INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Segmental Alignment of English Syllables with Singleton and Cluster Onsets

Zirui Liu, Yi Xu

University College London, UK
Zirui.liu.17@ucl.ac.uk, yi.xu@ucl.ac.uk

Abstract

Recent research has shown fresh evidence that consonant and vowel are synchronised at the syllable onset, as predicted by a number of theoretical models. The finding was made by using a minimal contrast paradigm to determine segment onset in Mandarin CV syllables, which differed from the conventional method of detecting gesture onset with a velocity threshold [1]. It has remained unclear, however, if CV co-onset also occurs between the nucleus vowel and a consonant cluster, as predicted by the articulatory syllable model [2]. This study applied the minimal contrast paradigm to British English in both CV and clusterV (CLV) syllables, and analysed the spectral patterns with signal chopping in conjunction with recurrent neural networks (RNN) with long short-term memory (LSTM) [3]. Results show that vowel onset is synchronised with the onset of the first consonant in a cluster, thus supporting the articulatory syllable model.

Index Terms: coarticulation, LSTM, RNN, consonant cluster, synchornisation

1. Introduction

Evidence for co-production between syllable initial segments has been shown from the very early days of phonetic research, e.g., the observation of lip rounding at the beginning of /ku/ [4], tongue rising at the start of /bi/ [5], and simultaneous articulation between onset cluster [6] [7]. This has led to models that predict CV synchrony at the syllable onset, including the articulatory syllable model [2], the Task Dynamics (TD)/Articulatory Phonology (AP) frameworks [8] and the synchronisation model of the syllable [9]. The strongest claim among these models is that the co-onset happens between the vowel and the very first consonant in an onset cluster, and the syllable is the domain of coarticulation [2]. The articulatory syllable hypothesis, however, has been disputed by the observation of anticipatory coarticulation [10] [7]. Further uncertainty about the hypothesis is brought about by the c-center effect [11], whereby the vowel appears to be aligned to the center of the consonant cluster rather than the onset of the first consonant. The c-center effect, together with more recent findings of CV asynchrony in CV syllables [12] [13] [14] have also brought uncertainties about the CV synchrony assumption in the strong version of the TD/AP framework [8]. It is not until recently that empirical evidence for full CV synchrony was shown for Mandarin Chinese [15] by applying a minimal triplet paradigm. The present study is a continuation of this line of work by examining CV alignment in consonant clusters in British English.

1.1. Current theories and findings on cluster and vowel alignment

Contemporary research on the time alignment of consonant clusters is predominantly focused on the c-center effect [16], developed under the AP framework [17]. In AP, the onset of C is said to be synchronous with the vowel due to their in-phase coupling relationship, while the coupling relationship between cluster components is anti-phase [18]. Therefore, a competitive coupling demand arises when both cluster components are coupled in-phase with the vowel, but anti-phase with each other. The competing demand is resolved through the c-center mechanism, in which the cluster components remain anti-phase with each other, but the rightmost component shifts more into the vowel to make space for the added consonants [11]. Under the c-center hypothesis, empirical studies have compared relative time lags between cluster and singleton conditions, where the time lag is calculated as the midpoint of singleton or cluster consonants subtracted from the time of an anchor point, usually the coda consonant or vowel offset [19] [20] [21] [22] [23] [24]. The c-center effect is said to be observed if the c-center to anchor lags are equal between the singleton and cluster condition.

The time lag measurements calculated this way are prone to confounds, such as prosodic effects including word edge lengthening [19] and potential variation in vowel duration. More importantly, observed c-center effects could be due to segment compression, as it is well known that consonants are shortened in clusters compared to singletons [25] [26] [27] [28]. Given the compression, it is predictable that the relative time lags would tend to be equal between singleton and cluster. The more appropriate approach would be to directly compare the onset time of the consonant with that of the vowel at the beginning of the syllable. For that purpose, however, it is critical to find a reliable way of determining segment onsets.

