Title: Coarticulation as synchronised CV co-onset – Parallel evidence from articulation and acoustics

Zirui Liu[a*], Yi Xu[a], Feng-fan Hsieh[b]
*Corresponding author: zirui.liu.17@ucl.ac.uk, Tel: +447360087940
[a] University College London, London, WC1E 6BT, United Kingdom
[b] National Tsing Hua University, Hsinchu City, 101, Section 2, Guangfu Rd, Taiwan

Abstract

This study tested the hypothesis that consonant and vowel are synchronised at the syllable onset, and that such synchronised co-onset is the essence of coarticulation. Articulatory data were collected for Mandarin Chinese, using Electromagnetic Articulography (EMA), and acoustic data were collected simultaneously. As a departure from conventional approaches, a minimal triplet paradigm was applied, in which divergence points between movement trajectories in contrastive pairs were used to determine segmental onsets. Triplets of disyllabic words consisting of two matching contrastive pairs in a $C_1V_1\#C_2V_2$ structure were used, whereby the consonant pair differed only in C_2 and the vowel pair differed only in V_2 (the numerical indices indicate syllable position). Both articulatory and acoustical results showed that the articulation of vowels and consonants started at about the same time, thus supporting the CV synchrony hypothesis. The realisation of CV synchronisation was dimension specific, however. For any particular articulator, only the dimensions free of consonantal requirement started their movements toward the vowel from the syllable onset, while the rest of the dimensions moved toward successive consonantal and vocalic targets. The finding of CV co-onset increases the amount of temporal overlap between C and V relative to the widely assumed CV asynchrony. The evidence of dimension-specific sequential articulation sheds further light on coarticulation by offering a timing-based explanation for the well-known phenomenon of coarticulation resistance.

Keywords
Coarticulation, CV synchrony, EMA, GAMMs, coarticulation resistance

Coarticulation as synchronised CV co-onset – Parallel evidence from articulation and acoustics

1.0 Introduction

Coarticulation has remained an unresolved issue since the term was first coined, which refers to the phenomenon that in a CV syllable, the vowel related articulatory postures can be observed during the articulation of the consonant (Menzerath & de Lacerda, 1933). The observation of the phenomenon, remarkably, can date back to as early as 1897, when Kymography was first used to study speech production. It was found that in a CV syllable, tongue movement for the vocalic target can be seen at the start of the consonantal movement (Rousselot, 1897-1901; Kühnert & Nolan, 1997). The exact nature of coarticulation, however, has remained a mystery and a matter of intense debate. One of the most critical uncertainties is the temporal domains of the segments involved, including both consonants and vowels. Due to this uncertainty, coarticulation as a scientific term has been widely used to refer only to influences of adjacent segments on each other (Hardcastle & Hewlett, 1999; Kühnert & Nolan, 1997), i.e., largely devoid of its original connotation. The present study attempts to tackle the difficulty of determining the temporal domain of segments by using a minimal triplet paradigm to determine the temporal alignment of C and V around syllable onset.

1.1 Historical Development of Coarticulation

Before spectrographic analysis became widely available, the earliest findings of extensive co-production between segments in CV syllables were based on articulatory data. During that time, many studies reported similar findings, such as the observation of labial or lingual articulation for the vowel at the start of consonant articulation in syllables such as /ku/ or /ba/ (Sievers, 1876; Stetson, 1951; Rousselot, 1897-1901; Kühnert & Nolan, 1999). Similar findings have also been reported recently in Polish, where vowel related articulation was detected near the beginning of /kV/ syllables (Gubian et al., 2019). A major theoretical development was the notion of the articulatory syllable, which was proposed based on recordings of lip movements during the production of /C_nu/ sequences in Russian (Kozhevnikov & Chistovich, 1965), where n denotes varying number of consonants. It was found that the lips begin protruding at the time of the first consonant regardless of the number of onset consonants. The articulatory syllable model, as shown in Figure 1a, states that the syllable is the domain of vowel articulation, and coarticulation occurs when no contradictory movements are involved between C and V.

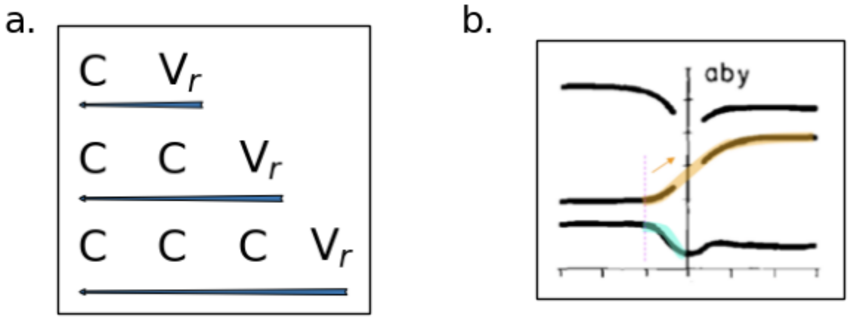

Figure 1

a. The articulatory syllable hypothesis (adapted from Kent & Minifie, 1977,). b. The anticipatory coarticulation that challenges it (Öhman, 1966, p.160).

The arrival of spectrographic analysis in the 1950s has given speech researchers a powerful tool for visualising the acoustic events of speech, and it has introduced many new assumptions, including the now widely accepted consensus that the start of a consonant is marked by sharp acoustic landmarks such as a sudden onset of stop closure, frication or nasal/lateral murmur (Turk et al., 2006). With these landmarks as the markers of consonant onsets, the articulatory syllable hypothesis was put in question by observations of spectral events associated with an upcoming vowel before the onset of a consonantal landmark, most notably by Öhman's classic spectrographic study in 1966. In Öhman (1966), continuous formant trajectories between contrastive V_1CV_2 sequences in terms of V_1 and V_2 are inspected. It is found that before the acoustic closure of the intervocalic C (e.g., the gap in F2 and F3 in Figure 1b), formant transitions toward the contrasting V_2 can already be observed (e.g., the upward movement of F2 before the gap in Figure 1b). This has made the notion of articulatory syllable appear too restricted (Kühnert & Nolan, 1999), because coarticulation seems to go *beyond* syllable boundaries. The alternative hypothesis is then formulated that the vowel properties occurring before the landmark-based consonant onset is due to *anticipatory coarticulation* (Daniloff & Hammarberg, 1973). This has since become an integral part of many coarticulation models, such as the coarticulation resistance model (Recasens, 1984), the time-locked model (Bell-Berti & Harris, 1979) and the hybrid model (Perkell & Chiang, 1986).

But treating acoustic landmarks as segmental boundaries is questionable, because it may not precisely reflect the relation between acoustics and articulation. It is true that the sharp landmarks are visually compelling, and they correspond well with the acoustic theory of speech production: stops involve momentary closure, fricatives involve frication, and sonorants involve abrupt shift of resonance cavities. But to produce these spectral patterns, the responsible articulators need to move in place from their prior positions associated with the preceding sounds. This means that the articulation of segments has to start before they manage to generate their prototypical landmarks. This is in line with arguments from the time-locked model that articulation starts not long before its acoustic landmarks (Bell-Berti & Harris 1979). Thus, a critical issue is whether the onset of a consonant is when its articulation commences, or when it has achieved its targeted closure or constriction. The articulatory syllable hypothesis seems to assume the former, while the anticipatory coarticulation hypothesis apparently assumes the latter. The view that articulation signals the onset of a segment also aligns with most contemporary work under the Articulatory Phonology (AP) framework (Nam et al., 2009; Shaw & Chen, 2019; Goldstein et al., 2006). Because of this difference, the presence of acoustic or articulatory movement toward the following vowel prior to the acoustic consonantal landmark does not necessarily contradict the articulatory syllable hypothesis.

In fact, based on the articulatory definition of segment onset, the finding of Öhman (1966) as well as the subsequent reports of anticipatory coarticulation (Rubertus & Noiray, 2018; Recasens & Pallarès, 2001; Magen, 1997; Mok, 2012) could be interpreted as evidence for the articulatory syllable. That is, the beginning of the formant movement toward the next vowel well before the consonant closure, as seen in Figure 1b, would suggest that the articulation of that vowel has started at that moment. Given also that F1 has started to move down toward a typical closure pattern as shown in Figure 1b as well as most of the tracked formant trajectories in Öhman (1966), the consonantal movement also started well ahead of the closure. In other words, if the onset of the formant movements toward the next segment is taken as the onset of the segment, the onset of the syllable itself can be taken as occurring well ahead of the consonant closure.

The articulatory syllable hypothesis as originally formulated does not fully specify how closely vowel and consonant are aligned with each other at the syllable onset, however. Yet it does state that coarticulation takes place only when no contradictory movements are required between C and V. This means that the articulation of some vowels may not start at the same time as the consonant if their movements are in conflict. This is consistent with the subsequent finding of coarticulation resistance (Bladon & Al-Bamerni, 1967). In a series of acoustic and electropalatographic studies, Recasens shows that the extent of the cross-consonantal vowel to vowel coarticulation is inversely proportional to the coarticulation resistance of the intervening C (Recasens, 1984; Recasens, 1987; Recasens, 1989). By focusing only on consonant resistance to vowel-to-vowel coarticulation, however, those studies did not examine the exact CV alignment at the syllable onset. In a more recent development of the AP framework, an explicit temporal overlap of gestures between onset C and the nucleus vowel is proposed, in the form of an "in-phase" coupling relationship (Nam et al., 2009). However, absolute synchrony in terms of articulation is not the current view in AP due to recent studies reporting positive CV lag (C preceding V) (Tilsen, 2020; Nam, 2007b; Shaw & Chen, 2019). To the best of our knowledge, direct evidence for CV synchrony has only been reported in German and Catalan in Mücke et al. (2012). The segmentation methods used in recent articulatory studies might have contributed to the finding of asynchronous CV onset, as will be reviewed next.

1.2 Determining Segment Onsets with a Minimal Triplet Paradigm

The method that has by now been conventionalised in articulatory studies (especially under the AP framework) is to use a peak velocity threshold, usually 20%, as the marker of segment onset and offset (Hoole et al., 1994). That is, given that the velocity profile of a unidirectional gesture is unimodal (Nelson, 1983), the onset of that gesture is said to be at the point when velocity has increased to 20% of its peak. This method is originally developed in Hoole et al. (1994) to investigate vowel production in German. It has subsequently been applied in studying various languages, including Mandarin Chinese (Gao, 2009; Marin & Pouplier, 2008; Marin & Pouplier, 2014; Shaw & Chen, 2019; Shaw et al., 2011;Yin et al., 2012), and is used as the default segmentation setting in the Mview program in MATLAB (i.e., the findgest() function; Tiede, Haskins Laboratories; Danner et al., 2018), which is one of the most widely used analytical tools for articulatory data. However, as pointed out from the beginning by Hoole et al. (1994), segment boundaries determined by the threshold method is sensitive to articulatory stiffness, i.e., the velocity threshold is achieved earlier for segments articulated with higher stiffness. For example, we simulated the process of approaching a numeric target of 100 with the qTA model (could be viewed as approaching a tonal target of 100 Hz or a spatial target of 100 mm) (Prom-on et al., 2009). All else being equal, the lambda value was varied from 20 to 80. The lambda value represents the stiffness parameter which is equivalent to that of the Task Dynamics (TD) model (Nam, et al., 2012). As Figure 2 shows, the higher the stiffness, the earlier the velocity peak and the earlier the achievement of the 20% threshold (marked by the colour coded triangles).

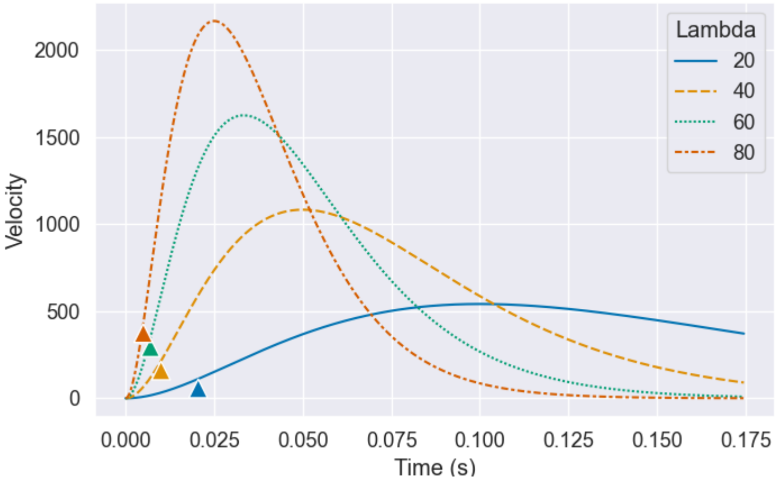

Figure 2

Simulated velocity profiles with the qTA model (Prom-on et al., 2009) with varying stiffness. The triangles mark the points where 20% of each contour's peak is achieved.

Indeed, a large body of research suggests that consonants are articulated with higher stiffness than vowels (Saltzman & Munhall, 1989; Pastätter & Pouplier, 2014; Nam, 2007a; Nam et al., 2012), which may explain the consistently reported earlier detection of C onset than V onset. Furthermore, articulatory stiffness may inherently vary between articulators across phonetic contexts, due to differences in muscle mass (Roon et al., 2021). Roon et al. (2021) used peak velocity over maximum displacement to measure stiffness and found that the tongue back has lower stiffness than both the lips and the tongue tip in certain onset positions. The intrinsic stiffness difference between CV and the articulators used to identify segment onsets may have jointly led to the findings of positive CV lags in a number of studies on Mandarin (Gao, 2009; Shaw & Chen, 2019), whereby movement of the lips and the back portion of the tongue were respectively used to determine the onsets of C and V in syllables consisting of labial consonants and mid or back vowels (e.g., /ma/). In Mücke et al. (2012), instead of applying a percentage threshold, the segment onsets were located at when velocity crossed the zero point, which may have avoided the stiffness confound and aided their finding of CV co-onset.

What makes the velocity method even more problematic is the potential confound introduced by gestural overlap (Saltzman & Munhall, 1989). The overlap would mean that any observed movement trajectory may consist of multiple gestures, which would make it hard to know which of the overlapped gestures is being segmented when applying the velocity threshold method. The problem is especially severe if the confounding gestures cannot be anticipated by the researcher. For instance, an articulatory study investigating CV alignment would avoid using syllables such as /ji/, as both C and V require the same articulators. However, confounding gestures might still be present for segments that seemingly have very different gestural specifications. To avoid such *covert* confounds, Gelfer et al. (1989) used minimal pairs as a control method to separate consonantal and vocalic effects in the orbicularis oris inferior (OOI) EMG activity. They showed that a portion of the OOI EMG activity after [i] in /iC_nu/ is not a rounding gesture in anticipation of /u/, because the same EMG activity also occurred in the control sequence /iC_ni/ (presumably to reverse the lip spreading gesture for /i/, Nalborczyk et al., 2020). Using a similar control method, Boyce et al. (1990) looked at velar lowering movement between the sequences /lasal/ and /lansal/. They showed that the same velar lowering can be observed for both sequences early on, but a second lowering gesture can be observed only in /lansal/ but not in /lasal/, suggesting that the initial lowering movement belongs to the vowel. Thus, what had been thought to be evidence of extensive anticipatory nasalisation was due to a velum lowering gesture associated with the vowel.

