INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Coarticulation as synchronised sequential target approximation: An EMA study

Zirui Liu[1], *Yi Xu*[1], *Feng-fan Hsieh*[2]

[1]Department of Speech, Hearing and Phonetic Sciences, University College London, UK
[2]Institute of Linguistics, National Tsing Hua University, Taiwan
zirui.liu.17@ucl.ac.uk, yi.xu@ucl.ac.uk, ffhsieh@mx.nthu.edu.tw

Abstract

In this study we tested the hypothesis that consonant and vowel articulations start at the same time at syllable onset [1]. Articulatory data was collected for Mandarin Chinese using Electromagenetic Articulography (EMA), which tracks flesh-point movements in time and space. Unlike the traditional velocity threshold method [2], we used a triplet method based on the minimal pair paradigm [3] that detects divergence points between contrastive pairs of C or V respectively, before comparing their relative timing. Results show that articulatory onsets of consonant and vowel in CV syllables do not differ significantly from each other, which is consistent with the CV synchrony hypothesis. At the same time, the results also show some evidence that articulators that are shared by both C and V are engaged in sequential articulation, i.e., approaching the V target after approaching the C target.

Index Terms: coarticulation, GAMMs, EMA, CV synchrony

1. Introduction

It has been long recognised that consonant and vowel are temporally overlapped at syllable onset, and that such overlap is the major source of coarticulation [1, 4]. For example, when the phoneme /b/ precedes a rounded vowel such as /u/, lip rounding can be observed around syllable onset, whereas the rounding gesture is absent when the succeeding vowel is unrounded [5]. It is much less clear, however, whether C and V are fully synchronised at syllable onset [1] or are only partially overlapped [6, 7]. The synchronisation view is a recent theoretical development in articulatory phonology [8], but there is not yet systematic empirical evidence. Empirical studies on gestural timing have so far found evidence only for partial CV overlap. For example, Shaw and Chen [7] have shown that vowel onset occurs later than consonant onset in Mandarin Chinese. Turk and Shattuck-Hufnagel [6, p.10] have cited multiple sources of evidence showing that "the timing of movement endpoint has higher priority than the timing of movement onset", due to failure of finding consistent alignment of gestural onsets.

A critical issue in this debate is how to determine the onset of a segment. The conventional method used in many studies is to locate the onset at where its movement velocity reaches 20% of its own peak velocity [2]. The problem with this method is the lack of experimental control, which makes it difficult to rule out confounds. One source of confound is the assumption that the interval of a gesture can be determined simply by identifying an articulatory trajectory moving in the direction of the gestural target. As demonstrated by Gelfer et al. [9], this is prone to error because not all movements, or portions of a movement, in the same direction have the same gestural sources. They showed that some of the lip-rounding movements are actually for un-spreading the spread lips during [i]. Another source of confound is that, as recognised by proposers of the threshold method [2], velocity timing is sensitive to articulatory stiffness. Specifically, when a segment is produced with higher stiffness, the 20% threshold is achieved earlier compared to when stiffness is lower. They found a 41 ms mean difference in durations of nucleus segments between lax and tense vowels obtained with the method, and attributed the difference to variation in stiffness. Variation in stiffness as a confound may be even more problematic when determining CV timing, given that it is known that consonants are likely to be articulated with greater stiffness than vowels [10, 11, 12, 13].

1.1. The minimal pair paradigm

The problem of hidden confounds with the velocity threshold method can be avoided by the minimal pair paradigm [3]. Gelfer et al. [9] used /iC$_n$u/ to contrast /iC$_n$i/ sequences in terms of lip movement (e.g., /itu/ vs. /iti/), and showed that an early lip activity can be observed in both sequences, which suggests that the movement is unrelated to the rounded vowel. Rather, it is a gesture for un-spreading the lips after [i]. This is made even clearer, as they showed, when a second rounding movement close to /u/ became increasingly separated from the first rounding movement as more and more consonants were added between [i] and [u] in the /iC$_n$u/ sequence.

