xu & Sun, 2002), speakers appeared to have reached their physiological limit, particularly in cases where duration was comparable to that of a contracted syllable (Figure 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 *when contraction occurs, properties of the original tone can still be found* was supported by examining F0 velocity profiles across different contraction types. Taking for example the tone dyad (H)#H*R*, evidence of the properties of targeted R tone under varying time pressures is shown in Figure 14. Unlike what is suggested by the Edge-in model, the absence of a final rise in contracted H*R* (Figure 10(b)) 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 F0 counter as is also seen in contracted F*F* (Figure 15). This analysis of F0 velocity profiles not only verifies the time-pressure account of extreme reduction, but also brings out the continuous nature of the effect of duration on target realisation.

4.1 The properties of tones in connected speech

The results presented in this study display two typical properties of tones in connected speech: (a) reduced pitch range as exhibited by a small F0 excursion size and (b) simplified F0 contours that 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 F0 contour (Kuo et al., 2007; Xu & Wang, 2005). It has also been argued above that even though the original tonal elements can be detected in the velocity profile, there is not enough time for the effort to result in large F0 movements.

Tseng (2005) 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 (see Table 4 in Appendix C), but it is shown in Figures 10–13 that an F tone is generally less susceptible to the loss of its dynamic features (i.e., a falling movement) than 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 by Tseng (2005). Furthermore, as a side note, this high dependency of target realisation on duration seems to further agree with the finding of Xu (1998) and Xu and Wang (2001) that R and F tones in Mandarin Chinese are more likely to have dynamic rather than static targets (as in the level tones).

Therefore, the residual tonal variants in contracted syllables are unlikely to be generated by rule (i.e., retaining only the edge portions of the tonal components), but are rather due to simple articulatory mechanisms—as duration is shortened, the movement towards the desired targets is gradually curtailed. Moreover, observations of F0 contours and their velocity profiles allowed us to see further evidence of articulatory movements towards the underlying targets of the four tones, as has previously been demonstrated (Gauthier et al., 2007).

There were nevertheless some limitations to the present study. Firstly, for the analysis of articulatory effort, only part of the data could be included because of the necessary evil associated with the very nature of extreme reduction. A similar scenario was also reported by Cheng and Xu (2013) regarding the severe destruction of formant trajectories in extreme segmental reduction. Secondly, there is a lack of analysis of individual variations (Fougeron & Jun, 1998), although such variability is assumed in the statistical analysis, that is, speakers differ in their strategy in producing speech at a fast rate. Nevertheless, to the best of our knowledge, no prior research has systematically examined variations in F0 patterns in relation to possible articulatory mechanisms underlying extreme reduction in any language. As mentioned in Section 1, although investigating a problem with such a large parameter space is difficult and poses many challenges, a great deal of research originating from differing perspectives has pointed towards the importance of duration in phonetic reduction.

4.2 Summary

In this study, we designed an experiment to evaluate the nature of tonal reduction. It involved the examination of tonal sequences produced under varying time pressure and in systematically varied tonal environments. We have demonstrated that tonal reduction is largely dependent on duration and that properties of the original tones can still be found when contraction occurs. 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. It is hoped that the empirical data on tonal reduction in Taiwan Mandarin presented here will help elucidate the relative importance of articulatory factors underlying phonetic reduction.

Acknowledgement

Part of this research was presented in Cheng, Xu, & Gubian (2010) 'Exploring the mechanism of tonal contraction in Taiwan Mandarin', *Proceedings of INTERSPEECH 2010*, pp. 2010–2013. Makuhari, Japan, 26–30 September 2010.

Funding

The first author was supported in part by UCL Graduate School Research Projects Fund and Overseas Elite Scholarship from the Ministry of Education, Taiwan.