1.2. Methods of determining segment onset

Most articulatory or acoustic studies examining dynamic/time series data use the velocity threshold method to determine movement onset [1]. Specifically, the onset of a gesture is located as when its associated movement (e.g., F2 or TTy—tongue tip height) reaches 20% of its peak velocity. However, onsets determined this way may be subject to confounds, such as articulatory stiffness difference between C and V [1], undershoot [29] and other concurrent articulations [15]. A more effective way to control these confounds is through the use of minimal contrasts [30], which has long been the norm in investigating many other aspects of speech [31]. A minimal pair method is shown to be able to control for confounds in studying dynamic articulatory movements [32] [30], in which the onset of a segment is defined as when articulation or acoustics becomes significantly different between members of a contrastive pair [32] [30] [33]. For example, the onset of /i/

http://dx.doi.org/10.21437/Interspeech.2021-187

in /li/ can be located as when /li/ becomes sufficiently different from /lu/ in terms of F2 or lip protrusion [15] [34] [30]. The minimal pair method was later adapted into a minimal triplet paradigm to investigate CV onsets in Mandarin [15] [34]. To detect onsets of both C and V, a CV triplet consisting of two minimal pairs are used to contrast both the consonant and the vowel. The consonant pair differed in terms of the onset C, and the vowel pair in terms of V. With the minimal triplet paradigm, the first articulatory evidence for CV co-onset was found [15].

1.3. Combining the minimal triplet paradigm with signal chopping and neural networks

Previous acoustic studies using the minimal triplet paradigm have so far used a single formant to track segmental movements [15] [34]. But spectral changes involve more than just one formant. To further improve the processing of acoustic contrast in minimal pairs, the present study explored Mel-frequency cepstral coefficients (MFCC) in place of formants. Another advantage of MFCCs is that they are not interrupted by an obstruent consonant, even if it is voiceless. To process the high dimensional MFCC data for tracking acoustic differences over time, we adapted the signal chopping method developed in Tilsen [3]. The main idea is that, when there is enough acoustic information in the data for a classifier to distinguish between the V or C/CL pair, movement towards the contrastive targets have started. For each contrastive pair in a triplet, neural network classifiers are trained to classify the word category in the training set then the accuracy of the model is obtained by the test set. To identify segment onset time, we truncated the time series data frame by frame and trained a classifier for each truncated dataset. Segment onset is located when the model accuracy falls to or below chance level on the truncated dataset. For example, when the truncated dataset consist of 0.05 s of the original time series data and the model accuracy falls to chance level, the segment onset is recorded as 0.05 s for this minimal pair.

The neural networks used are LSTM RNNs. RNNs are powerful models for sequential data such as time series speech data. However, naïve RNNs cannot grasp long distance dependencies in the data due to problems with vanishing/exploding gradients [35]. LSTM units can handle the vanishing gradient problem [36] and have been able to deliver high accuracy results in speech recognition [37] [38].

2. Method

2.1. Stimuli

A total of 18 triplets were designed in the study, 9 for the singleton condition and the other 9 for the cluster condition. As shown in Table 1, all words in each triplet have the same syllable structure. In each triplet, the first two words/syllables differ in terms of V and the second and third in terms of onset C/C_1. Therefore, the C/CL and V onsets in the second word can be determined as when acoustics features become distinctly different between the C and V pairs, respectively. The cluster design aims to test whether the vowel is synchornised with the start of the whole cluster. Note that for triplet 6 in the cluster block, there are no consonant clusters with a C_1 minimal contrast with /sk/ in English, so only a vowel pair is included. Around 20% of the stimuli contain simple pseudo words, which all participants pronounced successfully. The singleton and cluster conditions are matched as much as possible while keeping the number of pseudo words to a minimum. The target words were embedded in the carrier phrase 'see a _ today'. 10 repetitions were produced in randomised blocks, which yielded 5200 tokens in total (52 words × 10 repetitions × 10 speakers). 15 tokens were excluded from analysis due to mispronunciation or spontaneous pausing.