The minimal pair paradigm was successfully used in a number of subsequent studies. Boyce et al. [14] found that, among other things, velum lowering movements can be associated with a low vowel as well as with a nasal consonant, and this significantly reduces the temporal scope of nasal articulation. Chen and Xu [15], also using the method, showed that the neutral tone in Mandarin has a specified underlying target rather than being targetless as had been widely assumed [16]. Most relevantly, based on the minimal pair paradigm, Xu and Gao [3] developed a triplet method to contrast two minimal pairs, one for determining the consonant onset and the other for determining the vowel onset. They used triplets of the form $C_1V_1\#C_2V_2$, where numeric index indicates syllable number and # stands for syllable boundary. In each triplet, the first two words differ in terms of C_2: /l/ vs. /j/, and the second two words differ in terms of V_2: /i/ vs. /u/. By directly comparing the formant trajectories between the V and C pairs, they were able to show, for the first time, clear acoustic evidence that C and V are synchronised by their onsets at the beginning of the syllable. The present study is to use the triplet method to test the CV synchrony hypothesis by examining articulatory trajectories tracked by EMA.

http://dx.doi.org/10.21437/Interspeech.2020-1432

2. Method

2.1. Speakers

7 male and 3 female native speakers of Mandarin Chinese participated in the present study. All of the participants were studying at the National Tsing Hua University in Taiwan and were from the northern part of China (5 from Beijing and 5 from Liaoning). No speaking or hearing difficulties were reported prior to data collection.

2.2. Stimuli

A total of 6 triplets, consisting of 18 $C_1V_1\#C_2V_2$ words in Mandarin were used, as presented in Table 1. In each triplet, the vowel pair differs in terms of V_2 and the consonant pair differs in terms of C_2. All 18 words bear the rising tone on both syllables.

Table 1: *Stimuli in pinyin and IPA. IPA transcriptions are presented in square brackets.*

1	láilí [laɪli]	láilú [laɪlu]	láiyí [laɪji]
2	léilí [leɪli]	léilú [leɪlu]	léiyí [leɪji]
3	lóulí [loʊli]	lóulú [loʊlu]	lóuyí [loʊji]
4	málí [mali]	málú [malu]	máyí [maji]
5	máolí [maʊli]	máolú [maʊlu]	máoyí [maʊji]
6	nílí [nili]	nílú [nilu]	níyí [niji]

The target words were embedded in a carrier sentence – "bǐ ___ wěishàn" ([bi ___ weɪ ʂan]), meaning "more hypocritical than ___". Participants read aloud the sentences with 10 repetitions each in randomised blocks, yielding 1800 tokens in total.

2.3. Data collection and processing

Data collection was done by the third author at the Phonetics Laboratory at the National Tsing Hua University, Taiwan. The articulatory data were collected while subjects read aloud the stimuli using an NDI Wave system. Following procedures in [17], the kinematic data were sampled at a rate of 400 Hz, and the distance values were converted from voltage with a filter with a cut-off frequency of 20 Hz for the upper and lower lips and 40 Hz for the tongue tip. The origin of the coordinate system was placed between the upper incisors on the lower front position. Acoustic data were recorded simultaneously with a sampling rate of 24 kHz.

The auditory tokens were manually annotated at syllable boundaries in the format of [$C_1V_1\#C_2V_2\#$weɪ] with a Praat script, which segmented the formant trajectories and corresponded them with the EMA trajectories. The left-most and right-most boundaries were determined by acoustic onset of C_1 (e.g., nasal murmur in 'maliwei') and end of voicing of [weɪ], respectively. All trajectories were aligned at the first boundary and sampled at 5 ms intervals (i.e., with a sampling rate of 200 Hz). The formant data was calculated with standard parameters using the Burg algorithm (window length = 0.025 s; male maximum formant = 5000 Hz; female maximum formant = 5500 Hz; dynamic range = 30 dB; pre-emphasis from 50 Hz).

Speaker 4's data were excluded from analysis due to background noise present in the audio recording, which led to difficulty in discerning acoustic segmental boundaries for annotation. Out of the remaining 1620 tokens, due to mispronunciation or background noise, 20 were excluded for speaker 1, 1 was excluded for speaker 9, and 1 was excluded for speaker 10.

2.4. Analysis

2.4.1. *Articulatory dimension for detecting vowel and consonant onset*

For detecting consonant onsets by contrasting /l/ and /j/, the tongue tip in the vertical dimension was used. This is motivated by findings on coarticulation resistance of /l/. Recasens and Espinosa [18] have found that for the Spanish /l/, across speakers, tongue tip in the vertical dimension (TTy) shows the least variation between vowel contexts at consonant mid-point, indicating that this is the primary articulatory dimension for /l/. For detecting vowel onsets by contrasting /i/ and /u/, upper lip protrusion (LP) was used.