Notes

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.
2. The term 'syllable contraction', however, has been used to refer to two different phenomena. One is the extreme phonetic reduction (Tseng, 2005) that this research is concerned with. The other is the morphophonological process involving combinatory phonetic modifications of adjacent morphemes, for example, English contracted forms I'm or don't, which might be arguably fossilised cases of phonetic reduction (Suihkonen, 2005; Vance, 2008, p. 48).
3. Here, 'direct cause' implies two things. Firstly, 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 the speaker to realise a target fully despite the extra effort that might have been applied. Secondly, the commonly recognised 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.
4. In fact Lindblom (1990) mentioned that peak velocity is a reasonable measurement of effort, and that it is compatible with the duration-dependent undershoot model. However, Equation (1) above clearly shows that peak velocity is not independent of duration.
5. Research has shown that as the number of syllables in a syllable group increases, the durations of all individual duration shorten. However, as shown in Section 3, we found that only speed had a statistically significant effect on reduction. Position in a carrier sentence had no effect on reduction and therefore failed at manipulating time pressure. We thus do not discuss it further.
6. Thanks to one reviewer for pointing out that elicited reduction might be confounded by factors such as speech rate and speaking style, large variations of which should indeed be avoided if possible. The

'slow' and 'natural' scenarios requested of speakers were simply to help maintain (to some degree) a consistent speaking mode across various speakers (whose speech rates for each mode will of course vary). The focus of our experiment, however, is to examine the role of time pressure on tonal reduction, that is, the relation between local duration and pitch movement, rather than the influence of a specific factor such as speaking style on phonetic reduction. Nevertheless, to help disentangle a potential confounding effect between speech rate and speech style on reduction, two follow-up statistics suggested by the reviewer were also conducted and reported in Notes 8 and 10. Results remained similar to our original approach.

7. Note that the low inclusion rate for this particular analysis is due to the very nature of extreme reduction as well as that of this analysis. That is, unless clear signs of articulatory failure are observed, we cannot be sure that reduction to the level of contraction has actually occurred. On the other hand, tokens that have been fully reduced could no longer be analysed for their originally intended movements in terms of peak velocity over movement amplitude. So the eligible tokens are only those that have not yet been fully reduced, and their number cannot be very large. A further breakup of tokens used in the three predictions is shown in Appendix D, Figure 20.
8. An anonymous reviewer suggests a follow-up logistic regression analysis with CONTRACTION TYPE as the ordinal dependent variable and DURATION as the continuous predictor. Results remained similar to our original statistical analysis, showing a negative relation between contraction type (from non- to semi- to contracted) and duration (*Coef.* = −86.1, *S.E.* = 31.4, $p < 0.01$).
9. Note that Figure 8 also shows two exceptional patterns: speaker contracted at slow/natural rate (0.12% + 4.89%) and did not contract at fast rate (11.16%). Such an unexpected finding may be due to a local time-pressure effect. That is, the durations of these tokens are generally shorter than their non-contracted counterparts at the same speech rate. A breakdown of duration by speed and contraction type indicates that those contracted tokens are indeed shorter than the non-contracted ones at the same speech rate, and vice versa. The mean duration of non-contracted tokens is 189 ms at slow rate, 145 ms at natural rate and 133 ms and fast rate; the mean duration of semi-contracted tokens is 157 ms at slow rate, 142 ms at natural rate and 119 ms at fast rate; the mean duration of the contracted tokens is 92 ms at slow rate, 67 ms at natural rate and 46 ms at fast rate (cf. Table 2).
10. An anonymous reviewer suggests using SPEED instead as an independent variable. Results remained similar to the original ANOVAs analyses, showing a large effect of SPEED on all three dependent variables, that is, DURATION, SIZE and SLOPE (more details are shown in Appendix B, Table 3).
11. The plots of F0 velocity profiles of all tone dyads and contraction types are provided in Figures 16–19 in Appendix A. The F0 velocity profiles were calculated before the time normalisation so the original velocity values were preserved. 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 that the following syllable in the carrier sentence begins with a voiceless consonant (i.e., /tʂʰ/) that interrupts the continuous F0 and affects the smoothness of the velocity profiles.