Table 1: *stimuli.*

Triplets	Singleton			Cluster		
1	bar	bee	dee	blah	blee	glee
2	pit	pot	cot	plit	plot	clot
3	fee	four	sore	flee	floor	slaw
4	sit	sot	fot	slit	slot	flot
5	sal	seal	shiel	spal	speel	spiel
6	sah	see	fee	scar		ski
7	caught	keat	Pete	clawt	cleat	pleat
8	git	got	bot	grit	grot	brot
9	dip	dop	cop	drip	drop	crop

2.2. Data collection and processing

5 female and 5 male British English speakers participated as subjects. All the participants have a Standard Southern British accent and are aged between 20 and 40. Due to the COVID-19 pandemic, the recording took place over Zoom. Participants all used an external microphone or a headset during recording, and the original sound feature was turned on to avoid audio enhancement in the Zoom application. Participants read aloud the sentences with the embedded stimuli while being recorded at a sampling rate of 32 kHz. The tokens were annotated in Praat in the format of [siəC$_n$V]. The left and right boundaries are determined by the voice onset of /i/ in 'see' and the end of voicing for the vowel in the target word, respectively.

The 15-dimensional MFCC features were extracted using *python_speech_features*[1] with a 25 ms Hamming window in 5 ms intervals. Cepstral Mean Subtraction (CMS) was performed on the MFCC features for individual speakers, which can mitigate any noise or channel distortion in the recording [39]. The first order derivatives were calculated for all 15 MFCCs, which yielded a 30-dimensional feature vector per frame.

2.3. LSTM RNN analysis with signal chopping

2.3.1. Analysis procedure with signal chopping

All the tokens are aligned at the voice onset of /i/ in 'see' (i.e., the first boundary). For each minimal pair in a triplet, the training set consists of 7 repetitions of the minimal pair from all the speakers, which results in 140 tokens in total (7 repetitions × 2 words × 10 speakers). The testing set has the remaining 3 repetitions containing 60 tokens. The targets are binary class labels for each utterance. The original feature sequences (i.e., before signal chopping) are all trimmed to be the same duration as the shortest token overall for that minimal pair, so that no padding was necessary. A classifier is trained on the training set, then the model accuracy is obtained using the testing set. The training and testing sets with N original frames are truncated frame by frame until N = 5 (i.e., 0.025 s in duration). An accuracy score is recorded for each

[1] https://github.com/jameslyons/python_speech_features (v0.6.1).

truncated dataset from the classifier. Note that a classifier is trained from scratch for each truncated dataset. The entire chopping analysis procedure is repeated 10 times for each minimal pair, by randomly assigning repetition blocks to the training and testing sets.

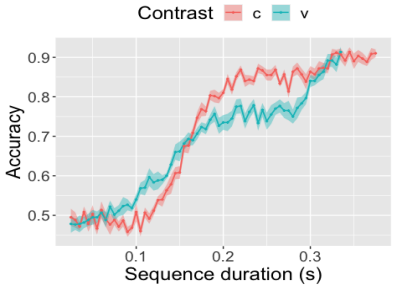

Figure 1: *Mean network accuracy across 10 trials for one triplet as a function of remaining sequence duration. Shaded ribbons indicate standard error of the mean. The C pair and V pair are colour coded as blue and red respectively.*

An example of a triplet is shown in Figure 1. The C and V onsets are recorded as the sequence duration when prediction accuracy drops to or below 0.6 without returning above it, as no categorical information can be detected from that point on. We used 0.6 instead of 0.5 as the threshold due to evidence for above chance classifier performance reported by studies using permutation testing. In a permutation test, the features and labels are permuted to destroy the underlying class structure, and a null accuracy distribution is estimated from repeatedly training classifiers on the permuted data. Ojala and Garriga [40] shows the upper confidence bound of the null accuracy distribution can go up to 0.4 for a 3-class classifier. In Tilsen [3], 0.1 above the chance level was used. Combrisson and Jerbi [41] reported that for a 2-class classifier with a similar training sample size to ours, the classifier needs to obtain an accuracy of around 0.6 or higher to reach significance compared to the null distribution. Therefore, 0.6 was chosen as the threshold for the current analysis to avoid Type I errors. In total, 350 onsets are collected from the analysis (35 minimal pairs × 10 randomised trials). Finally, Linear Mixed Effects Models (LMEMs) are constructed to test whether C/CL onsets differ from V onsets in time. For the LMEM, syllable type (cluster vs. singleton) and onset type (C/CL vs. V) are included as fixed effects and the repetition randomisation as a full random effect. Likelihood ratio tests are used to test for significance of the fixed effects. An interaction model is constructed by including an interaction term between syllable and onset types to test whether the effect of onset type differs between singleton and cluster conditions.