The application of the minimal pair paradigm in those studies has helped to reduce the amount of observed anticipatory coarticulation. But it has not addressed the issue of how exactly consonants and vowels are temporally aligned. To assess the CV alignment, the onset of the consonant and the onset of the vowel need to be first estimated respectively. Then the temporal locations of the estimated onsets of C and V can be compared to assess how closely they are aligned. This means that not only a minimal pair of consonants and a minimal pair of vowels are needed, but also the two minimal pairs need to be closely matched to form a minimal triad for the final estimation of CV alignment. A method that can achieve this has been developed in Xu & Gao (2018) (derived from Xu, 2007), whereby triplets of Mandarin disyllabic words were used, each consisting of a consonant minimal pair and a vowel minimal pair which together form a C-V minimal pair. The two minimal pairs are made to closely resemble each other by sharing a word that contrasts with the other two words in either consonant or vowel, respectively.

An example of the paradigm is shown in Figure 3, where trajectories of F2 in a triplet of Mandarin words of $C_1V_1\#C_2V_2$ structures are plotted. The first two words in a triplet differ in C_2 — /l/ vs. /j/ (between ‘louliw’ and ‘louyiw’), and the second two differ in V_2 — /i/ vs. /u/ (between ‘louliw’ and ‘louluw’). The design brings forth the difference between the F2 movements of /l/ and /j/ in the consonant contrast pair (dashed blue vs. solid black) and that of /i/ and /u/ in the vowel pair (solid black vs. dotted red), allowing direct estimation of the articulatory onsets of both C and V. As can be seen, the bifurcation due to the consonant contrast and that due to the vowel contrast start around the same time, suggesting temporal synchrony of consonant and vowel.

So far, the minimal triplet paradigm has been applied only to acoustic data (Xu, 2007; Xu & Gao, 2018), and without direct comparison with articulatory data. It could be the case that the F2 trajectories are not sufficiently detailed to provide enough precision in assessing the alignment of consonants and vowels. The present study applies the minimal triplet paradigm to articulatory data to investigate the temporal alignment of consonants and vowels in Mandarin. At the same time, F2 trajectories are obtained from the same speech utterances and directly compared to the articulatory trajectories.

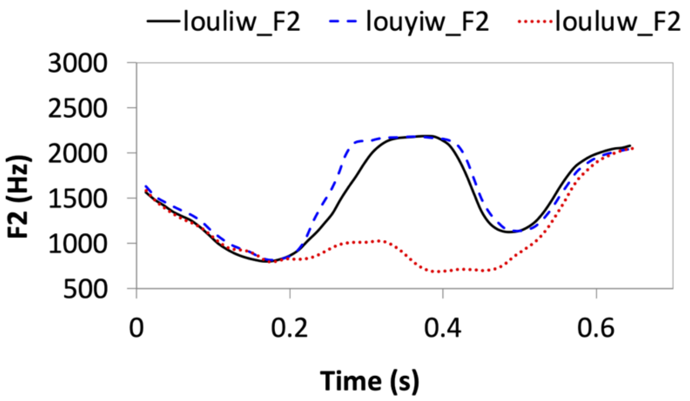

Figure 3

Mean F2 trajectories of CV contrastive pairs. Pinyin and corresponding IPA transcription: louliw – [loʊliw]; louyiw – [loʊjiw]; louluw – [loʊluw].

2.0 Experimental methods

Continuous Electromagnetic Articulography (EMA) and simultaneous formant trajectories were used to establish CV alignment in Mandarin syllables under the minimal triplet paradigm. The basic method is to find triplets of disyllabic words in which one word differs from the other two either in a single consonant or in a single vowel. The shared word ensures that the three members of the triplets form two closely matched minimal pairs, each differing only in one sound. The articulatory as well as the formant trajectories can then be compared to show when the contrasting trajectories start to deviate from each other. Because everything else is made identical, the bifurcation of the contrasting trajectories is unambiguous. The close match of the two minimal pairs within a triplet thus allows direct comparison of the timing of the consonant and vowel onsets.

2.1 Stimuli

A total of six triplets, consisting of 18 C_1V_1#C_2V_2 disyllabic words are used as stimuli, where the numerals indicate syllable number and # indicates syllable boundary. As shown in Table 1, in each triplet, there is a vowel contrast between the first and second word in terms of V2 – /i/ vs. /u/ and a consonant contrast between the first and third word in terms of C2 – /l/ vs. /j/. Note that there is both acoustic (Shih, 1995) and articulatory (Zheng & Bao, 2002) evidence that a consonantal glide /j/ is produced even before /i/ in Mandarin. All 18 words bear the Rising tone (as marked by ‘ ́ ’ in their Pinyin format. All target words were embedded in the carrier phrase “bǐ ___ wěi shàn” ([bi ___ weɪ ʂan]), meaning “more hypocritical than ___”. The words in the vowel pairs are made-up personal names. The third word in each triplet means Aunt C_1V_1 (i.e., the first word is a Surname in Chinese, and yí means Aunt). All the words used are novel combinations in Chinese. This minimised word frequency effects since they are all low frequency words. 10 participants were instructed to read aloud the sentences with 10 repetitions each in randomised blocks, which yielded 1800 (10×18×10) tokens in total.

Table 1 ***Stimuli***

Triplet	Pinyin	Chinese	IPA	Pinyin	Chinese	IPA	Pinyin	Chinese	IPA
1	láilí	来黎	[laɪli]	láilú	来卢	[laɪlu]	láiyí	来姨	[laɪji]
2	léilí	雷黎	[leɪli]	léilú	雷卢	[leɪlu]	léiyí	雷姨	[leɪji]
3	lóulí	娄黎	[loʊli]	lóulú	婁卢	[loʊlu]	lóuyí	娄姨	[loʊji]
4	málí	麻黎	[mali]	málú	麻卢	[malu]	máyí	麻姨	[maji]
5	máolí	毛黎	[maʊli]	máolú	毛卢	[maʊlu]	máoyí	毛姨	[maʊji]
6	nílí	倪黎	[nili]	nílú	倪卢	[nilu]	níyí	倪姨	[niji]

2.2 Speakers

7 male and 3 female native speakers of Mandarin Chinese living in Taiwan participated as subjects. All of them were studying at the National Tsing Hua University, and are from the northern part of

China (5 from Beijing and 5 from Liaoning who speak Mandarin fluently). No speech or hearing difficulties were reported from the subjects prior to data collection.

2.3 Data Collection and Processing

Data collection was done at the Phonetics Laboratory at the Institute of Linguistics, National Tsing Hua University. Articulatory data were collected while subjects read aloud the stimuli using the NDI Wave system. Kinematic data were sampled at a rate of 200 Hz, with the distance value converted from voltage with a filter cut-off frequency of 40 Hz for the tongue tip and 20 Hz for the lips. The origin of the coordinate system was placed on the lower front edge between the upper incisors. Acoustic data was recorded simultaneously with a sampling rate of 24 kHz. EMA receiver coils were glued onto the articulators (the upper and lower lips, tongue tip, tongue blade, and tongue dorsum) with an addition of four location reference receivers placed on the upper incisors (origin point), nasion, and the left and right mastoids. All participants sat next to the NDI Wave field generator with the receiver coils in place and read aloud the stimuli. The sentences were displayed in front of the speakers on a screen at a comfortable pace.

The acoustic data were manually annotated at the acoustic landmarks of syllable boundaries in the format of [$C_1V_1C_2V_2$weɪ] with a Praat script, which extracted the formant data and segmented the EMA data at the same time. The start and end boundaries were respectively located at the acoustic onset of C_1 (e.g., nasal murmur in /mali/) and the end of voicing in /weɪ/ (see Figure 4). The formant data were generated with a custom version of FormantPro in Praat with the default parameters (window length = 0.025 s; female maximum formant = 5500 Hz; male maximum formant = 5000 Hz; dynamic range = 30dB; pre-emphasis from 50 Hz) (Boersma & Weenink, 2005; Xu & Gao, 2018). The formants are visually checked during annotation and the FormantPro algorithm trimmed off any irregular spikes. Two measures were taken to ensure temporal consistency between the articulatory and acoustic data. First, all trajectories were aligned at the first acoustic syllable boundary. Second, the trajectories were all sampled at 5 ms intervals. Thus, the sample points between articulatory and formant trajectories all correspond in real time.

Speaker 4's data was excluded from analysis due to background noise, which made it difficult to determine the acoustic landmarks. Out of the remaining 1620 tokens, 22 were excluded due to mispronunciation or spontaneous pausing.

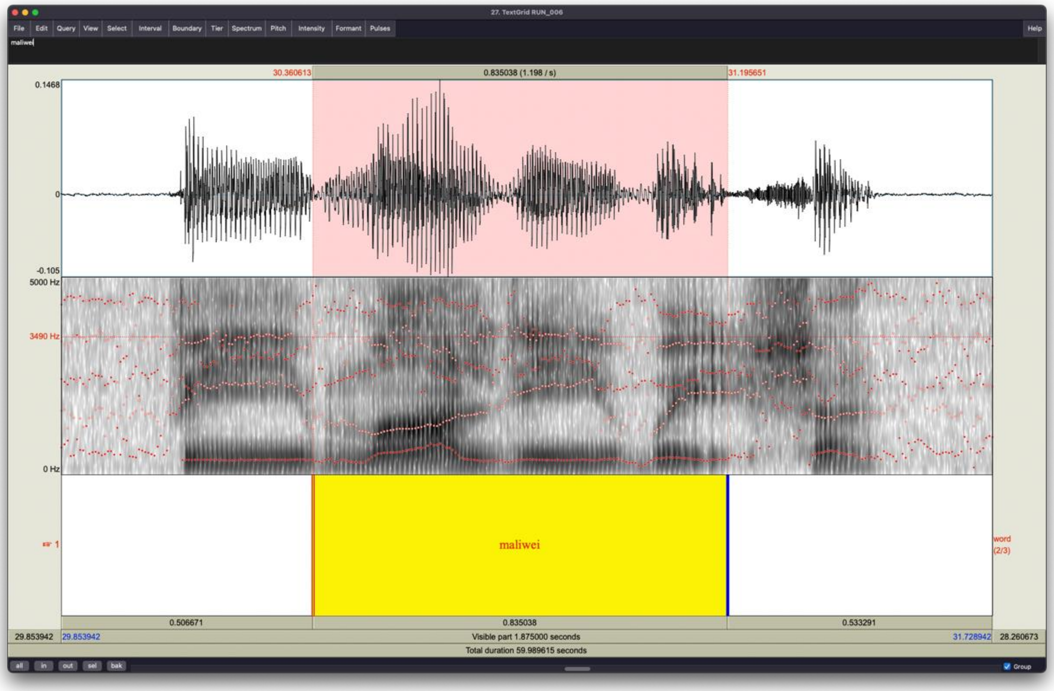

Figure 4

Annotation example.

2.4 Data Analysis

2.4.1 Measurements Used to Detect CV Onsets

To determine C onset based on divergence between the consonant pair, the tongue tip in the vertical dimension (TTy) is used. The choice is motivated by coarticulation resistance studies on /l/ in Catalan, as the tongue tip in the articulation of /l/ would be the least affected by the concurrent vowel. Previous ultrasound studies have shown that at the acoustic midpoint of /l/, the least variation is seen between different vowel contexts for the tongue tip, especially in the vertical dimension (Recasens & Rodríguez, 2016; Recasens & Espinosa, 2009). In terms of /j/, palatographic data show that the tongue tip is not crucial for its articulation (Recasens, 1990). Also, it is important to note that /l/ in Mandarin is always clear, which does not actively involve the tongue dorsum, unlike its dark counterpart in English (Smith, 2010). Therefore, as suggested by past studies, TTy divergence between /li/ and /yi/ can reliably reflect the articulatory onset of /l/.

The lip rounding contrast between /i/ and /u/ has been used in various studies to investigate the articulation of /u/ (Gelfer et al., 1989; Boyce et al., 1990; Bell-Berti & Harris, 1981). In addition, /y/ is a phoneme in Mandarin which contrasts /i/ in terms of lip rounding, therefore, for the sake of categorical distinction, lip spreading occurred during articulation of /i/ due to contraction of the zygomaticus major muscles (Nalborczyk et al., 2020). Thus, the lips are actively controlled for the articulation of /i/. To determine vowel onset by contrasting /li/ with /lu/, upper lip protrusion (LP) is used.

For acoustic analysis, F2 was used as the measurement, as it well reflects the contrasts between /i/ and /u/ and between /j/ and /l/, as shown in Xu & Gao (2018).

2.4.2 Determining Significant Divergence Time Point with Generalised Additive Mixed Models (GAMMs)

To obtain time points at which trajectories diverge significantly, GAMMs were used for their ability to model non-linear time series contours, while also accounting for random variabilities in the pattern, which is similar to the concepts of random effects in linear mixed effects models (Winter & Wieling, 2016). GAMMs were constructed for each minimal pair in each triplet (e.g., 'laili' vs. 'lailu' as the vowel minimal pair in t1) for each speaker, respectively, with TTy for C onset, LP for V onset and F2 for acoustic onset. In other words, for the same speaker, 10 repetitions of each word in a minimal pair are compared in each GAMM. C and V onset times were determined by when the model indicated a statistically significant difference between the trajectories in the C and the V pairs respectively. In order to satisfy the equal sampling point requirement of GAMMs, each utterance was trimmed to be the same length as the *shortest* utterance across all repetitions and speakers (465 ms). An example is shown in Figure 5 with the trimming point indicated by the vertical dotted line. Since movements towards the contrasting segments take place well before 0.465 s, the trimming procedure did not affect the final results. To ensure that all analysis is done on the real time scale, time normalisation (Wieling , 2018) was not used.

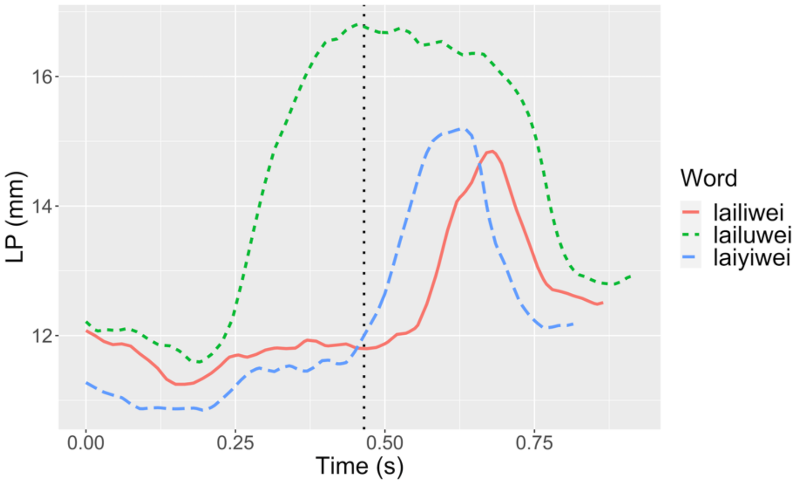

Figure 5

LP movements for the first repetition of t1 for speaker 1. The dotted black line corresponds to the trimming point at 0.465 s.

Prior to model construction, articulation and formant data was normalised for each speaker using z-score transformation (Lobanov, 1971). According to previous research, the contraction speed of the muscles in the tongue and lips averages around 40-50 ms per cycle (Blair, 1988; Ito et al., 2004). Therefore, to avoid type 1 error, only significant divergences that lasted for longer than 40 ms were recorded as an onset.