2.4.2. *Determining significant divergent time points between contrastive pairs using generalised additive mixed models (GAMM)*

To determine the articulatory onset of consonants and vowels, we used GAMM in *R* [19] to model articulatory trajectories of minimal pairs, and tracked statistical differences in articulation over time. GAMM is a kind of non-linear regression, which can model dynamic data while accounting for subject, item or time related variability [20]. For each speaker and each triplet, two GAMMs were built – one for LP for the vowel pair and one for TTy for the consonant pair. Due to GAMMs' requirement of time normalisation, and for the sake of retaining real time information, each token was trimmed to the same length as the shortest token across all tokens (465 ms). Prior to model construction, each speaker's data were normalised into z-scores separately.

GAMMs were constructed using the *bam* function from the *mgcv* package [20] in *R* [19]. Autocorrelation of the residuals was accounted for by incorporating an autoregression model of the error at lag 1, which is supported in the *bam* function. For each GAMM, word was used as the main effect, and a time by word smooth was included with the k parameter set to 15, which accounts for the strong non-linearity of articulatory data. In order to account for variation between repetitions of words, a random smooth modelling non-linear difference over time in relation to repetition was also included in the model. The random smooth is conceptually comparable to a full random effect in Linear mixed effects models, which accounts for the variability in the non-linear trajectories between repetitions (i.e., random slope), as well as the overall height of the trajectory (i.e., random intercept).

Vowel and consonant onset times were collected when the GAMMs indicated significant divergence of articulation between minimal pairs. To avoid type I errors, only divergence that lasted for 50 ms or longer were recorded as segment onset. In total, 108 (6 × 2 × 9) onset times were collected.

2.4.3. *Comparison of C and V onset time using linear mixed effects models (LMEM)*

Linear mixed effects models were built in *R* using the *lme4* package to compare C and V onset time [19, 21]. The models included a fixed effect of contrast/onset type (vowel vs. consonant), and random effects of speaker and triplet. In order to test the significance of the main effect, the *anova* function in

R was used to compare models with and without the fixed effect.

3. Results

Figure 1 shows mean articulation trajectory for LP and TTy of triplet 1. Note that in order to demonstrate systematic patterns of articulation, trajectories are averaged across speakers and repetitions. Only trajectories for triplet 1 are presented here. Other triplets show similar patterns of divergence. It can be observed in the upper panel, the vowel contrast between /i/ and /u/ in 'lailiwei' vs. 'lailuwei' starts to show substantial articulatory difference in LP by around 150 ms. In contrast, as expected, no substantial consonant divergence can be seen between 'lailiwei' and 'laiyiwei', as LP is presumably the same for the articulation of /j/ and /l/.

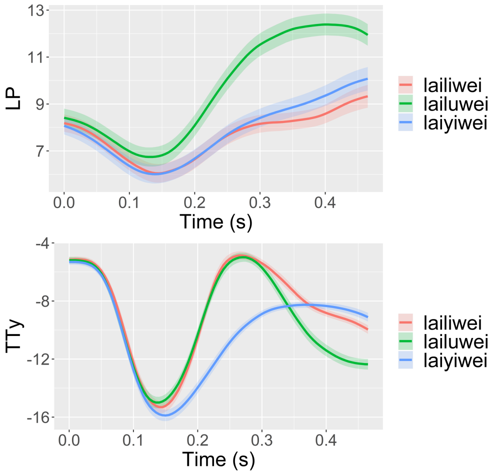

Figure 1: *Mean articulatory trajectories across speakers and repetitions for triplet 1. Shaded ribbons indicate standard error of the mean.*

The lower panel in Figure 1 shows mean articulation in the TTy dimension for triplet 1. Compared to LP, a reversed divergence pattern is shown, between around 150 ms to 300 ms. During this temporal interval, the articulation of the consonants in the second syllable is carried out. Therefore, as a primary parameter for /l/, tongue tip does not differ vertically between 'lailiwei' and 'lailuwei', and the divergent point between 'lailiwei' and 'laiyiwei' can be regarded as the articulatory onset of /l/.