2.3.2. LSTM RNN architecture

All the neural networks have identical structures and hyperparameters[2]. The N × 30 feature sequences are fed into the input layer. The hidden layers consist of 2 bi-directional LSTM layers with variational dropout. Variational dropout has been shown to prevent overfitting for gated RNN networks better than the conventional dropout method [42] [43]. The output layer is one dimensional with a sigmoid activation function to predict the probability of the word class. To avoid model bias, the network structure and hyperparameters were tuned by using grid search with 5-fold validation on data from a pilot study, i.e., the training and testing data in the analysis from section 2.3.1 were not used during model tuning.

The main purpose is to use machine learning (ML) methods to test for the presence of categorical acoustic information over time. Therefore, although network performance can be further optimised for each analysis individually, it is not essential for the current purpose.

3. Results

3.1. LSTM RNN Signal chopping results

Results of using signal chopping with LSTM RNNs are presented in Figure 2. Each plot in the Figure correspond to a triplet in Table 1. i.e., C1 correspond to the first cluster triplet – 'blee' vs. 'blah' vs. 'blee'. The shaded areas indicate the mean acoustic onset and offset of the voiceless, closure or frication interval of the onset C/CL for each individual triplet. In general, a similar trend to Figure 1 can be observed for all the triplets in Figure 2. As more frames are truncated, or in other words, the shorter the remaining sequence duration, the lower the model accuracy. For the consonant minimal pairs, model accuracy stays high and drops sharply before the acoustic closure/frication of the onset C/CL. Some of the vowel pairs follow the same pattern, such as triplet S8, S7, C7, S2, S1 and S3. On the other hand, classification accuracy for the other vowel pairs drops once the voiced portion of the contrasting syllable is fully truncated, which is indicated by the right boundary of the shaded rectangles. This is more so for triplets with fricative onsets (e.g., S5). However, the accuracy rate for these vowel pairs oscillates around or hover slightly above 0.6 during the voiceless or frication intervals, indicating the presence of some categorical information in the acoustic features. For all the triplets, the model accuracy for both the consonant and vowel pairs fall to and remain below 0.6 when the MFCC feature sequences only contain the voiced portion of [siɘ]. Most importantly, the neural network classifiers start performing below the 0.6 accuracy threshold at around the same truncation point for the C and V pairs in each triplet. For example, for C1, the blue and red lines correspond to model accuracy for the V and C pairs as a function of truncation, and they both fall below 0.6 at around 0.15 s. The C/CL and V onsets are collected as when accuracy falls to or under 0.6 for all triplets and randomised trials.

3.2. LMEM results for consonant and vowel onsets

Figure 3 shows the segment onset times collected in 2.3.1. According to the segment onsets determined by the current analysis, consonant and vowel onsets do not differ much from each other, for both syllables with singleton and cluster onsets. LMEMs shows that onset time does not differ between onset types (t = -0.83l; $X^2(1)$ = 0.67, p = 0.41). The model fit does not improve significantly by adding an interaction term for onset type and syllable type (t = -1.69; $X^2(1)$ = 2.82, p = 0.09). Therefore, onset type has no significant effect on onset time for either singleton or cluster syllables.

[2] Full network details can be found at https://github.com/Clara-liu/English_alignment_LSTM