References

Adank, P., & Janse, E. (2009). Perceptual learning of time-compressed and natural fast speech. *Journal of the Acoustical Society of America*, *126*, 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*, 3048–3058.

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

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

Chao, Y. R. (1968). *A grammar of spoken Chinese*. Berkeley, CA: University of California Press.

Cheng, C., & Xu, Y. (2013). Articulatory limit and extreme segmental reduction in Taiwan Mandarin. *Journal of the Acoustical Society of America*, *134*, 4481–4495.

Cooke, J. D. (1980). *The organization of simple, skilled movements*. In G. E. Stelmach & J. Requin (Eds.), *Tutorials in motor behavior* (pp. 199–212). Amsterdam, The Netherlands: North-Holland Publishing.

Duanmu, S. (2000). *The phonology of standard Chinese*. New York, NY: 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*, 1863–1875.
Ernestus, M., & Warner, N. (2011). An introduction to reduced pronunciation variants. *Journal of Phonetics*, *39*, 253–260.
Fougeron, C., & Jun, S. A. (1998). Rate effects on French intonation: Prosodic organization and phonetic realization. *Journal of Phonetics*, *26*, 45–69.
Fourakis, M. (1991). Tempo, stress and vowel reduction in American English. *Journal of the Acoustical Society of America*, *90*, 1816–1827.
Fulop, S. A. (2011). *Speech spectrum analysis*. New York, NY: Springer.
Gandour, J., Potisuk, S., & Dechongkit, S. (1994). Tonal coarticulation in Thai. *Journal of Phonetics*, *22*, 477–492.
Gauthier, B., Shi, R., & Xu, Y. (2007). Learning phonetic categories by tracking movements. *Cognition*, *103*, 80–106.
Gay, T. (1978). Effect of speaking rate on vowel formant movements. *Journal of the Acoustical Society of America*, *63*, 223–230.
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, Germany: Mouton de Gruyter.
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*, 523–536.
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*.
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*, 266–280.
Kirchner, R. M. (1998). *An effort-based approach to consonant deletion*. PhD, University of California.
Klatt, D. H. (1975). Vowel lengthening is syntactically determined in a connected discourse. *Journal of Phonetics*, *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*, 1208–1221.
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*, 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 C. Gussenhoven & T. Riad (Eds.), *Tones and Tunes Vol 2: Experimental Studies in Word and Sentence Prosody* (pp. 211–237). Berlin, Germany: Mouton de Gruyter.
Lindblom, B. (1963). Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America*, *35*, 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, The Netherlands: Kluwer Academic Publisher.
Local, J. (2003). Variable domains and variable relevance: Interpreting phonetic exponents. *Journal of Phonetics*, *31*, 321–339.
Malécot, A. (1955). An experimental study of force of articulation. *Studia Linguistica*, *9*, 35–44.
Mermelstein, P. (1975). Automatic segmentation of speech into syllabic units. *The Journal of the Acoustical Society of America*, *58*, 880–883.
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*, 40–55.

Myers, J., & Li, Y. S. (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.

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*, 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*, 640–648.

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

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*, 1627–1641.

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

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

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

Tatham, M., & Morton, K. (2006). *Speech production and perception*. New York, NY: 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. (2005). Syllable contractions in a Mandarin Conversational Dialogue Corpus. *International Journal of Corpus Linguistics*, *10*, 63–83.

Tseng, S. C., & Liu, Y. F. (2002). *Annotation of spontaneous Mandarin*. *Technical Report No. 02–1*. Taipei: Chinese Knowledge Processing Group, Academia Sinica (in Chinese).

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

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. (1990). Formant frequencies of Dutch vowels in a text, read at normal and fast rate. *Journal of the Acoustical Society of America*, *88*, 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*, 121–127.

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

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, UK: Wiley-Blackwell.

Wong, W. Y. P. (2004). Syllable fusion and speech rate in Hong Kong Cantonese. *Proceedings of Speech Prosody-2004*, pp. 255–258.