After 300 ms, divergence between the vowel pair ('lailiwei' vs. 'lailuwei') can be observed, which indicates that by this point, the approximation of the [l] target is over, so that the tongue tip can start to move toward the vowel target. This suggests that when an articulator is needed for both the vowel and the consonant on a particular dimension, articulation is sequential along that dimension (to be discussed in section 4).

In the following analysis, therefore, consonant and vowel onset in the syllable /li/ were identified as the point where articulatory trajectories diverged in the TTy dimension in the consonant pair and in the LP dimension in the vowel pair, respectively.

3.1. CV onset times determined by GAMMs

Figure 2 shows results of the two GAMMs model for triplet 1 and speaker 6. Analogous patterns were found for other triplets and speakers. The left two plots show the modelled articulation by GAMM, and the right plots show the difference between the minimal pairs calculated from the GAMM prediction. Intervals where significant differences can be found between trajectories are indicated by red lines in the difference plots. The shaded ribbons represent 95% of the confidence interval, such that the calculated difference become statistically significant when the grey bands are above or below zero. As Figure 2 shows, articulation between the consonant pair become significantly different at around 180 ms (top-right graph), and around the same time for the vowel pair (bottom-right graph). CV onset times were collected via GAMMs for all speakers and triplets, and results are shown in Figure 3. The average onset time in /li/ is 200 ms for the consonant, with a standard deviation of 38 ms, and 190 ms for the vowel, with a standard deviation of 67 ms.

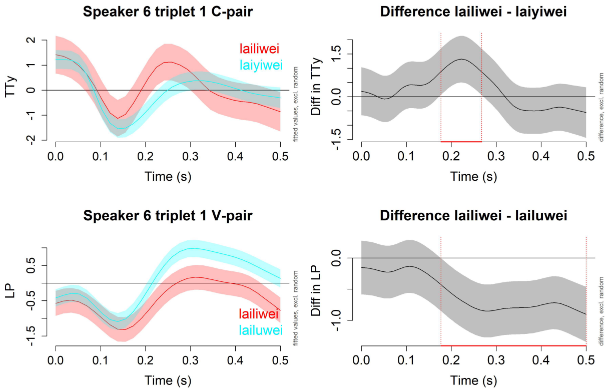

Figure 2: *Articulation modeled with GAMM and difference between the GAMM smooths for triplet 1 and speaker 6 for C and V pairs; The shaded bands indicate 95% of the confidence intervals.*

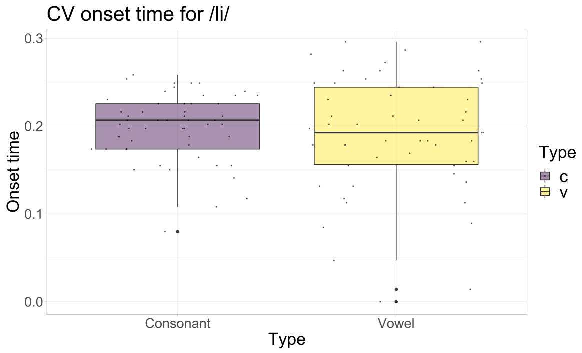

Figure 3: Box plot for *CV onset times.*

3.2. LMEMs for comparison of CV onsets

For analysis of the effect of onset type on segment onset time in CV, a likelihood ratio test was performed on the full model, with onset type as the fixed effect and a null model without the fixed effect. Result shows that there was no significant effect of onset type on onset time (t = -1.02 (LMEM output); X^2 (1) = 1.03, p = 0.31 (likelihood ratio test output)). In other words, consonant onset times did not differ significantly from vowel onset times in the current data.

3.3. Comparison between the threshold and minimal pair methods' results

To examine how different the results would be with the conventional approach, we applied the 20% threshold method to determine the vowel onsets in /lu/ based on LP. Words with rounded vowels in the first syllable (e.g., 'louluwei') were excluded to avoid the effects from the first rounding gesture for the threshold method. Overall, V onsets in /lu/ determined by the threshold method was 17 ms earlier than by the minimal pair method. Due to highly uneven sample sizes between the results from the two methods, only qualitative analysis is offered here.