Two examples of determining the onset times by GAMMs are shown in Figure 6. The top-left graph shows that the LP position between the vowel pair becomes significantly different over time when the confidence intervals of the trajectories do not overlap. The second column shows differences between the trajectories for each pair, and the windows of significant difference are highlighted by the red lines. C and V onsets in /li/ in each triplet and for each speaker are identified when significant difference can be detected, e.g., 0.17 s for the vowel pair and 0.18 s for the consonant pair in Figure 6.

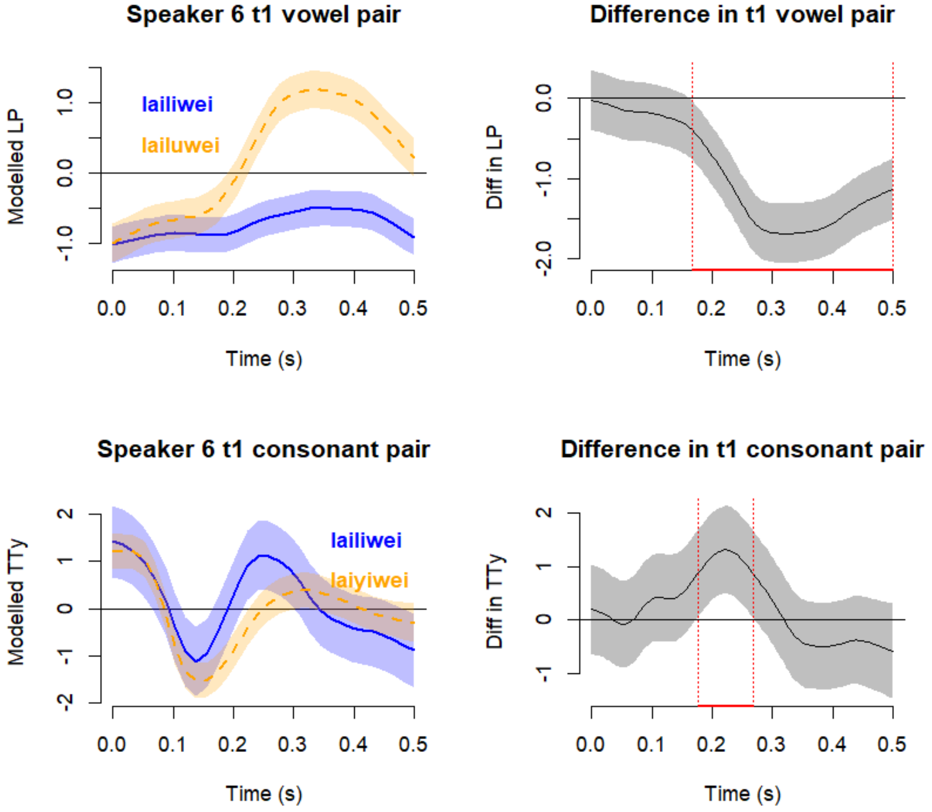

Figure 6

Articulatory trajectories modelled by GAMMs for speaker 6 and triplet 1. The shaded ribbons represent 95% of the confidence interval.

Models were constructed in *R* with the bam() function provided by the *mgcv* package (Wood, 2019). For each model[1], word was included as the main effect, and the measurement of interest was specified

[1] R syntax for the lower model of Figure 6:
model<- bam(TTy ~ Word + s(Time, by=Word, k=15) + s(Time, Repetition, by=Word, bs='fs', m=1), data=data)
# account for autocorrelation
model_ACF<- acf_resid(model)
model<- bam(TTy ~ Word + s(Time, by=Word, k=15) + s(Time, Repetition, by=Word, bs='fs', m=1), data=data, rho=model_ACF[2], AR.start=data$start.event)

as the dependent variable (e.g., TTy position). The models included a by-word smooth function through time to investigate articulatory or acoustic changes over time, and a random smooth (i.e., analogous to a full random effect in linear mixed models) to account for non-linear variation between repetitions. The *k* parameter was set to be 15 following model diagnosis by the gam.check() function. To rectify the violation of independent model errors due to autocorrelation in the model residuals, an AR-1 correlation parameter was estimated and incorporated into the models using the scf_resid() function provided by the package *itsadug* (Van Rij et al., 2017). 108 onset times (12 minimal pairs × 9 speakers) were collected from GAMMs for further analysis.

2.4.3 Comparison of C and V Onset Time

Linear mixed effects models[2] (LMEMs) were fitted in R using the *lme4* package (Bates et al., 2019) to compare C and V onset time collected from the GAMMs. The models were fitted with the Maximum Likelihood criterion, and all included a fixed effect of contrast/onset type (vowel vs. consonant). Speaker and triplet were included as random intercepts, and no random slopes were included to avoid singular fit. To test whether the fixed effect was significant, likelihood ratio tests were performed by comparing model likelihoods with and without onset type as the fixed effect using the anova() function in R. The fixed effect's *t*-value from the LMEM output and the X^2 value along with its associated *p*-value from the likelihood ratio test are reported in the results section.

2.4.4 Validating the Null Effect with Bayes Factors

A non-significant result from the LMEM analysis can either mean that there is a true absence of effect, or that the data is insufficient for the analysis to detect an effect. In other words, even if the effect of onset type is non-significant, we cannot conclude that the null hypothesis is true (i.e., evidence for synchrony). However, Bayesian statistics can help us make statements about a hypothesis given the observed data, and the evidence can go both ways – the result can indicate if the evidence is for the null hypothesis, the alternative hypothesis or that there is not enough evidence for either (Dienes, 2016; Dienes, 2014; Lakens et al., 2020; Harms & Lakens, 2018). This can be achieved with the Bayes factor (BF), which can be derived from the posterior distribution from the Bayes' theorem (Stone, 2013):

$$\mathrm{p}(\boldsymbol{\theta}|\mathrm{Data}) = \frac{\mathrm{p}(\mathrm{Data}|\boldsymbol{\theta}) \cdot \pi(\boldsymbol{\theta})}{\mathrm{p}(\mathrm{Data})}$$

$\mathrm{p}(\boldsymbol{\theta}|\mathrm{Data})$ is the posterior distribution of the parameters vector $\boldsymbol{\theta}$ given the data, $\pi(\boldsymbol{\theta})$ is the prior distribution of the parameters, $\mathrm{p}(\mathrm{Data}|\boldsymbol{\theta})$ is likelihood function of the model and p(Data) is the normalising constant (Harms & Lakens, 2018). The Bayes factor for indicating how the data support the *null (0) over the alternative model (1)* is calculated as:

$$BF_0 = \frac{\text{Posterior odds}_{01}}{\text{Prior odds}_{01}}$$

where:

$$\text{Posterior odds}_{01} = \frac{\mathrm{p}(\boldsymbol{\theta}_0|\mathrm{Data})}{\mathrm{p}(\boldsymbol{\theta}_1|\mathrm{Data})}$$

Since we do not have any strong prior belief about how likely the null model (for synchrony) or the alternative/full model (for asynchrony) is true, we can set the prior odds to be equally likely, i.e., 1 (Dienes, 2016). Note that the prior odds are not model priors. Therefore, BF_0 indicates that given our data, how much more likely the null hypothesis is true compared to the alternative. A BF close to 1 suggests that the evidence is not discriminative between the two, and a BF between 1 to 3 provides only "anecdotal" evidence, while a BF larger than 3 is considered valid evidence (Schönbrodt et al., 2017; Dienes, 2016; Jeffreys, 1961; Lee & Wagenmakers, 2013). Therefore, we consider the synchrony hypothesis to be supported by the data if BF_0 is larger than 3. According to Dienes (2016),

[2] R syntax for LMEM:

```
full_model<- lmer(Diverge ~ Type + (1|Speaker)+ (1|Triplet), data=data, REML=FALSE)
```

a number of high-powered studies reporting replication failures actually only achieved a BF of around 1, which illustrates the sensitivity of BF at distinguishing H_0 from H_1.

Similar to the LMEM analysis, Bayesian hierarchical modelling was used to include random intercepts for speaker and triplet[3]. Model construction is done in R using the *brms* package (Bürkner et al., 2021). For model priors (i.e., $\pi(\theta)$), due to the lack of assumptions about any effects on onset times, we used weakly informative Gaussian priors with a mean of zero for both models (Nalborczyk et al., 2019). For the Bayesian analysis, we will report BF_0, the mean and 95% confidence intervals of the effect of onset type in the full model, as well as the intercepts for both models.

2.4.5 Comparison between the velocity threshold method and the minimal triplet method

To demonstrate how using velocity profiles could give rise to incorrect segmentation, we analysed the data from two triplets (t1 and t5) with both methods and compared their respective results. Specifically for the velocity method, the onset of /u/ in /lu/ was determined by locating the onset for each token at when velocity reaches 20% of the peak velocity for the upper lip. For the minimal contrast method, results from section 2.4.2 were used.

2.5 Control for Speech Rate Variation

Because all the trajectories were aligned at the acoustic closure of C_1 in the current study, speech rate variation could confound the divergence analysis. For example, if the word 'lailu' is spoken consistently slower than the word 'laili' and 'laiyi' for a particular speaker, the movement towards V_2 in 'lailu' would be later than that in 'laili'. This mismatch in speech rate would result in delayed detectable divergence in the vowel pair but not the consonant pair. Moreover, it is well known that word frequency has an effect on duration – low frequency words are spoken with longer duration (Wright, 1979). By checking the durations of all the tokens, we can determine whether word frequency is a compromising factor in the current stimuli design.

To see how much variation there is between words in each triplet, durations of all tokens were calculated. For each token, the duration is measured as the time lapse from the acoustic onset of C_1 to the end of voicing in /weɪ/. The duration distribution for each word is plotted for each triplet and speaker in Figure 7. It can be seen that for each GAMM/divergence analysis, the duration differences between minimal pairs are very small. For instance, the top right subplot shows the duration distribution for triplet 6 (t6: 'nili' vs. 'nilu' vs. 'niyi') and speaker 1. Duration between the V pair is shown by the red and green distributions and the C pair by the red and blue distributions. It can be seen that the distributions are heavily overlapped between both contrastive pairs. Furthermore, the mean values are also very close together. Most importantly, Figure 7 shows that there are no systemic variations in token duration, e.g., tokens with /li/ as the target syllable is not consistently shorter or longer than the those with /lu/ as the target. Such minimum random variation between speech rates should be accounted for by the GAMMs, and the current stimuli design should suffice for the purpose of this study.

[3] R syntax for Bayesian Hierarchical model:
full_model <- brm(Diverge ~ Type + (1|Triplet) + (1|Speaker), data=data, family=gaussian(), prior=prior_full, warmup=2000, iter=7000, save_pars=save_pars(all=TRUE))

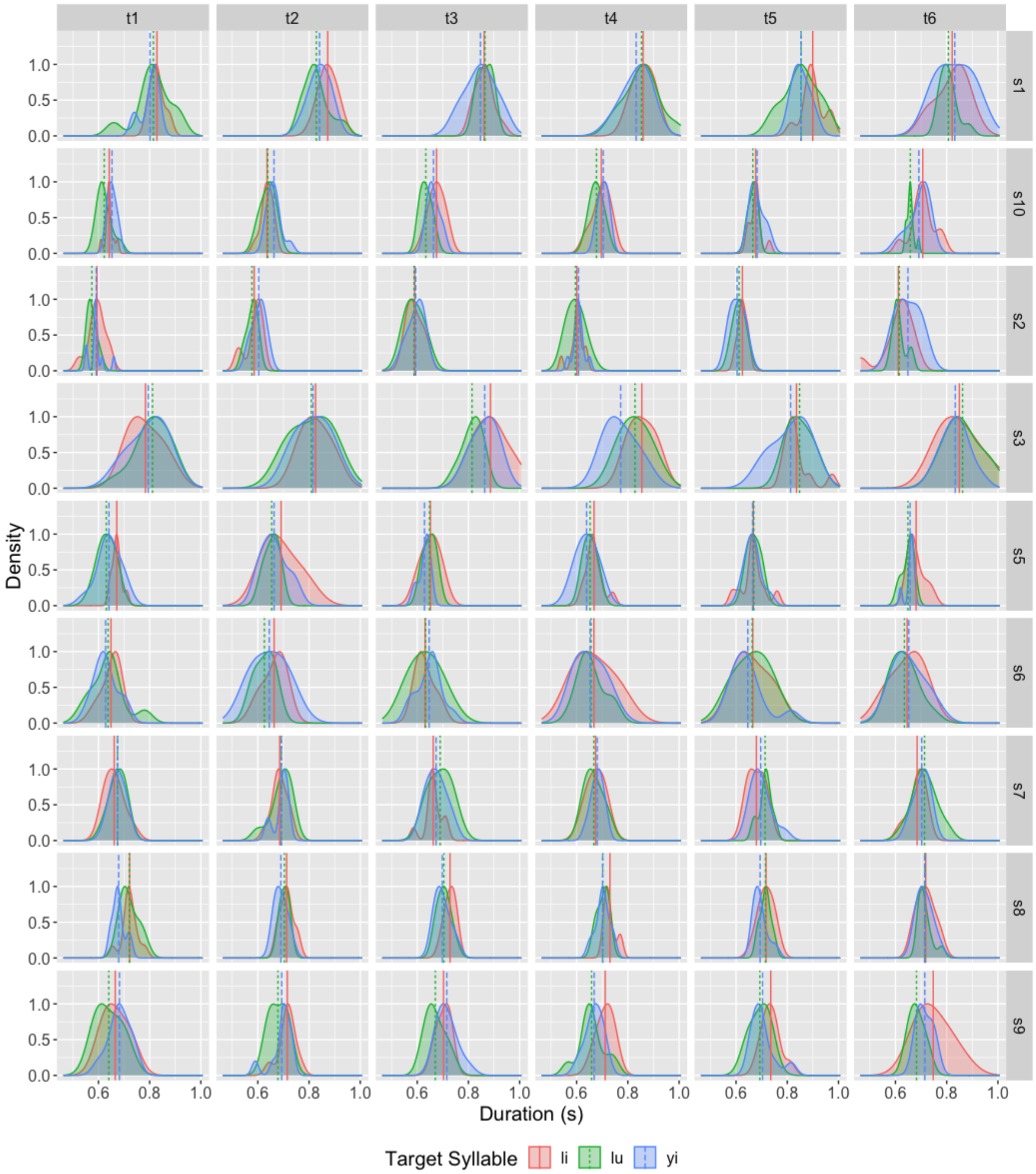

Figure 7

Probability density plot of token durations for all speakers and triplets. The columns correspond to triplets and rows correspond to speakers. The mean values are indicted by the colour coded lines.

3.0 Results

3.1 Acoustic Results with F2

Figure 8 shows the mean F2 movements of all 6 triplets. Overall, similar patterns to Xu & Gao (2018) can be observed. For each triplet, F2 between the consonant pair (e.g., t1: 'laili' vs. 'laiyi') start to diverge around the same time as it does for the vowel pair (e.g., t1: 'laili' vs. 'lailu'). Note that the F2 trajectories plotted here are not speaker normalised like the data used in GAMMs.