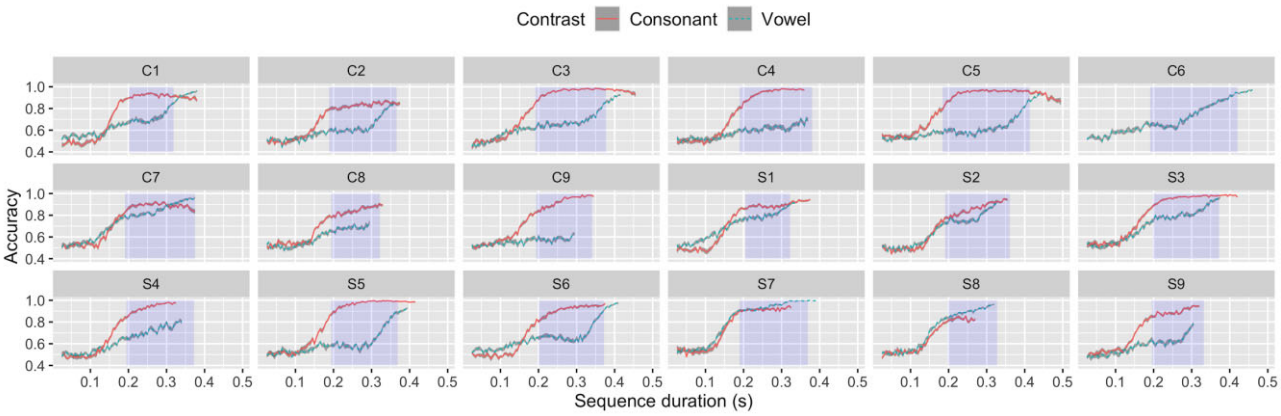

Figure 2: *Mean prediction accuracy as a function of remaining sequence duration. The grey ribbons indicate standard error of the mean. The shaded areas indicate mean intervals of frication, closure or aspiration of C/ CL. The facet labels correspond to the onset type and triplet number in Table 1. e.g., C1 – first triplet in the cluster condition.*

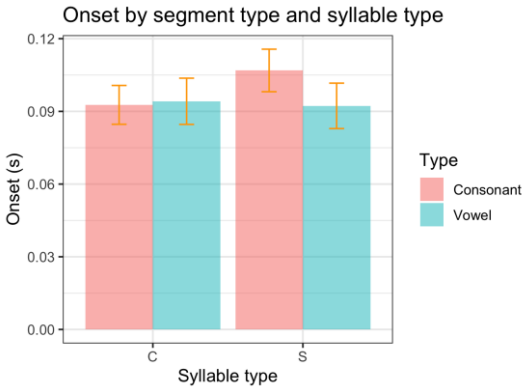

Figure 3: *Mean consonant and vowel onsets by syllable type. The error bars represent 95% of the confidence interval. Onset type is indicated by colour and syllable type by the x-axis.*

4. Discussion

4.1. Acoustic evidence for onset C/CL and V synchrony

By combining signal chopping with neural networks and the minimal triplet paradigm, we were able to track overall spectral movement over time. The results showed that a neural network classifier can detect categorical information in the acoustic signal around the same time for the onset C/CL and the nucleus vowel. Therefore, the results support the notion of synchronisation – that the start of the vowel is synchronised with the onset consonant [9], for both singleton and cluster onsets [2]. Most importantly, contrary to the prediction of the c-center hypothesis, the results do not show that the cluster onset precedes the vowel onset.

Figure 2 shows that movement towards a contrasting vowel target start well before the acoustic landmark of the onset consonant (e.g., start of frication in /s/). As reported in the classic spectrographic study of Öhman [10], this is widely believed to be anticipatory coarticulation with the vowel. However, by adding a consonant contrast in the minimal triplet paradigm, it has become clear that the movement toward the onset consonant also starts well before the acoustic landmark, as can be seen in Figure 2. Tilsen [3] has found similar results regarding articulatory onset for the onset consonant, although no vowel contrast was examined. The current results therefore suggest that the widely reported anticipatory coarticulation is the true onset of the syllable, since the vowel and consonantal movements begin simultaneously before the commonly assumed syllable boundary (i.e., acoustic landmark of the onset C).

4.2. Degree of coarticulation in relation to coarticulation resistance

It is important to note that CV synchrony does not entail complete coproduction of onset consonant and vowel. If an articulator is needed for both C and V, articulation needs to be sequential for that particular articulator [44] [15] [9] [2]. Therefore, the more articulators the onset C or cluster segment requires, the less overall coarticulation occurs. This is related to the phenomenon of coarticulation resistance [45] [46] [47]. For example, ultrasound imaging studies have shown that tongue shape varies less at consonant midpoint in fricatives than in plosives [48] [49]. This is also reflected in Figure 2. For the vowel minimal pairs, the classifiers achieve better accuracy for triplets with plosive onsets (triplet 1,2,7,8 and 9) than fricative onsets (triplet 3,4,5,6), indicating more overall vowel coarticulation in the former.