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. (2013). *ProsodyPro — A tool for large-scale systematic prosody analysis*. *Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP 2013)*. Aix-en-Provence, France.

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*, 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 F0 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.

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

Zhao, Y., & Jurafsky, D. (2007). The effect of lexical frequency on tone production. *Proceedings of the 16th International Congress of Phonetic Sciences*, pp. 477–480.

Appendix A

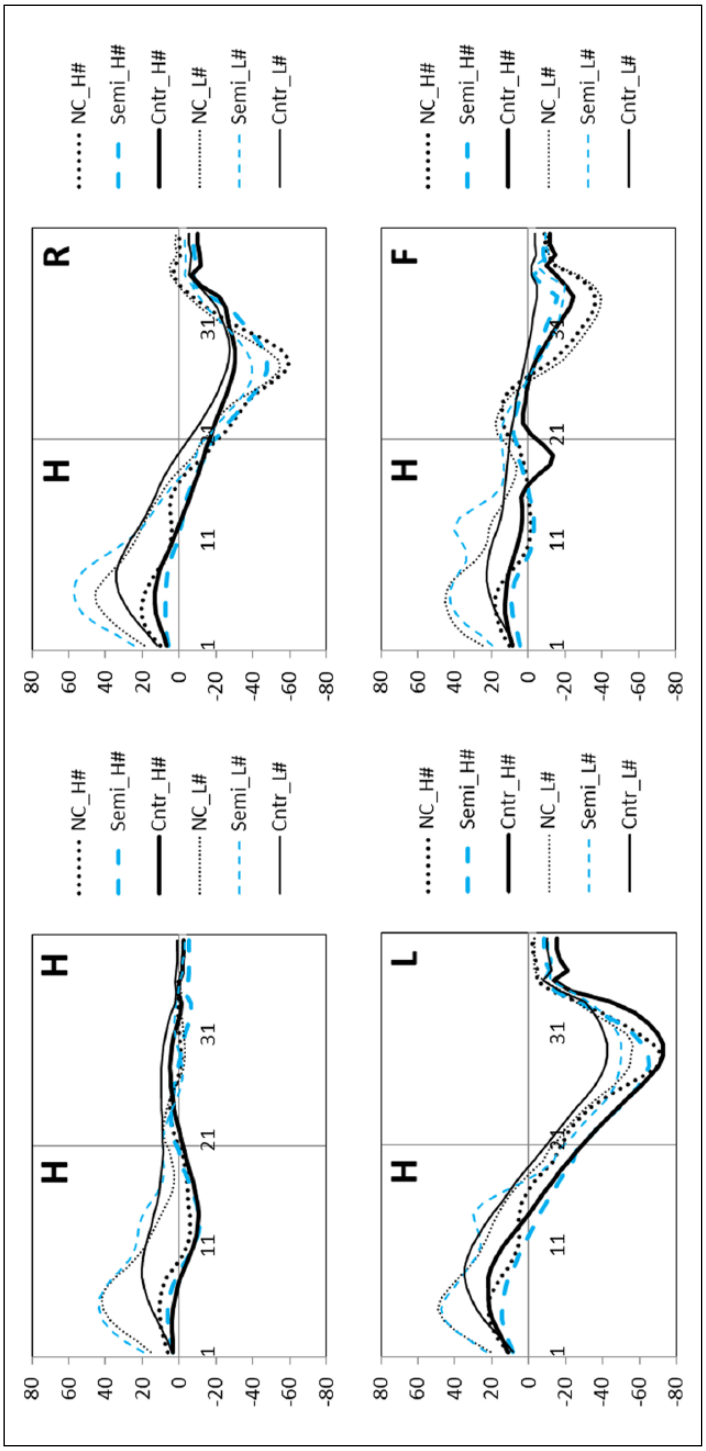


Figure 16. F0 velocity profiles of tone dyad HH, HR, HL and HF. The *x*-axis is normalised 40 measurement points and the *y*-axis is in units of semitones per second.