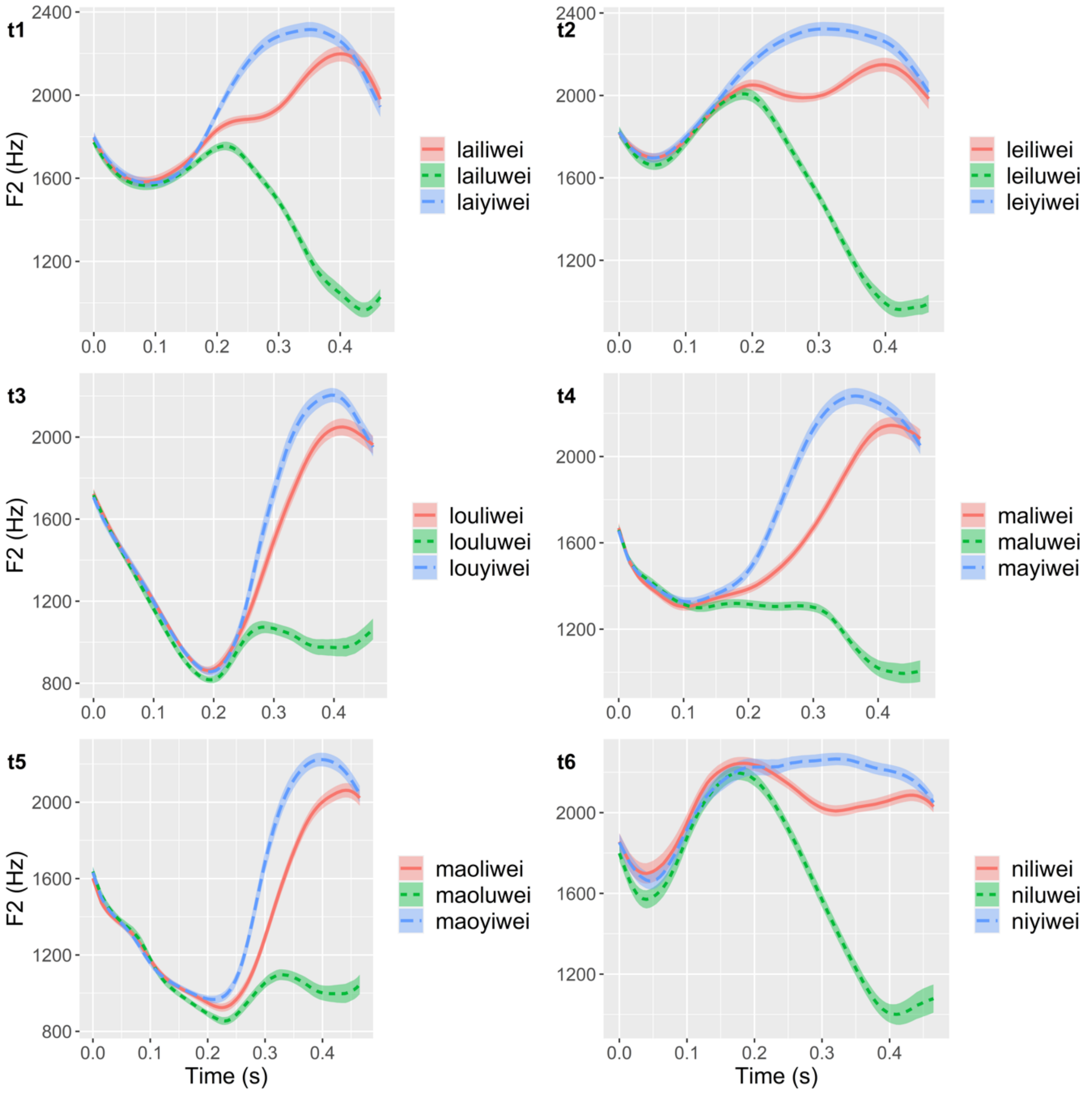

Figure 8

Mean F2 trajectories in Hz over time for each triplet, averaged across all speakers and repetitions. The shaded ribbons indicate standard error of the mean.

Figure 9 shows mean F2 onset time collected from the GAMMs, which is averaged across speakers. Although some variability can be seen between C and V onsets across triplets, there is no consistent effect of segment type. In other words, C and V onsets are temporally very close to each other. A likelihood ratio test indicates that the effect of onset type on onset time is not significant (t = 0.97 (LMEM output); $X^2(1)$ = 0.93; p = 0.34).

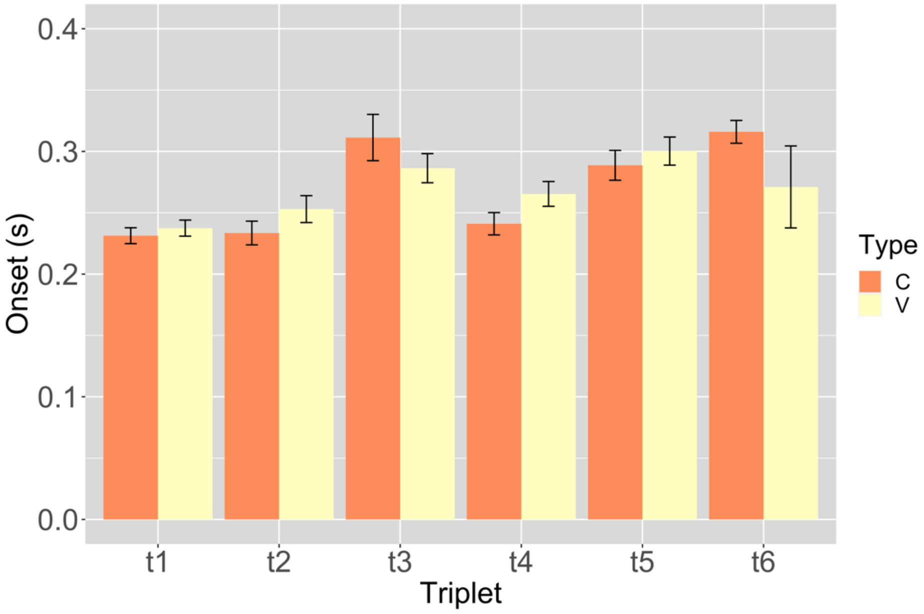

Figure 9

Acoustic CV onset time by segment type and triplet determined by GAMMs. The error bars show standard error of the mean.

Results from the Bayesian analysis are plotted in Figure 10 (top panel for the full model and bottom for the null model). In order to check for divergent transitions which could arise during model construction, the parallel coordinates plots are plotted on the right in Figure 10. Divergent transitions refer to when the sampling process goes wrong for certain iterations and corrupts the parameter estimation process (Gabrys et al., 2019). The parallel coordinates plots visualise each iteration as a line connecting the estimated parameters. Potential divergences diagnosed by the nuts_params() function are highlighted in green (Bürkner et al., 2021). Indications of true divergence should result in a single-point convergence of the estimated parameters among the highlighted trajectories (Gabrys et al., 2019). As Figure 10 shows, the highlighted iterations do not converge around a specific point for any of the parameters, indicating successful convergence for both models.

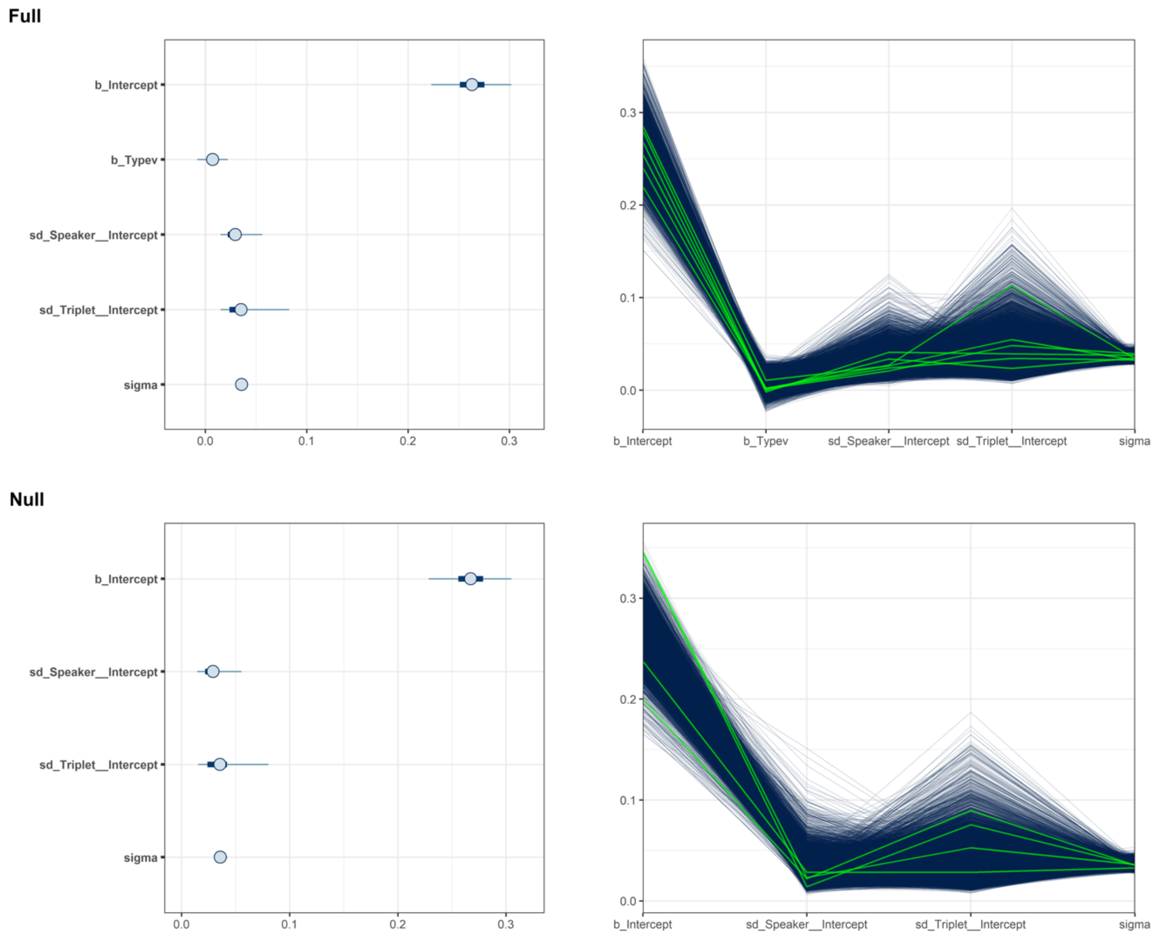

Figure 10

Bayesian analysis results (left) and parallel coordinates plots (right) for all iterations for F2. For the results plots, the thick and thin solid lines represent 50% and 95% of the confidence intervals, and the dots indicate mean values. For the parallel coordinates plots, each line corresponds to one iteration during model construction.

The effect of onset type in the full model (b_Typev in Figure 10) is located very close to zero (μ = 0.01 [-0.01, 0.02]) while the intercept estimations are very similar between the full and null model (full: μ = 0.26 [0.22, 0.30]; null: μ = 0.27 [0.23, 0.31]). Results show that BF_0 is 413.54, which suggests that given the current data, the null model is 413.54 times more likely than the full model.

3.2 Articulatory Results with EMA Data

3.2.1 Main results from the minimal triplet paradigm

Figure 11 shows mean articulatory movements for all 6 triplets. The left column shows LP patterns, where divergence between the vowel pairs can be clearly seen. The right column shows TTy movements and divergence between the consonant pairs can be seen. The trend for CV synchrony can be observed by comparing the left and right columns. In Figure 11, the vowel onset for t6 is around 0.2 s. Similarly, the consonant onset for t6 is also around 0.2 s when the red and blue lines move away from each other. However, for t3 and t5, the vowel pair divergence in LP seems later than that of the consonant pair in terms of TTy. This is likely caused by the rounded portion of the diphthong in /aʊ/ and /oʊ/ in the first words, as the inertia of the rounding gesture at the end of the diphthong is transferred across the syllable boundaries (Xu & Prom-on, 2019). The inertia effect might delay the LP movement for /li/. Yet, the amount of delay is highly variable between speakers and repetitions, as reflected by the error bands, which are slightly wider for both triplet 3 and 5 than for other triplets. Note, however, the continuous LP movement would be more problematic if the threshold method were used, as the onset would be located in the diphthong of the first word, as demonstrated in the following section. For the present study, the variable inertia effect should be taken care of by both the GAMMs and LMEMs, since repetition, item and speaker were included as random effects.

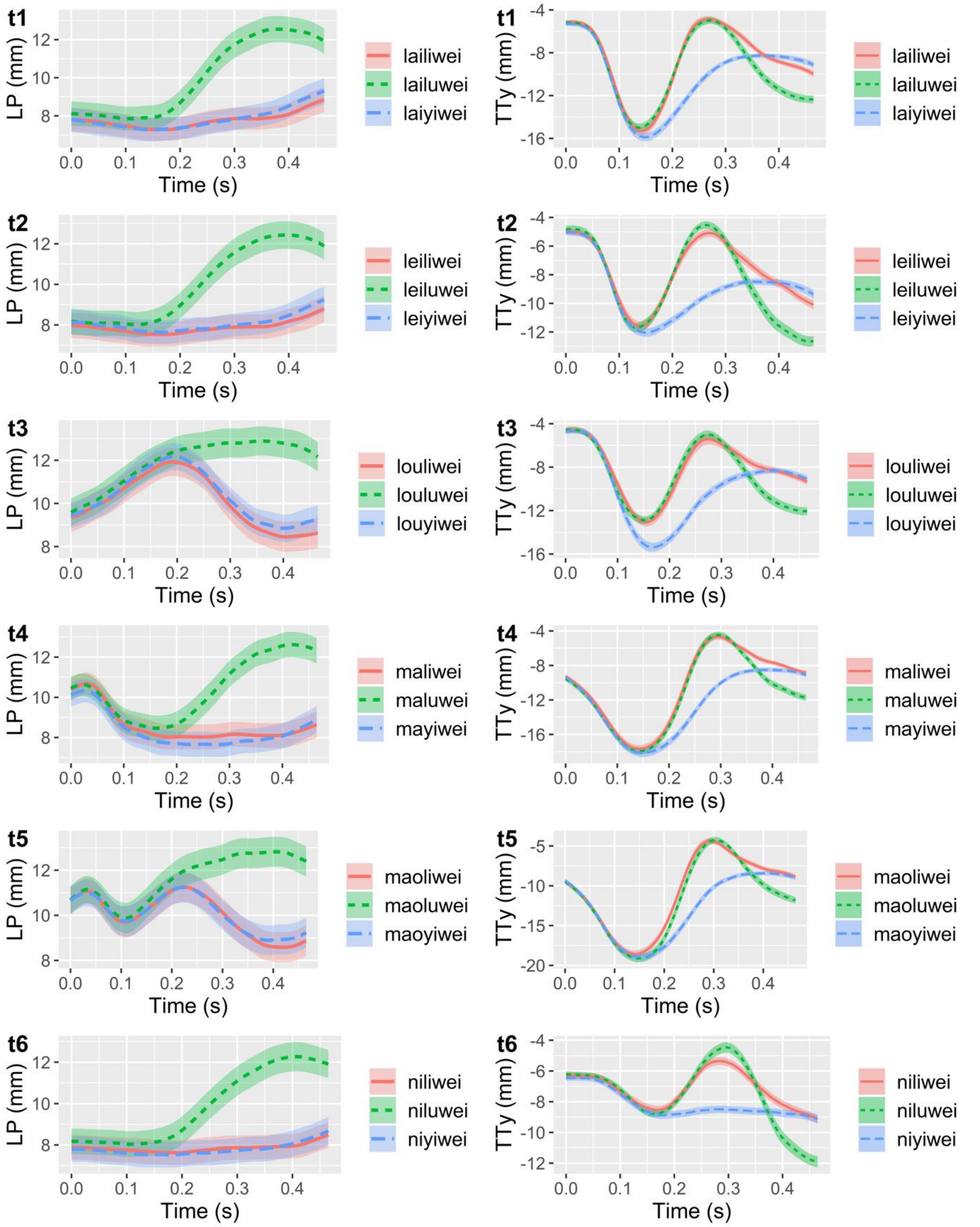

Figure 11

Mean articulatory trajectories averaged across all speakers and repetitions (LP in the first column; TTy in the second column). Articulatory positions are measured in mm. The coloured ribbons indicate standard error of the mean.