5. Conclusions

This study has found acoustic evidence for onset C/CL and V synchrony, by combining a novel signal chopping method [3] and the minimal triplet paradigm [34] [15]. The finding not only replicates, for British English, previous findings of CV synchrony in Mandarin [15] [34], but also offers fresh support for the articulatory syllable model [2] whereby vowel articulation starts with the very first consonant in an onset cluster. The results also demonstrate the effectiveness of using high dimensional acoustic measurements such as MFCC in combination with ML for phonetic research.

6. References

[1] P. Hoole, C. Mooshammer, and H. G. Tillmann, "Kinematic analysis of vowel production in German," in *ICSLP*, 1994, no. 94, pp. 53–56.

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

[3] S. Tilsen, "Detecting anticipatory information in speech with signal chopping," *J. Phon.*, vol. 82, pp. 1-27, 2020.

[4] E. Sievers, *Grundzüge der Lautphysiologie zur Einführung in das Studium der Lautlebre der Indogermanischen Sprachen*. Leipzg: Breitkopf and Härtel, 1876.

[5] R. Steston, *Motor phonetics: A study of speech movements in action*. Amsterdam: North-Holland, 1951.

[6] P. J. Rousselot, *Principes de phonétique experimentale*. Paris: H. Welter.

[7] B. Kühnert and F. Nolan, "The origin of coarticulation," in *Coarticulation: Theory, Data and Techniques*, Cambridge: Cambridge University Press, 1999, pp. 7–30.

[8] H. Nam, L. Goldstein, and E. Saltzman, "Self-organization of syllable structure: A coupled oscillator model," *Approaches to Phonol. Complex.*, no. December 2009, pp. 299–328, 2009.

[9] Y. Xu, "Syllable is a synchronization mechanism that makes human speech possible," pp. 1–44, 2020.

[10] S. E. G. Öhman, "Coarticulation in vcv utterances: Spectrographic measurements," *J. Acoust. Soc. Am.*, vol. 39, no. 1, pp. 151–168, 1966.

[11] D. Byrd, "C-Centers Revisited," *Phonetica*, vol. 52, pp. 285–306, 1995.

[12] M. Gao, "Gestural Coordination among Vowel, Consonant and Tone Gestures in Mandarin Chinese," *Chinese J. Phonetics*, p. Beijing: Commercial Press, 2009.

[13] J. A. Shaw and W. R. Chen, "Spatially Conditioned Speech Timing: Evidence and Implications," *Front. Psychol.*, vol. 10, no. December, pp. 1–17, 2019.

[14] S. Tilsen, "Exertive modulation of speech and articulatory phasing," *J. Phon.*, vol. 64, pp. 34–50, 2017.

[15] Z. Liu, Y. Xu, and F. F. Hsieh, "Coarticulation as synchronised sequential target approximation: An EMA study," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2020-Octob, pp. 1381–1385, 2020.

[16] C. P. Browman and L. Goldstein, "Some Notes on Syllable Structure in Articulatory Phonology," *Phonetica*, vol. 45, pp. 140–155, 1988.

[17] C. P. Browman and L. Goldstein, "Articulatory phonology: An overview," *Phonetica*, vol. 49, no. 3–4, pp. 155–180, 1992.

[18] S. Marin and M. Pouplier, "Articulatory synergies in the temporal organization of liquid clusters in Romanian," *J. Phon.*, vol. 42, pp. 24–36, 2014.

[19] 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.

[20] M. Pastätter and M. Pouplier, "The temporal coordination of Polish onset and coda clusters containing sibilants," *Proc. International Seminar on Speech Production*, 2014.

[21] M. Pastätter and M. Pouplier, "Articulatory mechanisms underlying onset-vowel organization," *J. Phon.*, vol. 65, pp. 1–14, 2017.

[22] S. Marin and M. Pouplier, "Articulatory synergies in the temporal organization of liquid clusters in Romanian," *J. Phon.*, vol. 42, pp. 24–36, 2014.

[23] S. Marin and M. Pouplier, "Organization of complex onsets and codas in American English: Evidence for a competitive coupling model," in *Proceedings of ISSP 2008 - 8th International Seminar on Speech Production*, 2008, pp. 433–436.

[24] A. Hermes, M. Grice, D. Mücke, and H. Niemann, "Articulatory coordination and the syllabification of word initial consonant clusters in Italian," in *Consonant Clusters and Structural Complexity*, P. Hoole, L. Bombien, M. Pouplier, C. Mooshammer, and B. Kühnert, Eds. Berlin: De Gruyter Mouton, 2012, pp. 157–176.