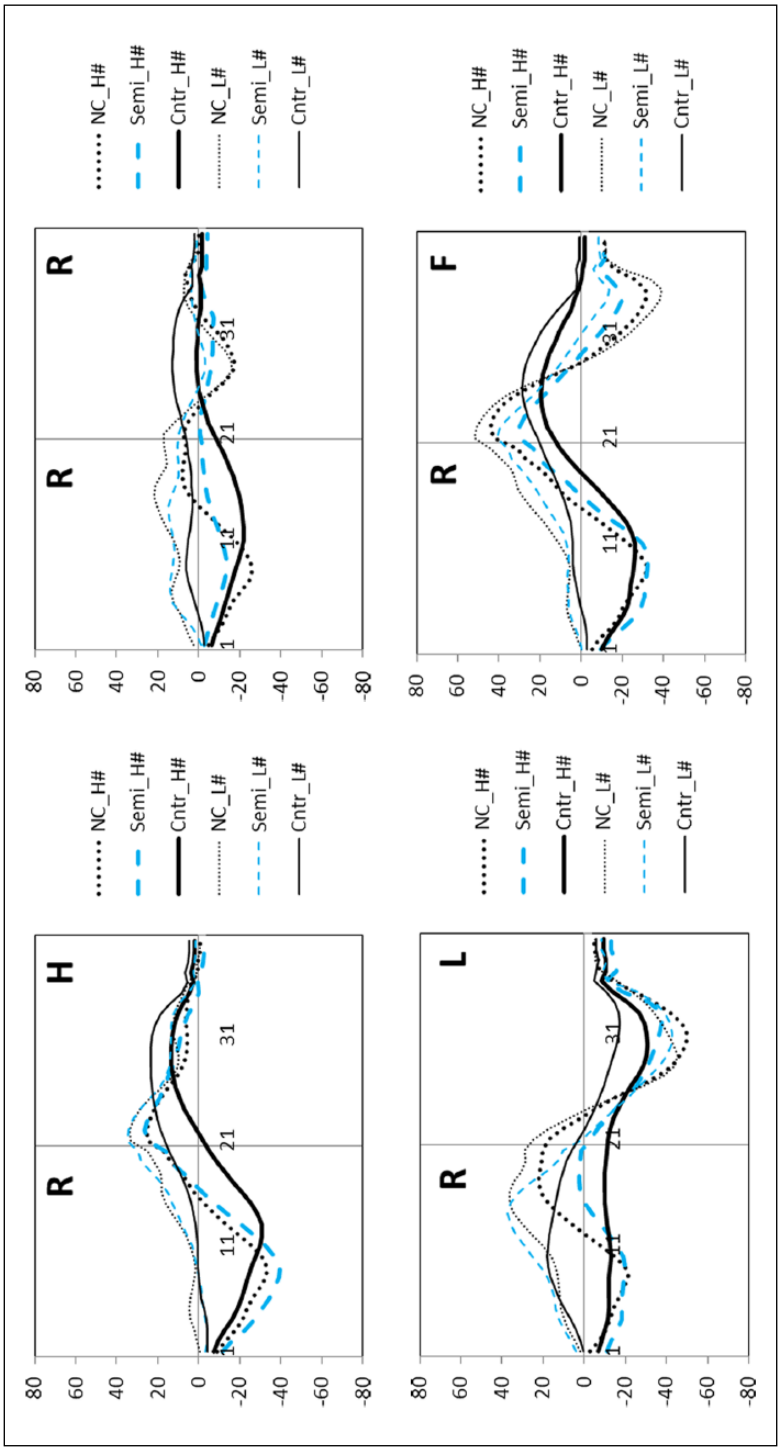


Figure 17. F0 velocity profiles of tone dyad RH, RR, RL and RF. The *x*-axis is normalised 40 measurement points and the *y*-axis is in units of semitones per second.