Interestingly, the right column shows that for the vowel pairs, a later divergence emerges after the consonant divergence. For example, in the top right graph in Figure 11, TTy starts to differ between 'laili' and 'laiyi' around 0.15 s, while remaining identical between 'laili' and 'lailu' until around 0.3 s, when it finally starts to diverge between the vowels, presumably due to the shared /l/. The later divergence between 'laili' and 'lailu' should be due to the vowel contrast. Thus, this temporal interval between 0.15 s and 0.3 s is likely the duration of /l/, which is shared by the vowel pair. The patterns in Figure 11 thus demonstrate that despite global synchrony, at the level of single articulatory dimensions, articulation is sequential (to be further discussed in section 3.3).

Figure 12 shows mean tongue tip raising (for C) and LP (for V) onset times collected from the GAMMs for /li/, averaged across speakers. Overall, articulatory onsets are earlier than acoustic onsets, possibly due to later onsets of other articulatory gestures resulting in later significant acoustic effects. A likelihood ratios test was performed to compare LMEMs with and without onset type (C vs. V) as the fixed effects. The results indicate that onset time was not significantly affected by onset type (t = 0.69 (LMEM output); $X^2(1)$ = 0.47; p = 0.49). Therefore, the LMEM results support the synchrony patterns shown in Figure 11, namely, articulatorily speaking, the effect of segment type (C or V) on onset time is not significant.

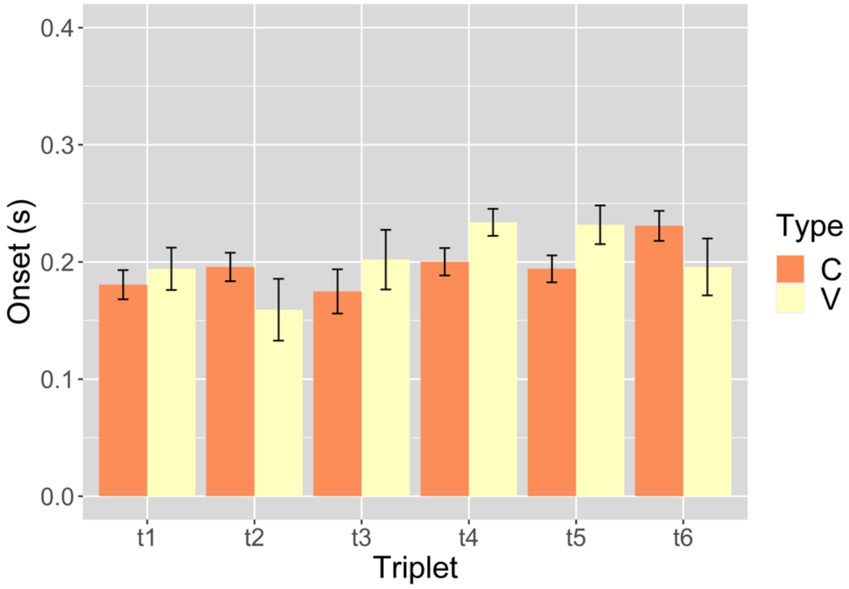

Figure 12

Articulatory CV onset time by segment type and triplet determined by GAMMs. The error bars show standard error of the mean.

Bayesian analysis results and model diagnostics are shown in Figure 13. Similar to results from F2, model convergence was achieved for both the full and null models, as the highlighted trajectories do not converge to a single point. The effect of onset type in the full model is also centred around zero (μ = 0.01 [-0.01, 0.02]) and the intercepts are very similar between the two models (full: μ = 0.19 [0.16, 0.23]; null: μ = 0.20 [0.17, 0.23]). BF_0 is 428.09 which indicates very strong support for the null model by the data.

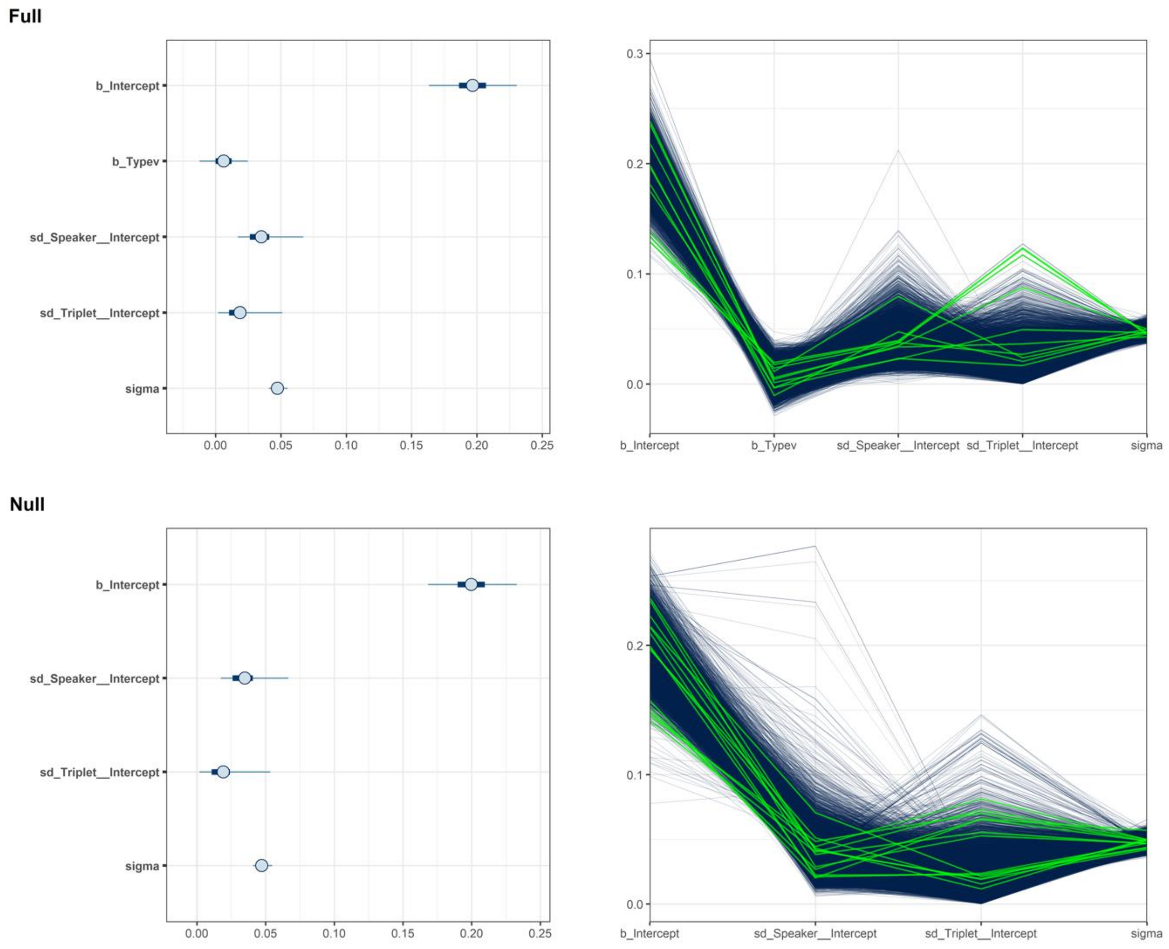

Figure 13

Bayesian analysis results (left) and parallel coordinates plots (rights) for all iterations for the EMA data. For the results plots, the thick and thin solid lines represent 50% and 95% of the confidence intervals, and the dots indicate mean values. For the parallel coordinates plots, each line corresponds to one iteration during model construction.

3.2.2 Method comparison

The onsets of /u/ in ‘maolu’ determined by both methods are potted in Figure 14 for comparison. As can be seen, the velocity method (with 20% as the threshold) would have located the onset of /u/ to be very early at 0.1 s, where the velocity profiles are identical between the minimal pair, likely due to the rounding gesture at the later part of /maʊ/. In contrast, the confounding gesture from the previous syllable is controlled for by the minimal pair method, as the onset is determined to be when the two velocity trajectories move away from each other at around 0.23 s, as shown in Figure 12 for the vowel onset in t5. Note that the standard error of the mean is lower for the velocity method due to its larger sample size.

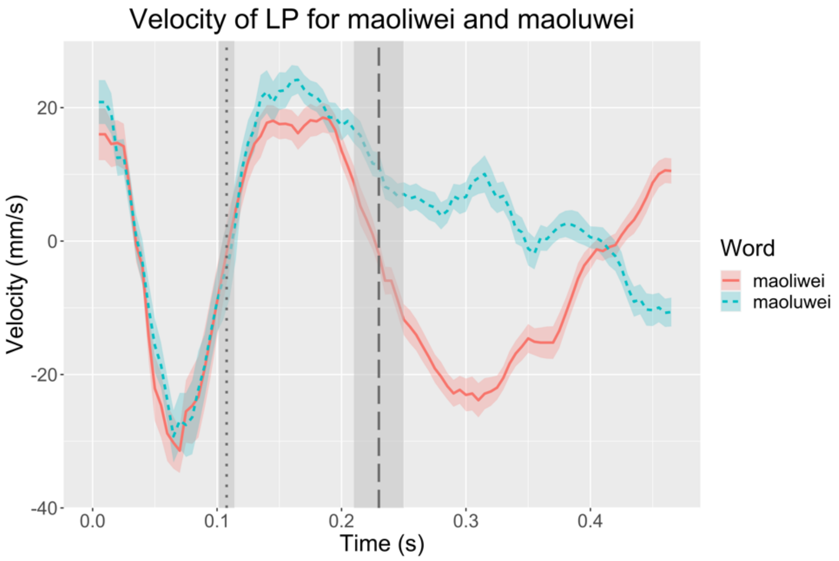

Figure 14

Mean velocity over time for the vowel pair in t5 (‘maolu’ /maʊlu/ vs. ‘maoli’ / maʊli/). The mean onsets determined by the velocity and the minimal contrast method are marked by the dotted and dashed black lines respectively. The shaded ribbons represent standard error of the mean.

However, as mentioned in the introduction, studies will likely avoid stimuli design with obvious gestural confounds such as those in Figure 14. The more misleading gestural confounds are those that are difficult to predict. An example is shown in Figure 15. Here the threshold method would have determined the onset of /u/ to be around 0.16 s. Yet, the velocity profile between ‘laili’ and ‘lailu’ are still similar at that point (Note that the shaded ribbons in the figure represent standard error of the mean which do not show the full extent of the variance. the distributions of raw velocity profiles are heavily overlapped between ‘laili’ and ‘lailu’ at the point of the dotted vertical line). It is not until slightly later the two diverge, which is where the current method located the onset to be. In this case, there are no predictably similar LP gestural specifications for the previous syllable, therefore, covert confounds can only be avoided by using a minimal contrast as reference.

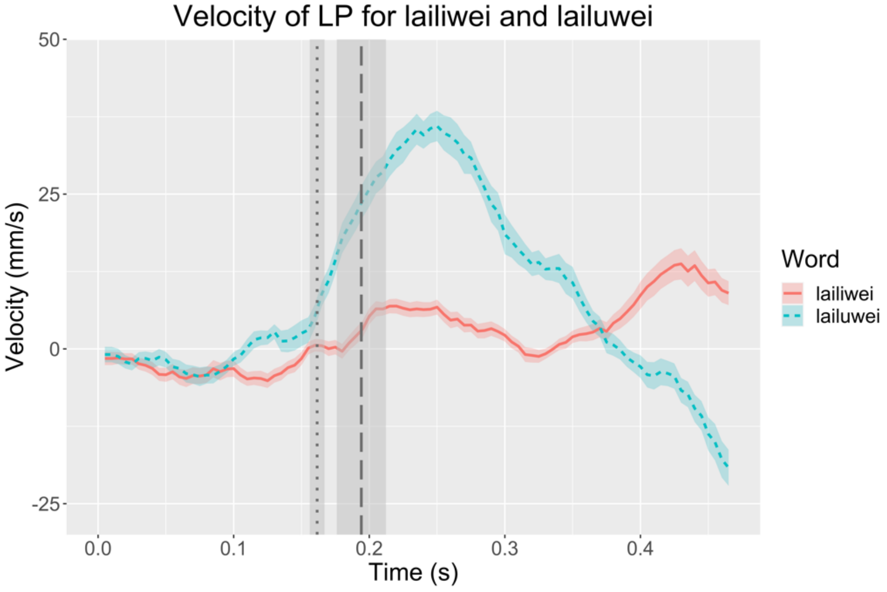

Figure 15

Mean velocity over time for the vowel pair in t1 (‘laili’ /laɪli/ vs. ‘lailu’ /laɪlu/). The mean onsets determined by the velocity and the minimal contrast method are marked by the dotted and dashed black lines respectively. The shaded ribbons represent standard error of the mean.

3.2.3 Control for spatial variation

It is possible that articulatory variability varies systemically between consonant and vowel. Similar to the potential problem of systemic duration variation, if vowels are consistently articulated with more variability than consonants, onsets determined through GAMMs might be systemically delayed for vowels. To assess whether between repetition variability is consistently higher or lower for C or V, we collected all the standard errors of the effect of word contrast from the GAMMs. Due to the inclusion of repetition as a random smooth (i.e., full random effect), the standard error of the effect of word contrast reflects how certain the model is about the effect while taking repetition variability into account (Wieling, 2018). As Figure 16 shows, the two distributions do not differ much between the vowel and consonant pairs in the current analysis. In addition, if /l/ or /i/ were articulated with high variability, a consistent difference between CV onsets would have been reflected in Figure 12. Therefore, the onset times collected from the GAMMs are not likely confounded by systemic spatial variability.

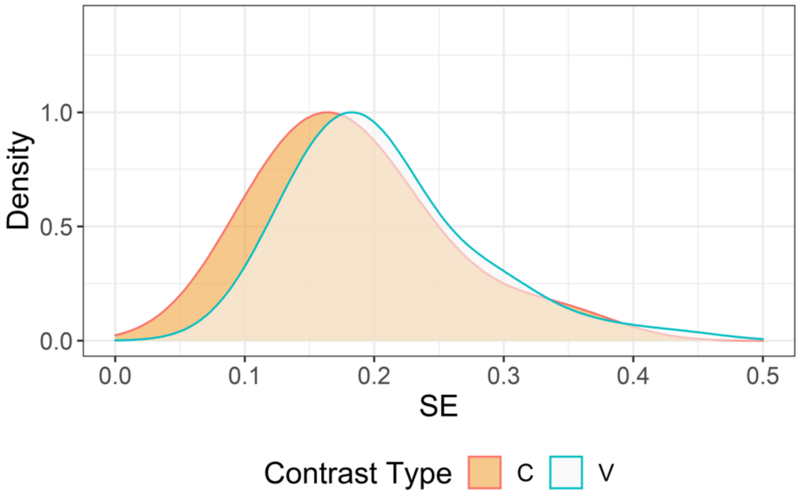

Figure 16
Probability density plot of the standard errors (SE) of the effect of word contrast from all the GAMMs.