[25] N. Umeda, "Consonant duration in American English," *J. Acoust. Soc. Am.*, vol. 61, no. 3, pp. 846–858, 1977.

[26] F. Kügler, "Timing of legal and illegal consonant clusters in Swedish," *Proc. FONETIK,* 2007.

[27] D. H. Klatt, "Linguistic uses of segmental duration in English: Acoustic and perceptual evidence," *J. Acoust. Soc. Am.*, vol. 59, no. 5, pp. 1208–1221, 1976.

[28] D. O'Shaughnessy, "A study of French vowel and consonant durations," *J. Phon.*, vol. 9, no. 4, pp. 385–406, 1981.

[29] Y. Xu and S. Prom-on, "Economy of Effort or Maximum Rate of Information? Exploring Basic Principles of Articulatory Dynamics," *Front. Psychol.*, vol. 10, no. November, pp. 1–22, 2019.

[30] 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. 2443–2445, 1989.

[31] J. Laver, *Principles of phonetics*. Cambridge: Cambridge University Press, 1994.

[32] S. E. Boyce, R. A. Krakow, F. Bell-Berti, and C. E. Gelfer, "Converging sources of evidence for dissecting articulatory movements into core gestures," *J. Phon.*, vol. 18, no. 2, pp. 173–188, 1990.

[33] F. Bell-Berti and K. S. Harris, "A temporal model of speech production," *Phonetica*, vol. 38, pp. 9–20, 1981.

[34] 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," *Rev. Estud. Da Ling.*, vol. 26, no. 4, p. 1435, 2018.

[35] Y. Bengio, P. Simard, and P. Frasconi, "Learning Long-Term Dependencies with Gradient Descent is Difficult," *IEEE Trans. Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.

[36] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[37] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM networks," in *IEEE International Joint Conference on*, 2005, vol. 4, pp. 2047–2052.

[38] H. Soltau, H. Liao, and H. Sak, "Neural speech recognizer: Acoustic-To-word LSTM model for largevocabulary speech recognition," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2017, vol. 2017-Augus, pp. 3707–3711.

[39] R. A. Bates and M. Ostendorf, "Reducing the effect of linear channel distortion on continous speech recognition," in *IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING*, 1999, vol. 7, no. 5, pp. 594–597.

[40] M. Ojala and G. C. Garriga, "Permutation tests for studying classifier performance," *J. Mach. Learn. Res.*, vol. 11, pp. 1833–1863, 2010.

[41] E. Combrisson and K. Jerbi, "Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy," *J. Neurosci. Methods*, vol. 250, pp. 126–136, 2015.

[42] A. M. and G. H. Alex Graves, "Speech Recognition with Deep Recurrent Neural Networks, Department of Computer Science, University of Toronto," *Dep. Comput. Sci. Univ. Toronto*, vol. 3, no. 3, pp. 45–49, 2013.

[43] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *33rd International Conference on Machine Learning, ICML 2016*, 2015, vol. 3, pp. 1651–1660.

[44] A. Xu, P. Birkholz, and Y. Xu, "Coarticulation as synchronized dimension-specific sequential target approximation : An articulatory synthesis simulation," in *ICPhS*, 2019, no. July, pp. 2–6.

[45] D. Recasens, "Vowel-to-vowel coarticulation in Catalan VCV sequences," *J. Acoust. Soc. Am.*, vol. 76, no. 6, pp. 1624–1635, 1984.

[46] D. Recasens, "An acoustic analysis of V-to-C and V-to-V coarticulatory effects in Catalan and Spanish VCV sequences," *J. Phon.*, vol. 15, no. 4, pp. 299–312, 1987.

[47] D. Recasens, "Long range coarticulation effects for tongue dorsum contact in VCVCV sequences," *Speech Commun.*, vol. 8, no. 4, pp. 293–307, 1989.

[48] 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. 2288–2298, 2009.

[49] D. Recasens and C. Rodríguez, "A study on coarticulation resistance and aggressiveness for front lingual consonants and vowels using ultrasound," *J. Phon.*, vol. 59, pp. 58–75, 2016.