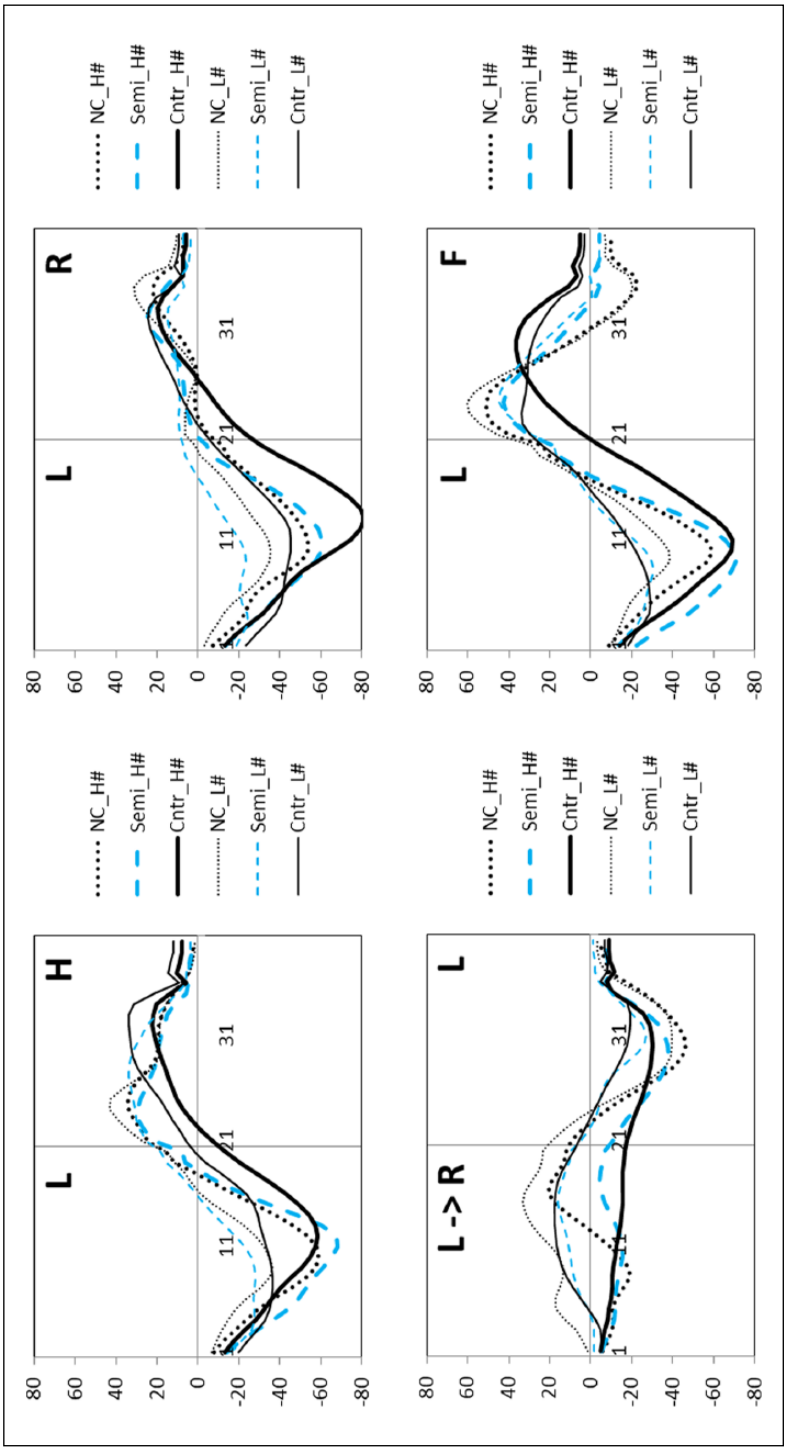


Figure 18. F0 velocity profiles of tone dyad LH, LR, LL -> RL and LF. The *x*-axis is normalised 40 measurement points and the *y*-axis is in units of semitones per second.