3.3 Dimensional Differences in Movements of the Same Articulator

As already seen in Figure 11, it is not the case that CV synchrony occurred in all the articulatory measurements, as vowel articulation seems absent from the TTy dimension during /l/. This means that the vowel is sequentially articulated after the consonant for the TTy dimension. There is therefore a likely general tendency that the greater the conflict between C and V for a particular articulatory dimension, the later the V onset can be detected for that dimension. To test this possibility, we used the GAMM-detected acoustic onsets as the reference point to compare articulatory onsets between different articulatory dimensions for the syllable /li/. Movement onset towards the contrasting targets were determined by additional GAMMs for other EMA measurements for the vowel and consonant pairs separately. Temporal lags were calculated by subtracting articulatory onset time from acoustic onset time for each EMA measurement, and for the consonant and vowel pair separately. For example, the articulatory to acoustic lag for TTy for the consonant would be F2 C onset – TTy C onset. A larger lag value thus means an earlier articulatory onset compared to the acoustic onset for a specific articulatory dimension. Results from the consonant pairs are shown in Figure 17a, and the vowel pairs in Figure 17b. For example, Figure 17a suggests that in reference to the F2 divergence point between the consonant minimal pair (/li/ vs. /yi/), articulatory divergence is detected the earliest for TTy and TDy, since they have the greatest lags. This is consistent with the finding that TTy is the primary articulatory dimension for /l/ (Recasens & Espinosa, 2009) and TDy for /j/ (Recasens, 1990). Parallel to patterns shown in Figure 11, the negative tongue tip lags in Figure 17b indicate that the tongue tip moved towards the vowel targets *after* the acoustic onset of the vowel, in order to first fulfil the gestural requirement for /l/. More importantly, the TTx dimension moves towards the V target earlier than TTy. This confirms Recasens & Espinosa's (2009) finding that the tongue tip in the vertical dimension shows greater coarticulation resistance with the vowel than the horizontal dimension. In other words, TTy is more crucial for the articulation of /l/ than TTx, or any other EMA measurements in the current study.

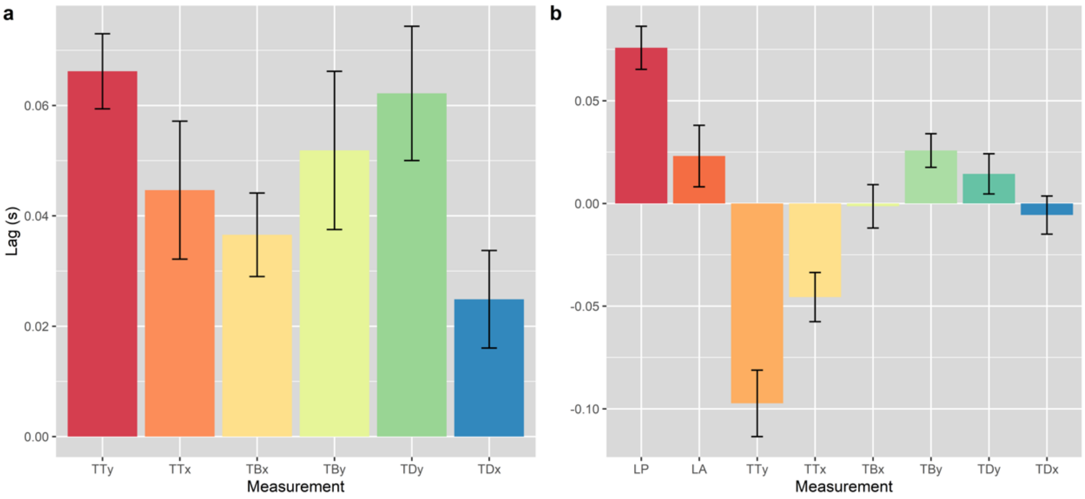

Figure 17

Mean articulation to acoustic lag for EMA measurements. Figure (a) shows temporal lags for the consonant and (b) for the vowel. LA – lip aperture; LP – lip protrusion; TT – tongue tip; TB – tongue blade; TD – tongue dorsum; x – front/back dimension; y – up/down dimension.

The analysis here seems to offer an important tip on a common struggle in EMA studies, i.e., the choice of measurement among the many different articulators and their respective dimensions. It is critical to determine which articulator and dimension is the most pertinent for the segment of interest. For the present study, based on previous reports on coarticulation resistance and other related studies (Recasens & Espinosa, 2009; Gelfer et al., 1989), LP was chosen to assess the contrast between /i/ and /u/, and TTy was chosen for /l/ and /j/. In addition, the analysis above also shows an alternative way to assess the relevance of each EMA measurement for different segments. By using acoustic onset as the reference, the main articulator for a given segment would move towards the associated target first. For the vowel pair, Figure 17b shows that LP is the most useful parameter for detecting the contrast between /i/ and /u/. Similarly, TTy is shown to be the likely primary dimension for the articulation of /l/.

4.0 Discussion

We have applied the minimal triplet paradigm (Xu & Gao, 2018) on articulatory as well as acoustic data to investigate the temporal alignment of consonants and vowels in Mandarin. Triplets of disyllabic words consisting of two matching contrastive pairs in a $C_1V_1\#C_2V_2$ structure were used, whereby the consonant pair differed only in C_2 and the vowel pair differed only in V_2. The onset times of consonants and vowels were determined by detecting articulatory and acoustic (F2) divergence points in the consonant pairs and vowel pairs, respectively, using GAMMs. The onset times of C and V were then compared with LMEMs to determine whether they differed significantly from each other. In addition, Bayesian analysis was used to determine whether the data supports the synchrony view (null model) or the asynchrony hypothesis (full model). Results from both articulatory and acoustical analyses showed no significant difference in onset time between the consonants and vowels, as well as robust support for the synchrony hypothesis. More specifically, in the time course of continuous articulatory movements in all triplets, when LP started to differ between the contrastive vowel pair, TTy also started to differ between the consonant pair. Meanwhile, F2 trajectories in the consonant pair started to diverge around the same time as it did in the vowel pair. This is despite the overall temporal delay relative to the articulatorily-determined onsets. Therefore, contrary to previous findings (Shaw & Chen, 2019; Gao, 2009), the current results provide clear support for the synchrony hypothesis for CV syllables in Mandarin Chinese.

The study also demonstrates the effectiveness of the minimal triplet paradigm. The method can be adopted to investigate segmental timing across difference segmental combinations and languages. For example, Liu & Xu (2021) used the minimal triplet design to examine consonant and vowel onsets in CV and CCV syllables in British English. Minimal triplets (e.g., ‘plit’ vs. ‘plot’ vs. ‘clot’) were embedded in a carrier phrase ‘see a _ today’ and overall differences were tracked in Mel-frequency cepstral coefficients over time between the minimal pairs. The results not only support synchrony between CV but also for the CCV syllables.

4.1 Determination of CV Synchrony

The early findings of co-production of C and V which led to the proposal of the term coarticulation (*Koartikulation* in German) (Menzerath & de Lacerda, 1933) and the articulatory syllable hypothesis (Kozhevnikov & Chistovich, 1965) did contemplate the idea of full CV synchronisation. Indeed, the methods available for examining articulation at the time were limited to visual observation, variations of palatography, kymography, oscillography or photography. Those methods had limited temporal and spatial resolution, and had problems with accuracy and quantification (Gósy, 2011; Rousselot, 1897-1901; Kozhevnikov & Chistovich, 1965). Clear evidence of CV synchrony would have been hard to find even if there had been an effort to look for it.

There have been theoretical proposals of CV synchrony (Goldstein et al., 2006; Xu & Liu, 2006). Goldstein et al. (2006) expanded the articulatory phonology framework to incorporate the idea of CV synchrony into a model of syllable structure in terms of phasing relationship. In this model, consonantal and vocalic gestures are aligned according to their planning oscillators. When the oscillators are coupled in-phase, such as at the beginning of a CV syllable, gestural onsets are synchronous. Nam et al. (2009) cited the finding of Löfqvist & Gracco (1999) that the onset of lip movement for /p/ or /b/ occur within 50 ms of the onset of tongue body movement for the vowel. But Tilsen (2020) claims that there has been a revision of AP in this regard, and the newly accepted generalisation is that the vowel gesture starts somewhere *after* the onset of the consonant closure gesture but *before* the release gesture. As evidence, he cited Nam’s (2007b) observation of the X-ray microbeam data in Browman & Goldstein (1995), which use a velocity threshold to determine movement onset time. Likewise, more recent EMA studies investigating CV timing using the threshold method all reported V onset lagging behind C onset (Gao, 2009; Shaw & Chen, 2019; Yi & Tilsen, 2016). Only one study has reported CV synchrony in support of the in-phase coupling relationship between CV, which used 0 velocity as the segmentation criteria rather than a percentage threshold (Mücke et al., 2012).

As mentioned in the introduction, it is difficult to determine gestural onsets using the velocity threshold method because it is hard to eliminate confounding factors such as adjacent or concurrent gestures and intrinsic difference in stiffness between gestures. Such confounds can be more effectively controlled by the use of minimal pairs for determining the temporal scope of individual gestures (Boyce et al., 1990; Gelfer et al., 1989). To further determine the relative alignment of consonants and vowels, however, two minimal pairs are needed, one for determining the C onset and the other for determining the V onset. Additionally, the C minimal pair and the V minimal pair need to be similar enough to each other to make the estimated C and V onsets comparable. Such a *double minimal pair* method was first proposed in Xu (2007). In Xu & Gao (2018) the method is simplified into a minimal triplet paradigm, which is adopted in the present study.

A further methodological issue is how to statistically determine when the formant or articulatory movements begin. One method is to use running t-test to find out the earliest time at which two mean trajectories becomes significantly different. This has been proposed for analysing articulatory data (Gelfer et al., 1989) and was used for f_0 trajectories (Xu et al., 2004). However, running t-tests have the disadvantage of increasing the possibility of type I error. For example, while comparing trajectories with 10 sampling points, 10 t-tests need to be conducted throughout the sampling interval, and the number of tests would increase with the size of the dataset and the duration of the tokens. By using GAMM, the number of statistical models is limited to one for each minimal pair while also accounting for repetition variation. Furthermore, only a significant divergence that lasts for longer or equal to 40 ms was determined as a valid onset. Therefore, the possibility of type I error can be reduced. Another pivotal part of the statistical method in the current study is to construct GAMMs for separate speakers, rather than modelling with the pooled data. This ensures that the sample size of the final dataset (i.e., CV onset times) is sufficient for further statistical testing.

4.2 CV Synchrony and Coarticulation

Note that the synchronised onsets of C and V determined in the present study, as described in the previous section, are temporally well ahead of the conventional syllable onset, namely, the acoustic onset of consonants, including stop closure, frication or nasal or lateral murmur, etc. (Lehiste & Peterson, 1961; Turk et al., 2006). For the vowels, this is a leftward shift of roughly half of a syllable from the conventional vowel onset. Take the utterance in Figure 4 as an example, the rise of F2 near the end of the /ma/ interval would correspond to what is reported by Öhman (1966) as the anticipatory coarticulation with /i/ in /li/. Given the present finding that the consonant also starts its articulation at the same time, as can be seen in Figure 8, F2 in /li/ starts to deviate from /ji/ also from that time point. Hence, /i/ in /li/ does not start before /l/, and so there is no cross-consonant vowel to vowel coarticulation. Note that this is exactly what is proposed by the articulatory syllable hypothesis (Kozhevnikov & Chistovich, 1965) as well as the synchronised coproduction hypothesis (Xu, 2020). The finding of CV synchronisation in the present study therefore suggests that local vowel to vowel anticipatory coarticulation is actually congruent with the view of CV synchronisation.

However, the original definition of coarticulation by Menzerath & de Lacerda (1933) may be considered to be too restrictive based on the now widely accepted definition, namely, “the influence of one speech segment upon another” (Daniloff & Hammarberg, 1973:239). These influences would include all the carryover or anticipatory effects. Again, back to Figure 4 for example, although the F2 rise toward the end of the /ma/ interval is now considered as part of the /i/ articulation based on the new interpretation, it can still be considered as a case of carryover coarticulation. If, however, based on the present finding, V_2 (i.e., /i/) actually starts at the F2 rise in the /ma/ interval, the carryover effect of /a/ on the next syllable is simply inertia. If so, its effect can be accounted for by models that incorporate it as a core mechanism, such as the Fujisaki model (Fujisaki, 1983), task dynamic model (Saltzman & Munhall, 1989) and target approximation model (Xu & Wang, 2001). In the target approximation model, for example, each articulatory movement is a process of approaching an underlying target, as illustrated in Figure 18. Each articulatory movement is therefore necessarily a process of continually departing from the initial state left by the prior articulation. In such a process, there is no need to additionally model the influence of the preceding articulation in the form of overlap with the current articulation.

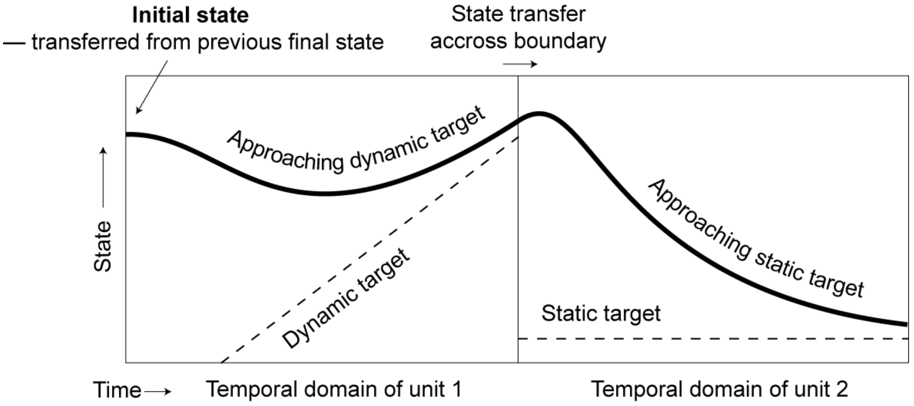
Initial state
— transferred from previous final state
State transfer
accross boundary
Approaching dynamic target
Approaching static target
Dynamic target
Static target
State
Time
Temporal domain of unit 1
Temporal domain of unit 2

Figure 18

The target approximation model. A schematic illustration of hypothetical phonetic targets (dashed lines) and their surface realisation (solid curve). The three vertical lines represent the boundaries of the two consecutive target intervals. The level dashed line on the right represents a static target, and the oblique dashed line on the left represents a dynamic target. Adapted from the original version for tone and intonation (Xu & Wang, 2001).

4.3 Dimension-Specific Sequential Target Approximation and Coarticulation

The question about carryover coarticulation raised above is also critical for another core issue of coarticulation, namely, is it possible for two articulatory movements or gestures involving the same articulator to be overlapped during coarticulation? Such overlap is allowed in AP/TD in the form of blending (Saltzman & Munhall, 1989). For example, the initial portion of the F2 rise in Figure 4 could potentially involve the blending of tongue body gestures for /a/, /l/ and /i/ in various proportions (in terms of /mali/). But there are also views against such blending. Wood (1996:139) claims that "potentially conflicting gestures are not blended but are produced sequentially." However, there are many cases, including /li/, where the same articulator is needed by both the consonant and the vowel. The present data as presented in section 3.3 show that the conflict can be resolved by allowing different dimensions of the same articulator to first start their movements toward either the consonant or the vowel target, respectively. That is, if an articulatory dimension is critical for the consonant (hence its primary dimension), e.g., TTy for /l/, it can approach the consonant and the vowel targets in succession. At the same time, other dimensions of the same articulator, e.g., TTx for /l/, can approach the vowel target simultaneously with the movement of all the other vowel-relevant articulatory dimensions. In this way, as far as any *specific articulatory dimension* is concerned, its target approximation movements are always sequential: approaching one target at a time. This *dimension-specific sequential target approximation* (Xu et al., 2019; Xu, 2020) can therefore be the strategy that resolves many of the C-V conflicts to make temporally synchronised CV coproduction possible.