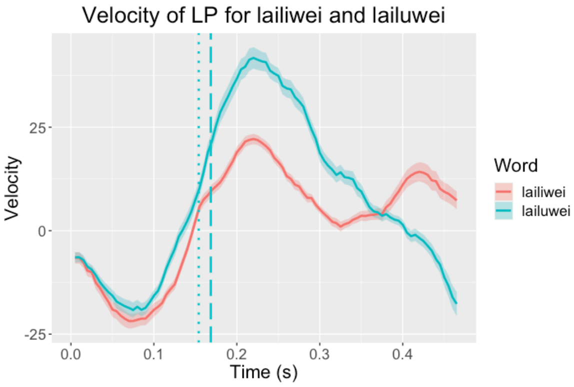

Figure 4: *Velocity of LP as a function of time for 'lailwei' vs. 'lailuwei'. The shaded ribbons indicate standard error of the mean. The blue dotted vertical line marks mean /u/ onset determined by the threshold method and dashed line indicate that of the minimal pair method.*

Figure 4 shows velocity contours of LP for the words 'lailiwei' (pink) and 'lailuwei' (blue). For the blue line, it may seem that the entire rising section of the velocity trajectory belongs to /u/ in 'lailuwei'. Indeed, the threshold method indicates that the /u/ onset is around the start of the rising movement, as marked by the dotted vertical line. However, the velocity trajectory of 'lailiwei' also shows a similar LP movement on a smaller scale. This suggests that the start of the LP movement for 'lailuwei' is not for the rounding of /u/, but for an unspreading movement after /lai/. The true onset of /u/ should be later: at the point when the trajectories of /lu/ and /li/ move away from each other. Indeed, that is exactly the mean onset determined by the minimal pair method, as marked by the dashed vertical line. That onset time occurs right after the LP velocity of 'lailiwei' makes a slight turn toward a reduced velocity slope.

4. Discussion

4.1. Articulatory evidence for CV synchrony

With the triplet method based on the minimal pair paradigm, we tested whether there is evidence of synchronous onset of C and V in CV syllables in Mandarin. The results show that, for all the triplets, the moment when the LP trajectories start to diverge toward the contrasting vowels is no different from the moment when the TTy trajectories start to diverge toward the contrasting consonants. This suggests that consonant and vowel in CV syllables in Mandarin Chinese are synchronised at syllable onset. This finding is consistent with the previous finding based on formant trajectories, also using the triplet method [3]. We have therefore seen both acoustic and articulatory evidence that the production of consonants and vowels are synchronised at the onset of CV syllables in Mandarin.

We have also found that, with the velocity threshold method widely used in previous articulatory studies, vowel onsets would have been located at an earlier point than the those determined by the triplet method. This is because the velocity threshold method is prone to confounds due to lack of experimental control. The major confounds in articulatory analysis may include articulatory movements that are in similar directions as the movement in question but associated to an adjacent segment [9, 14], and intrinsic differences in stiffness between consonants and vowels [2, 10, 11, 12, 13].

4.2. Evidence for articulator-specific sequential target approximation

Synchronised onset of C and V does not mean that all the associated articulatory movements are in synchrony, however. Those that are shared by both C and V may need to be sequentially executed, i.e., C before V [1]. Signs of sequential articulation can be seen in the current data, e.g., in the lower graph of Figure 1. There the divergence of TTy between /ji/ vs. /li/ at around 150 ms indicates that the consonant articulation has started. However, no divergence of TTy can be seen between the vowel pair /li/ vs. /lu/ until after around 300 ms. This means that the tongue tip has to first meet the requirement of /l/ for making an alveolar contact, which is terminated only at the end of the consonant around 300 ms.

Sequential articulation as a solution to direct articulatory conflicts has been reported before [22]. But it is inconsistent with the hypothesis that articulatory gestures involving the same articulator can be overlapped through blending [23, 13]. The present data are not enough to resolve the difference, which would require further research.

5. Conclusion

In this study we used EMA data to test for evidence of synchronised articulation of consonant and vowel at syllable onset in Mandarin. We applied a triplet method previously developed for acoustic analysis [3] based on a minimal pair paradigm to detect divergence points in contrastive pairs of C and V before comparing their relative timing. Results show that articulatory onsets of consonant and vowel in CV syllables do not differ significantly from each other. These results provide the first clear articulatory evidence in support of the CV synchrony hypothesis. In addition, we have found evidence of sequential articulation at syllable onset for articulators that are shared by consonants and vowels. These findings demonstrate the effectiveness of the minimal pair paradigm as a means to implement full experimental control in articulatory investigation, as suggested long ago [9, 14].