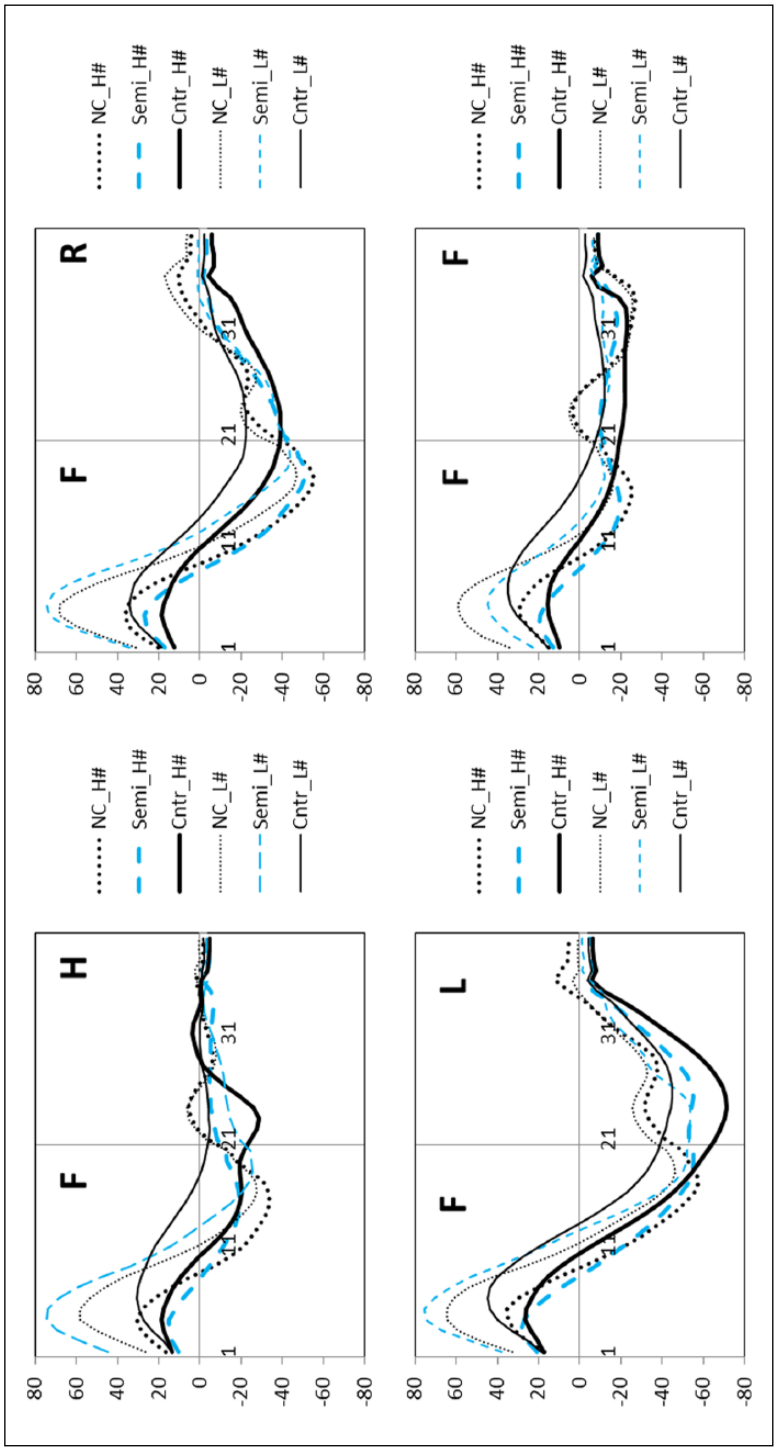


Figure 19. F0 velocity profiles of tone dyad FH, FR, FL and FF. The *x*-axis is normalised 40 measurement points and the *y*-axis is in units of semitones per second.