Some might argue that the movement patterns for TTy in Figure 11 could be due to the lack of vowel specification for TTy. Although the tongue body is believed to be the main articulatory for vowels, articulatory studies do show that TTx and TTy vary systemically between vowels under different consonant context (Recasens & Espinosa, 2009; Recasens & Rodríguez, 2016). MRI data of vowel production also shows that tongue tip constriction varies between vowels, especially between front and back vowels (i.e., /i/ vs. /u/), a significant negative correlation was also identified between tongue tip constriction and F2 (Zourmand et al., 2014). Zourmand et al. (2014)'s findings correspond to that of Mac Neilage & Sholes (1964), in which Electromyography was used and strong muscle activity was found for the TT for /i/ and /u/. Therefore, it is likely that the tongue tip first fulfilled the consonantal requirement of /l/ then went on to articulate the vowel for the reminder of the syllable in Figure 11.

Dimension-specific sequential target approximation also provides an explanation for coarticulation resistance, the well-known phenomenon that the degree at which a segment can resist coarticulatory effects from adjacent segments depends on the amount of lingual requirement for its own articulation (Bladon & Al-Bamerni, 1967; Recasens, 1984; Recasens, 1987; Recasens, 1989). /ʃ/, for example, is more resistant than /p/ or /t/ to coarticulatory effects from vowels, because its articulation requires the tongue body for its production, while the labial and dentoalveolar stops do not (Recasens, 2018). Based on dimension-specific sequential target approximation, any articulator dimension primarily specified by a consonant needs to approach the consonant target first, before the execution of the vowel target. But dimensions that are not primarily specified by the consonant, even if it belongs to the same articulator, can approach the vowel target from the syllable onset. For instance, for the syllable /gV/, the tongue body can move both upwards to make the required velar contact for /g/, and horizontally toward the desired tongue position for the vowel, resulting in a well-documented context dependent contact point along the soft palate (Recasens & Espinosa, 2009). This articulation pattern has been successfully simulated using dimension-specific sequential target approximation in an articulatory modelling study (Xu et al., 2019). Therefore, approximations of the vowel and consonant targets can simultaneously take place on separate spatial dimensions. Given this mechanism, the more articulatory dimensions are obligated by a segment, the more sequential the segment need to be relative to the neighboring segments, and the less it would show coarticulatory influences from its neighbors.

Given dimension-specific sequential target approximation, some segments may even be largely insusceptible to coarticulation. Glides like /j/, /w/, etc., in particular, are specified for almost the whole shape of the vocal tract (i.e., the entire tongue shape in both the vertical and horizontal dimensions) because they are semivowels in nature (Ladefoged & Maddieson, 1996), thus engaging almost all the articulatory dimensions (Recasens & Espinosa, 2009). They may therefore have to be produced only sequentially with adjacent vowels, as demonstrated by articulatory synthesis (Prom-on et al., 2014). For the same reason, adjacent vowels can only be produced sequentially. This is consistent with the hypothesis that speech articulation consists of a relatively slow flow of sequential (i.e., diphthongal, in Öhman's (1966) term) vowel movements with a separate flow of intermittent consonants superimposed on its transitional portion (Fowler, 1977; Öhman, 1966).

4.4 Coarticulation and the Syllable

The discussion so far has demonstrated that the synchronised onset of C and V at the start of the syllable is most consistent with the articulatory syllable hypothesis (Kozhevnikov & Chistovich, 1965). The evidence found in the present study has provided an alternative view on anticipatory vowel to vowel influence (Öhman, 1966), by demonstrating that it is actually part of syllable-internal coproduction. This coproduction pattern is also consistent with the strong version of articulatory phonology that assumes full synchrony of CV onsets (Goldstein et al 2006; Nam et al., 2009). As discussed in section 4.1, the observed lagging of V onset in some of the articulatory studies (Gao, 2009; Shaw & Chen, 2019; Yi & Tilsen, 2016) that led to the revised version of articulatory phonology might be due to the method of using velocity threshold to determine movement onsets. The only previous direct evidence for CV synchrony might have avoided the stiffness confound by using 0 velocity as the segmentation point (Mücke et al., 2012). We have therefore seen an emergent consensus on the relation between coarticulation and the syllable (Goldstein et al., 2006; Kozhevnikov & Chistovich, 1965; Xu, 2020; present data), that is, a large part of the observed coarticulation might be the result of synchronised co-onset of C and V at the beginning of the syllable. The only major disagreement is whether CV synchrony is the result of a planning process based on coupled oscillation that happens before the execution of each syllable according to AP/TD, or it is achieved by all the involved articulators simply starting at the same time without such an oscillation mechanism (Xu, 2020). The latter hypothesis is based on the need to minimise degrees of freedom in motor control (Bernstein, 1967). Synchronisation of C and V as well as tone and phonation could be an effective mechanism of solving this problem by eliminating most of the temporal degrees of freedom in speech motor control (Xu, 2020). The exact mechanism of CV coarticulation, however, can be solved only in future research.

4.5 Acoustic versus articulatory measurements

There have often been concerns that acoustic measurements may not provide relevant articulatory information due to a lack of one-to-one relations between articulation and acoustics. But it has also been argued that acoustic measurements such as continuous formant trajectories provide information about the underlying articulatory movements no less relevant than direct articulatory measurements (Cheng & Xu, 2013). The time-continuous acoustic and articulatory measurements of the same speech data in the present study have made it possible to make direct comparisons between the two kinds of measurements. As discussed in 3.3, it is often tricky to decide which of the many measurable articulators would provide the most relevant information for answering a particular research question. As explained in Cheng & Xu (2013), this is because, to generate sufficient audible effects for linguistic contrasts, the articulators need to work together to produce acoustic patterns that can be tracked by measurements like formant trajectories. But only the first few formants can be effectively manipulated in articulation according to perturbation theory (Fant, 1960; Stevens, 1998). As shown in Figure 8, F2 alone can provide enough information for determining the time alignment of C and V. In contrast, at least two articulatory measurements, TTy and LP are needed to make the same determination, and the two measurements are among many others that can provide only partial information about CV alignment, as seen in Figure 17. In general, therefore, acoustic measurements can be directly relevant for examining articulation in speech.

4.6 Caveats

The current data are from Mandarin Chinese only. Therefore, they do not indicate anything directly about other languages. However, by using the minimal triplet design from the current study, new evidence for synchrony has surfaced in British English for both CV and cluster syllables (Liu & Xu, 2021). Future studies using the minimal contrast design will also have to consider potential variability confounds such as speech rate and articulatory variation. These variational confounds can be measured and analysed as outlined in sections 2.5 and 3.2.3.

Also, the present study addresses the phenomenon of *local* anticipatory coarticulation as shown in Figure 1b (Öhman, 1966; Rubertus & Noiray, 2018; Recasens & Pallarès, 2001; Mok, 2012), and not *long distance* anticipatory coarticulation reported in Grosvald (2009) and Magen (1997), vowel harmony in languages such as French (Chiu et al., 2015), or the type of preparatory effects reported in Tilsen (2020). However, long distance vowel to vowel coarticulation can also be tested with the minimal triplet method, which might provide valuable insights into how much earlier the vocalic activity start before the consonantal onset.

Arguably one of the disadvantages of the current method is that it cannot be used to determine the onset of individual movements on their own, since it requires at least two contrasting phonetic groups to determine a common onset point. But given that the underlying articulatory goal/target of a phonetic unit is always hidden (Saltzman & Munhall, 1989), and its articulatory movements and acoustic patterns vary extensively with surrounding units, its temporal alignment simply cannot be directly determined from its own surface form alone.

Furthermore, a recent study suggests that CV alignment in Mandarin might be subject to stress variations. Zhang et al. (2019) used gestural plateau to measure CV lags and found that CV lag is significantly greater for syllables with a full tone compared to tonesless syllables. However, /li/ is not a toneless syllable in the current study, and any potential stress induced effect on CV alignment can be tested with the minimal contrast paradigm in future studies.

Finally, the current findings do not refute, nor do they confirm the preparation effects reported by Tilsen (2020), in which consonant related preparation effects were observed to go beyond gestural initiation (determined by the threshold method). The methodology of eliciting speech in the study is very different from that of the present study. In the current design, there is no preparation stage (which is as long as 2000 ms) when the participant can see the stimuli but cannot start the articulation of the target syllable before the go cue. It is therefore hard for us to compare his finding with the present results.

5.0 Conclusion

The present study offers an alternative segmentation method that can better control for various gestural and articulatory confounds. With the proposed method, we tested the hypothesis that consonant and vowel start at the same time at syllable onset (Goldstein et al 2006; Xu & Liu, 2006; Xu, 2020). Methodologically, we extended the minimal triplet paradigm first applied to acoustic data (Xu & Gao, 2018) to both articulatory and acoustic data. In this paradigm, the onset of consonants or vowels is identified as the divergence point between the articulatory or F2 trajectories of a contrasting pair of segments. The temporal alignment of consonant and vowels are then determined by comparing the onset times of minimally contrasting C and V pairs, respectively. The results of analysing EMA and F2 data showed clear evidence of CV synchrony, although the acoustically determined onset lag behind the articulatorily determined onset by a largely constant amount of time. Detailed analysis of the articulatory data further demonstrates that the realisation of CV synchrony is based on dimension-specific sequential target approximation. That is, for any particular articulator, only the dimensions not essential to the consonant started their movements toward the vowel target from the syllable onset, while the consonant-essential dimensions had to move toward the vocalic target only after the termination of consonant approximation.

The newly determined co-onset points are much earlier than the conventional landmark-defined onsets, i.e., start of acoustic closure for consonants as well as syllables. This means that most of the previously reported anticipatory coarticulation (Öhman, 1966), happens within the syllable, and is in fact part of the formation of the syllable. This is not a fully novel idea, as it is consistent with the articulatory syllable hypothesis (Kozhevnikov & Chistovich, 1965). But the findings of the present study have answered many of the uncertainties followed the initial proposal of the articulatory syllable hypothesis (Kühnert & Nolan, 1999). Also, the new evidence for dimension-specific sequential articulation addressed the conundrum of coarticulation resistance (Bladon & Al-Bamerni, 1967; Recasens, 1984, 1987, 1989). The present findings have therefore presented an alternative perspective on some of the most critical issues about coarticulation.

The new perspective has raised important questions about the range and type of coarticulation. First, in the strict sense of coproduction, coarticulation can happen at the level of segments or articulators, but likely not at the level of single articulatory dimensions (e.g., TTy for the vowel pairs in the current study). Second, most of the previously reported local anticipatory coarticulation is likely no longer there, given that articulatory movements toward C and V targets both start at syllable onset, but not before it (except for the possible postural preparation reported by Tilsen, 2020). Third, based on both the task dynamic model and the target approximation model, the articulation of a segment is in essence a process of moving away from the final articulatory state of the preceding segment, following the basic law of physics. Under this view, the articulation of the preceding segment therefore does not need to overlap with that of the following segment, which raises questions about the definition of carryover coarticulation. Finally, the current study is limited to data from CV syllables in Mandarin Chinese only. The method proposed here however, is applicable to a wide range of languages and segment combinations, as has already been tested in Liu & Xu (2021).

Reference List

Bates, D., Maechler, M, Bolker, B. & Walker, S. (2019). lme4: Linear Mixed-Effects Models using 'Eigen' and S4. R Package version 1.1-21.

Bell-Berti, F. & Harris, K. S. (1981). A temporal model of speech production, *Phonetica,* 38**,** 9-20.

Bell-Berti, F., & Harris, K. S. (1979). Anticipatory coarticulation: some implications from a study of lip rounding. J Acoust Soc Am, 65(5), 1268-1270.

Bernstein, N. A., 1967. The co-ordination and regulation of movements. Pergamon Press, Oxford.

Bladon, R. A. W., & Al-Bamerni, A. (1976). "Coarticulation resistance of English /l/," *J. Phonetics* **4**, 135-150.

Blair, C. (1988). Firing & Contractile Properties of Human Lower Lip Motor Units during Sustained Isometric Contractions. *Experimental Neurology, 99*, 269-280.

Boersma, Paul & Weenink, David (2019). Praat: doing phonetics by computer [Computer program]. Version 6.1.08.

Boyce, S. E., Krakow, R. A., Bell-Berti, F., & Gelfer, C. E. (1990). Converging sources of evidence for dissecting articulatory movements into core gestures. *Journal of Phonetics, 18*, 173-188. doi:10.1016/S0095-4470(19)30400-0

Browman, C. P., & Goldstein, L. (1995). Dynamics and articulatory phonology. In R. F. Port & T. van Gelder (Eds.), Mind as motion: Explorations in the dynamics of cognition (p. 175–193). The MIT Press.

Bürkner, P. C, Gabry, J., Weber, S., Johnson, A., & Modrak, M. (2021). Brms: Bayesian Regression Models Using 'Stan'. R package version 2.15.0.

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

Chiu, F., Fromont, L., Lee, A. & Xu, Y. (2015). Long-distance anticipatory vowel-to-vowel assimilatory effects in French and Japanese. In *proceedings of the 2015 International Congress of Phonetic Sciences.* Glasgow, UK.

Daniloff, R. G., & Hammarberg, R. E. (1973). On defining coarticulation. *Journal of Phonetics 1*, 239-248.

Danner, S. G., Barbosa, A. V., & Goldstein, L. (2018). Quantitative Analysis of Multimodal Speech Data. *Journal of Phonetics*, 71, 268-283. doi:10.1016/j.wocn

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. Frontiers in Psychology, 5(July), 1–17. https://doi.org/10.3389/fpsyg.2014.00781

Dienes, Z. (2016). How Bayes factors change scientific practice. Journal of Mathematical Psychology, 72, 78–89. https://doi.org/10.1016/j.jmp.2015.10.003

Fant, G. (1960). *Acoustic theory of speech production*. The Hague.

Fowler, C. A. (1977) *Timing control in speech production*. Indiana University Linguistics Club.

Fujisaki, H. (1983), Dynamic characteristics of voice fundamental frequency in speech and singing. *The production of Speech*, 39-55

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., & Gelman, A. (2019). Visualization in Bayesian workflow. *Journal of the Royal Statistical Society*. Series A: Statistics in Society, 182(2), 389–402. https://doi.org/10.1111/rssa.12378

Gao, M. (2009). Gestural coordination among vowel, consonant and tone gestures in Mandarin Chinese. *Chin. J. Phonet, 2*, 43-50.

Gelfer, C. E., Bell-Berti, F., &Harris, K. S. (1989). Determining the extent of coarticulation: effects of experimental design. *Journal of the Acoustical Society of America,* Acoustical Society of America 86(6), 2443-2445.

Goldstein, L., Byrd, D., & Saltzman, E. (2006). The role of vocal tract gestural action units in understanding the evolution of phonology. *Action to language via the mirror neuron system***,** 215-249.