6. References

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

[2] P. Hoole, C. Mooshammer, and H. G. Tillman, "Kinematic analysis of vowel production in German," presented at the

International Conference on Spoken Language Processing, Yokohama, 1994.

[3] Y. Xu and H. Gao, "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,* vol. 26, no. 4, 2018, doi: 10.17851/2237-2083.26.4.pp. 1435-1454.

[4] H. Nam, L. Goldstein, and E. Saltzman, "Self-organization of Syllable Structure: a Coupled Oscillator Model," in *Approaches to Phonological Complexity*. Germany: De Gruyter Mouton, 2009, pp. 299-328.

[5] V. Kozhevnikov and L. Chistovich, "Speech: Articulation and Perception " in *Translation by Joint Publications Research Service*. Washington, DC, 1965.

[6] A. Turk and S. Shattuck-Hufnagel, "Timing Evidence for Symbolic Phonological Representations and Phonology-Extrinsic Timing in Speech Production," *Front Psychol,* vol. 10, p. 2952, 2019, doi: 10.3389/fpsyg.2019.02952.

[7] J. A. Shaw and W. R. Chen, "Spatially Conditioned Speech Timing: Evidence and Implications," *Front Psychol,* vol. 10, p. 2726, 2019, doi: 10.3389/fpsyg.2019.02726.

[8] E. Saltzman and D. Byrd, "Task-dynamics of gestural timing: Phase windows and multifrequency rhythms," *Human Movement Science,* vol. 19, no. 4, pp. 499-526, 2000, doi: 10.1016/s0167-9457(00)00030-0.

[9] C. E. Gelfer, F. Bell-Berti, and K. S. Harris, "Determining the extent of coarticulation: effects of experimental design," *J Acoust Soc Am,* vol. 86, no. 6, pp. 2443-5, Dec 1989, doi: 10.1121/1.398452.

[10] H. Nam, "Articulatory Modeling of Consonant Release Gesture," presented at the International Congress of Phonetic Sciences, Saabrücken, Germany, 6 - 10 August, 2007.

[11] C. P. Browman and L. Goldstein, "Articulatory phonology: an overview," *Phonetica,* vol. 49, no. 3-4, pp. 155-80, 1992, doi: 10.1159/000261913.

[12] M. Pastätter and M. Pouplier, "The articulatory modelling of German coronal consonants using TADA," in *Proceedings of the 10th International Seminar on Speech Production*, Cologne, Germany, 5-8 May 2014. [Online]. Available: https://www.phonetik.uni-muenchen.de/universals/pub/ISSP2014_TADA.pdf.

[13] E. L. Saltzman and K. G. Munhall, "A Dynamical Approach to Gestural Patterning in Speech Production," *Ecological Psychology,* vol. 1, no. 4, pp. 333-382, 1989, doi: 10.1207/s15326969eco0104_2.

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

[15] Y. Chen and Y. Xu, “Production of weak elements in speech — Evidence from f0 patterns of neutral tone in standard Chinese,” *Phonetica* vol. 63, pp. 47-75, 2006, doi:10.1159/000091406

[16] Y. R. Chao, *A Grammar of Spoken Chinese*. Berkeley: University of California Press, 1968.

[17] J. A. Shaw, A. I. Gafos, P. Hoole, and C. Zeroual, "Dynamic invariance in the phonetic expression of syllable structure: a case study of Moroccan Arabic consonant clusters," *Phonology,* vol. 28, no. 3, pp. 455-490, 2011, doi: 10.1017/s0952675711000224.

[18] D. Recasens and A. Espinosa, "An articulatory investigation of lingual coarticulatory resistance and aggressiveness for consonants and vowels in Catalan," *J Acoust Soc Am,* vol. 125, no. 4, pp. 2288-98, Apr 2009, doi: 10.1121/1.3089222.

[19] R. (2014), R Core Team. Available: https://www.r-project.org

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

[21] Linear mixed-effects models using eigen and s4. (2019), LME4 authors. Available: https://cran.r-project.org/web/packages/lme4/lme4.pdf

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

[23] C. P. Browman and L. Goldstein, "Articulatory gestures as phonological units," *Phonology,* vol. 6, no. 2, pp. 201-251, 1989.