Appendix B

Table 3. Three separate one-way repeated measures analyses of variance using SPEED as the independent variable with respective dependent variables: DURATION (in seconds), F0 excursion SIZE (in semitones) and SLOPE of the regression line of F0 peak velocity over F0 movement amplitude of the three contraction types.

	DURATION	SIZE	SLOPE
Slow	0.188	5.51	12.56
Natural	0.137	2.71	17.85
Fast	0.088	1.70	23.59
F value	$F(2,10) = 24.06$	$F(2,10) = 14.99$	$F(2,10) = 14.77$
p value	$p < 0.001$	$p = 0.001$	$p = 0.001$

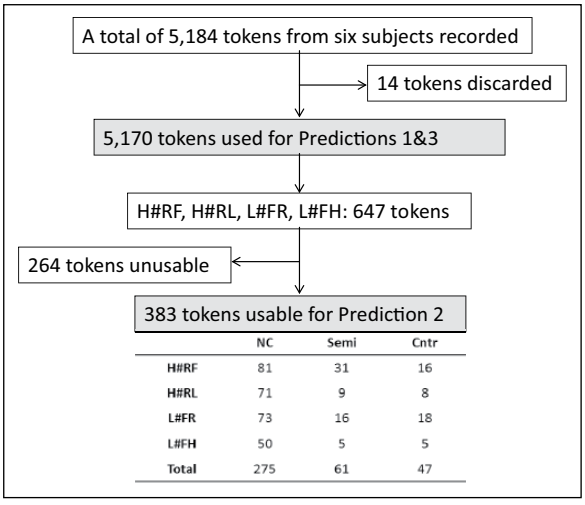
Note. Results of the post-hoc (Tukey) analyses on each dependent variable were:
(1) DURATION ([Slow > Natural], $p < 0.01$**; [Slow > Fast], $p < 0.001$***; [Fast > Natural], $p < 0.05$*); (2) SIZE ([Slow > Natural], $p < 0.05$*; [Slow > Fast], $p < 0.01$**; [Fast > Natural], $p = 0.53$); (3) SLOPE ([Slow < Natural], $p = 0.12$; [Slow < Fast], $p < 0.01$**; [Fast < Natural], $p = 0.09$).

Appendix C

Table 4. Percentage (%) with which contraction type occurred across different tone combinations. Row represents the tones (H, R, L, F) in the first syllable and column represents the tones of the second syllable. (NC: non-contracted; Semi: semi-contracted; C: contracted; H: high tone; R: rising tone; L: low tone; F: falling tone).

1st sylb. \ 2nd sylb.		H		R		L		F	
		(H)#	(L)#	(H)#	(L)#	(H)#	(L)#	(H)#	(L)#
H	NC	73.5	62.1	78.6	62.4	76.5	63.6	80.9	63.5
	Semi	10.5	11.8	8.2	12.4	11.1	13.0	7.4	15.7
	C	16.1	26.1	13.2	25.3	12.4	23.5	11.7	20.8
R	NC	66.7	60.5	64.4	56.2	73.5	56.8	75.3	56.8
	Semi	11.7	11.7	10.4	10.5	9.3	10.5	12.4	14.8
	C	21.6	27.8	25.2	33.3	17.3	32.7	12.4	28.4
L	NC	77.8	55.3	74.1	57.6	63.3	50.7	75.3	58.0
	Semi	8.6	14.0	13.0	10.4	14.6	12.7	6.2	13.5
	C	13.6	30.7	13.0	32.1	22.2	36.6	18.5	28.6
F	NC	75.9	67.9	67.3	58.4	63.6	63.0	75.3	57.4
	Semi	9.9	7.4	17.3	15.5	19.8	13.0	11.1	14.2
	C	14.2	24.7	15.4	26.1	16.7	24.1	13.6	28.4

Appendix D



	NC	Semi	Cntr
H#RF	81	31	16
H#RL	71	9	8
L#FR	73	16	18
L#FH	50	5	5
Total	275	61	47

Figure 20. Flowchart of the total number of tokens used for various data analyses.