Gósy, M. (2011). From stomatoscopy to BEA: The history of Hungarian experimental phonetics. In *proceedings of the International Congress of Phonetic Sciences (2011),* HK, China.

Grosvald, M. (2009). Interspeaker variation in the extent and perception of long-distance vowel-to-vowel coarticulation. *Journal of Phonetics*, 37(2), 173–188. https://doi.org/10.1016/j.wocn.2009.01.002

Gubian, M., Pastätter, M., & Pouplier, M. (2019). Zooming in on spatiotemporal V-to-C coarticulation with functional PCA. In *proceedings of the 2019 Annual Conference of the International Speech Communication Association, INTERSPEECH.* Graz, Austria.

Hardcastle, W., & Hewlett, N. (Eds.). (1999). *Coarticulation: Theory, Data and Techniques*. Cambridge University Press.

Harms, C., & Lakens, D. (2018). Making "null effects" informative: statistical techniques and inferential frameworks. *Journal of Clinical and Translational Research*, 1–24. https://doi.org/10.18053/jctres.03.2017s2.007

Hoole, P., Mooshammer, C., & Tillman, H. G. (1994). Kinematic analysis of vowel production in German. In *proceedings of the 3rd International Conference on Spoken Language Processing*. Yokohama.

Ito, T., Murano, E. Z., & Gomi, H. (2004). Fast force-generation dynamics of human articulatory muscles. *J Appl Physiol, 96*, 2318-2324.

Jeffreys, H. (1961). *The Theory of Probability* (3rd ed). Oxford University Press.

Kent, R. D., & Minifie, F. D. (1977). Coarticulation in recent speech production models. *Journal of Phonetics*, 5(2), 115–13

Kozhevnikov, V. A. & Chistovich, L. A. (1965). *Speech: Articulation and Perception*. Washington, DC: Translation by Joint Publications Research Service. JPRS 30543.

Kühnert, B. & Nolan, F. (1999). The origin of coarticulation. In *Coarticulation: Theory, Data and Techniques*. W. J. Hardcastle and N. Newlett. Cambridge University Press.

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

Lakens, D., McLatchie, N., Isager, P. M., Scheel, A. M., & Dienes, Z. (2020). Improving Inferences about Null Effects with Bayes Factors and Equivalence Tests. *Journals of Gerontology - Series B Psychological Sciences and Social Sciences*, 75(1), 45–57. https://doi.org/10.1093/geronb/gby0653.

Lee, M. & Wagenmakers, E. J. (2013*). Bayesian Cognitive Modelling*. Cambridge University Press.

Lehiste, I., & Peterson, G. E. (1961). Some basic consideration in the analysis of intonation. *J Acoust Soc Am*, 33(419), 419-425. doi:10.1121/1.390399

Liu, Z. & Xu, Y. (2021). Segmental Alignment of English Syllables with Singleton and Cluster Onsets. To appear in *proceedings* of *the 2021 Annual Conference of the International Speech Communication Association, INTERSPEECH.* Brno, Czechia.

Lobanov, B. M. (1971). Classification of Russian Vowels Spoken by Different Speakers. *The Journal of the Acoustical Society of America, 49*(2B), 606-608. doi:10.1121/1.1912396

Löfqvist, A., & Gracco, L. (1999). Interarticulator programming in VCV sequences: Lip and tongue movements. *Journal of the Acoustical Society of America*, *105***,** 1864-1876.

Mac Neilage, P. F. & Sholes, G. N. (1964). AN ELECTROMYOGRAPHIC STUDY OF THE TONGUE DURING VOWEL PRODUCTION. *Journal of Speech, Language, and Hearing Research*, 7, 209–232.

Magen, H. S. (1997). The extent of vowel-to-vowel coarticulation in English. *Journal of Phonetics*, 25, 187-205.

Marin, S. & Pouplier, M. (2008). Organization of complex onsets and codas in American English. In *proceedings of the 8th International Seminar on Speech Production*. Strasbourg, France.

Marin, S., & Pouplier, M. (2014). Articulatory synergies in the temporal organization of liquid clusters in Romanian. *Journal of Phonetics, 42*, 24-36. doi:10.1016/j.wocn.2013.11.001

Menzerath, P. & de Lacerda, A. (1933). *Koartikulation, Seuerung und Lautabgrenzung*. Fred. Dummlers.

Mok, P. P. K. (2012). Effects of consonant cluster syllabification on vowel-to-vowel coarticulation in English. *Speech Communication*, 54(8), 946-956. doi:10.1016/j.specom.2012.04.001

Mücke, D., Nam, H., Hermes, A., & Goldstein, L. (2012). Coupling of tone and constriction gestures in pitch accents. In *Consonant Clusters and Structural Complexity*. https://doi.org/10.1515/9781614510772.205

Nalborczyk, L., Batailler, C., Lœvenbruck, H., Vilain, A., & Bürkner, P.-C. (2019). An Introduction to Bayesian Multilevel Models Using brms: A Case Study of Gender Effects on Vowel Variability in Standard Indonesian. *Journal of Speech, Language, and Hearing Research*, 62(5), 1225–1242. https://doi.org/10.1044/2018_JSLHR-S-18-0006

Nalborczyk, L., Grandchamp, R., Koster, E. H. W., Perrone-Bertolotti, M., & Loevenbruck, H. (2020). Can we decode phonetic features in inner speech using surface electromyography? *PLoS ONE*, 15(5), 1–27. https://doi.org/10.1371/journal.pone.0233282

Nam, H. (2007a). Articulatory modelling of consonant release gesture. In *proceedings of the 16th International Congress of Phonetic Sciences*. Saarbrücken, Germany.

Nam, H. (2007b). Syllable-level intergestural timing model: Split-gesture dynamics focusing on positional asymmetry and moraic structure. In I. J. Cole & J. I. Hualde (Eds.). *Laboratory phonology* (Vol. 9, pp. 483–506). Walter de Gruyter. Nelson, 1983

Nam, H., Goldstein, L., & Saltzman, E. (2009). Self-organization of syllable structure: a coupled oscillator model. In *Approaches to Phonological Complexity*. De Gruyter Mouton.

Nam, H., Mitra, V., Tiede, M., Hasegawa-Johnson, M., Espy-Wilson, C., Saltzman, E., & Goldstein, L. (2012). A procedure for estimating gestural scores from speech acoustics. *J Acoust Soc Am, 132*(6), 3980-3989. doi:10.1121/1.4763545

Öhman, S. E. (1966). Coarticulation in VCV utterances: spectrographic measurements. J Acoust Soc Am, 39(1), 151-168. doi:10.1121/1.1909864

Pastätter, M., & Pouplier, M. (2014). The articulatory modelling of German coronal consonants using TADA. In *proceedings of the 12thInternational Seminar on Speech Production.* Cologne, Germany.

Perkell, J. & Chiang, C. M. (1986). Preliminary support for a 'hybrid' model of anticipatory coarticulation. In *proceedings of the 12th International Congress of Acoustics*. Toronto.

Prom-on, S., Birkholz, P., & Xu, Y. (2014). Identifying underlying articulatory targets of Thai vowels from acoustic data based on an analysis-by-synthesis approach. *EURASIP Journal on Audio, Speech, and Music Processing*, *23*.

Prom-on, S., Xu, Y., & Thipakorn, B. (2009). Modelling tone and intonation in Mandarin and English as a process of target approximation. *The Journal of the Acoustical Society of America*, 125(1), 405–424. https://doi.org/10.1121/1.3037222

Recasens, D. (1984). V-to-C coarticulation in Catalan VCV sequences: an articulatory and acoustical study, *Journal of Phonetics*, 12, 61-73.

Recasens, D. (1987). An acoustic analysis of V-to-C and V-to-V coarticulatory effects in Catalan and Spanish VCV sequences, *Journal of Phonetics*, 15, 299-312.

Recasens, D. (1989). Long range coarticulation effects for tongue dorsum contact in VCVCV sequences, *Speech Communication*, 8, 293-307.

Recasens, D. (1990). The articulatory characteristics of palatal consonants. *Journal of Phonetics, 18*(2), 267-280. doi:10.1016/s0095-4470(19)30393-6

Recasens, D. (2018) Coarticulation. In *Oxford Research Encyclopedia of Linguistics*. https://oxfordre.com/linguistics/view/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-416.

Recasens, D., & Espinosa, A. (2009). An articulatory investigation of lingual coarticulatory resistance and aggressiveness for consonants and vowels in Catalan. *J Acoust Soc Am*, 125(4), 2288-2298. doi:10.1121/1.3089222

Recasens, D., & Pallarès, M. (2001). Coarticulation, assimilation and blending in Catalan consonant clusters. *Journal of Phonetics*, 29, 273-301. doi:10.006/jpho.2001.0139

Recasens, D., & Rodríguez, C. (2016). A study on coarticulation resistance and aggressiveness for front lingual consonants and vowels using ultrasound. Journal of Phonetics, 59, 58–75.

Roon, K. D., Hoole, P., Zeroual, C., Du, S., & Gafos, A. I. (2021). Stiffness and articulatory overlap in Moroccan Arabic consonant clusters. *Laboratory Phonology*, 12(1), 1–23. https://doi.org/10.5334/LABPHON.272

Rousselot, P .-J. (1897-1901). *Principes de phonétique experimentale*, I-II. Paris: H. Welter.

Rubertus, E., & Noiray, A. (2018). On the development of gestural organization: A cross-sectional study of vowel-to-vowel anticipatory coarticulation. *PLOS ONE*, 13(9), e0203562

Saltzman, E. L., & Munhall, K. G. (1989). A Dynamical Approach to Gestural Patterning in Speech Production. *Ecological Psychology, 1*(4), 333-382. doi:10.1207/s15326969eco0104_2

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*, 22(2), 322–339. https://doi.org/10.1037/met0000061

Shaw, J. A., & Chen, W. R. (2019). Spatially Conditioned Speech Timing: Evidence and Implications. *Front Psychol,* 10, 2726. doi:10.3389/fpsyg.2019.02726

Shaw, J. A., Gafos, A. I., Hoole, P., & Zeroual, C. (2011). Dynamic invariance in the phonetic expression of syllable structure: a case study of Moroccan Arabic consonant clusters. Phonology, 28(3), 455-490. doi:10.1017/s0952675711000224

Shih, C. (1995). STUDY OF VOWEL VARIATIONS FOR A MANDARIN SPEECH SYNTHESIZER. *Eurospeech*, 3–6.

Sievers, E. (1876). Grundzüge der Lautphysiologie zur Einführung in das Studium der Lautlebre der Indogermanischen Sprachen. Leipzig: Breitkopf and Härtel.

Smith, J. G. (2010). *Acoustic properties of english /l/ and /ɹ/ produced by Mandarin Chinese speakers* (Master's thesis, University of Toronto: http://individual.utoronto.ca/jgsmith/content/papers/LIN1290Y_jsmith2010_forum_master.pdf).

Stetson, R. (1951). *Motor phonetics: A study of speech movements in action*. Amsterdam: North-Holland (2nd ed.).

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

Stone, J. V. (2013). *Bayes' rule: a tutorial introduction to Bayesian analysis*. Sebtel Press.

Tilsen, S. (2020). Detecting anticipatory information in speech with signal chopping. *Journal of Phonetics*, *82*, 100996.

Turk, A., Nakai, S., & Sugahara, M. (2006). "Acoustic Segment Durations in Prosodic Research: A Practical Guide," in *Methods in Empirical Prosody Research*, edited by S. Sudhoff, D. Lenertová, R. Meyer, S. Pappert, P. Augurzky, I. Mleinek, N. Richter, & J. SchlieBer. De Gruyter.

Van Rij, J., Wieling, M., Baayen, R. H. & Van Rijn, H. itsadug: Interpreting Time Series and Autocorrelated Data Using GAMMs. https://cran.r-project.org/web/packages/itsadug/index.html

Wieling, M. (2018). Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between L1 and L2 speakers of English. *Journal of Phonetics*, 70, 86–116. https://doi.org/10.1016/j.wocn.2018.03.002

Winter, B., & Wieling, M. (2016). How to analyze linguistic change using mixed models, Growth Curve Analysis and Generalized Additive Modeling. *Journal of Language Evolution, 1*(1), 7-18. doi:10.1093/jole/lzv003

Wood, S. (2019). Mixed GAM Computation Vehicle with Automatic Smoothness Estimation. https://cran.r-project.org/web/packages/mgcv/mgcv.pdf

Wood, S. A. J. (1996). Assimilation or coarticulation? Evidence from the temporal co-ordination of tongue gestures for the palatalization of Bulgarian alveolar stops. *Journal of Phonetics*, 24, 139-164.

Wright, C. E. (1979). Duration differences between rare and common words and their implications for the interpretation of word frequency effects. *Memory & Cognition*, 7(6), 411–419. https://doi.org/10.3758/BF03198257

Xu, A., Birkholz, P., & Xu, Y. (2019). Coarticulation as synchronized dimension-specific sequential target approximation: An articulatory synthesis simulation. In *proceedings of the 2019 International Congress of Phonetic Sciences.* Melbourne, Australia.

Xu, Y. (2007). Speech as articulatory encoding of communicative functions. In *proceedings of the 2007 International Congress of Phonetic Science*s, Saarbrucken, Germany.

Xu, Y. (2020). Syllable is a synchronization mechanism that makes human speech possible. *PsyArXiv*. doi:10.31234/osf.io/9v4hr.

Xu, Y., & Gao, H. (2018). FormantPro as a Tool for Speech Analysis and Segmentation / FormantPro como uma ferramenta para a análise e segmentação da fala. *Revista De Estudos Da Linguagem, 26*(4). doi:10.17851/2237-2083.26.4.1435-1454

Xu, Y., & Liu, F. (2006). Tonal alignment, syllable structure and coarticulation: Toward an intergrated model. *Italian Journal of Linguistics, 18*, 125 – 159.

Xu, Y., & Prom-On, S. (2019). Economy of Effort or Maximum Rate of Information? Exploring Basic Principles of Articulatory Dynamics. *Front Psychol*, 10, 2469. doi:10.3389/fpsyg.2019.02469

Xu, Y., Larson, C. R., Bauer, J. J., & Hain, T. C. (2004). Compensation for pitch-shifted auditory feedback during the production of Mandarin tone sequences. *Journal of the Acoustical Society of America*, *116*, 1168–1178.

Yi, H., & Tilsen, S. (2016) Interaction Between Lexical Tone and Intonation: An EMA Study. In *Proceedings of Interspeech 2016*, pp. 2448-2452.

Yin, J., Shaw, J., Kroos, C., & Best, C. T. (2012). Relations between acoustic and articulatory measurements of /l/. In *proceedings of the 2012 Australasian International Conference on Speech Science and Technology.* Sydney, Australia

Zhang, M., Geissler, C., & Shaw, J. (2019). Gestural Representations of Tone in Mandarin: Evidence From Timing Alternations. *ICPhS* 2019, August, 1803–1807.

Zheng, Y., & Bao, H. (2005). Research on the semivowel by dynamic palatogram in Standard Chinese. *ISCSLP*, 249–258.

Zourmand, A., Mirhassani, S. M., Ting, H. N., Bux, S. I., Ng, K. H., Bilgen, M., & Jalaludin, M. A. (2014). A magnetic resonance imaging study on the articulatory and acoustic speech parameters of Malay vowels. *BioMedical Engineering Online*, 13(1). https://doi.org/10.1186/1475-925X-13-103