Ambisyllabicity as decoupling of constriction formation and release

Sam Tilsen
Cornell University
tilsen@cornell.edu

word count: 18,576

Abstract

Many phonological theories allow for the possibility that an intervocalic consonant may be associated with both a preceding and following syllable, a pattern called ambisyllabicity. This study investigated constriction formation and release movements of bilabial and alveolar stops in phonological environments where coda-, onset-, or ambisyllabic syllabification are expected. Normal productions were compared with intersyllable pause productions, in which speakers deliberately paused between syllables. The results show two distinct timing patterns in paused productions: one in which both formation and release of constriction were coordinated with the preceding vowel (coda-syllabification), and another in which formation and release were temporally dissociated (ambisyllabicity). These results are hard to reconcile with the conventional assumption that segmental units are associated with syllables, but can be readily understood in a gestural framework.

1. Introduction

The constructs of segments and syllables underlie a vast range of work in phonological theory and serve as useful heuristics in numerous empirical studies. A fundamental question that arises in regard to these hypothetical units is that of *syllabification*: how are segments associated with syllables? An important theoretical claim (that remains contested) is that in some contexts, an intervocalic consonant may be *ambisyllabic*, associated with both a preceding and following syllable. However, there is not much empirical work on the articulatory consequences of such dual association: what specifically do we expect to observe in actions to form and release the constrictions associated with purportedly ambisyllabic segments? A number of prior studies have compared acoustic and articulatory properties of potentially ambisyllabic consonants to their properties in onset and coda positions, but these studies have reached a variety of conclusions and suffer from various confounds. Other studies designed to elicit metalinguistic intuitions about syllabification have found evidence for ambisyllabic organization, but these studies have not conducted detailed articulatory/acoustic analysis of production data.

The novelty of the current study is that it investigates ambisyllabicity by employing articulatory imaging in a metalinguistic task of intersyllable pause production (henceforth *paused production*), where speakers deliberately pause between syllables. Midsagittal ultrasound and webcam video were used to analyze lingual and labial articulation for a small set of intervocalic stops (/p/, /b/, /t/, and /d/). Target stops occurred as intervocalic singletons (VCV) or as the first member of a cluster (VCCV) in disyllabic words in environments where, on the basis of phonotactic constraints and structural/metrical factors, phonological analyses predict either coda-syllabification, onset-syllabification, or ambisyllabicity. Both normal and paused productions of the target words were elicited. It was hypothesized that evidence for ambisyllabic organization in a subset of the environments would manifest in the temporal dissociation of gestures to form and release the consonantal closure. After the production task, syllabification intuitions were obtained using an orthography-based procedure.

The results from paused production revealed two distinct timing patterns in paused production: (i) in environments with a relevant sequencing constraint (*.CC), a coda-syllabification pattern was observed, in which constriction formation and release gestures were temporally coordinated and proximal to the preceding vowel; (ii) in all other environments, an ambisyllabic dissociation pattern was observed: constriction formation gestures were temporally displaced with the preceding vowel, while releases remained proximal to the following vowel. Unexpectedly, constriction formation movements in ambisyllabic environments sometimes occurred twice, both before and after the intersyllable pause, suggesting that the constriction formation gesture was activated in association with both syllables. I argue that the experimental results are hard to reconcile with conventional theories of phonological organization

in which segments are associated with syllables, but they can be readily interpreted in the gestural framework of Selection-coordination theory (Tilsen, 2016).

The paper is structured as follows: Section 1 reviews conceptual understandings of syllabification and ambisyllabicity in segmental and gestural phonological frameworks, and surveys relevant work from previous empirical studies. Sections 2, 3, and 4 present hypotheses, methods, and results, respectively. Section 5 considers implications for our understanding of phonological organization, provides a theoretical interpretation, and discusses the extent to which the interpretation generalizes to typical modes of production.

1.1 Syllabification and ambisyllabicity in segmental phonology

The notion that speech is comprised of segments (or phones), and that segments are associated with syllables, is undoubtedly a very old idea. One might speculate that it originates with the development of phonographic writing systems and metrically organized poetic verse. English dictionaries from a couple centuries ago, such as Thomas Sheridan's (1790) *A general dictionary of the English language*, presented quasi-phonological transcriptions of words in which syllable boundaries intervene between sounds. Furthermore, the dictionary explicitly depicted ambisyllabic phonological representations, although the term "ambisyllabic" itself was not yet in use. There are numerous entries like the ones below, where an intervocalic consonant or the first consonant of a cluster appeared on both sides of a syllable boundary (—). Statistical analyses show that ambisyllabic representations in the dictionary were motivated not by orthographic doubling but rather by the avoidance of open stressed syllables with lax/short vowels (Ballier & Pouillon, 2013).

ACTIVITY, a1 k — t i1 **v' — v** i1 — t y1
ATROPHY, a1 **t' — t** r o2 — f y1

The earliest use of the term *ambisyllabic* that I have found through online searches is from Swadesh (1936), who stated that the liquid consonant [l] "is ambisyllabic in positions where other cases of vowel plus consonant are ambisyllabic (thus in *balloon*; cf. [əd] in *adopt*)". Notably, Swadesh used the term *ambisyllabic* without defining it, as if presupposing that it would be familiar to readers. We can therefore infer that the concept of ambisyllabicity was in play in phonological analyses prior to this time, although its original use remains elusive.

To understand modern analyses of ambisyllabicity from a segmental phonological perspective, we must examine several factors that are commonly held to govern syllabification of segments. One of these is the *sonority sequencing principl*e (SSP; (Hooper, 1976; Zec, 1995)), which dictates that segments are associated with syllables such that each syllable has a single sonority peak and no sonority plateaus. The SSP is based on the premise that segments are distinguished by their ranking in a sonority hierarchy, from most to least sonorous: vowels, glides, liquids, nasals, fricatives, and stops/affricates. Language-specific phonotactic constraints can also play a role in syllabification, particularly in cases where sonority sequencing constraints do not apply. Some relevant examples are */.tl/ and */.dl/ in English.

A second major factor in syllabification is the *maximum onset principle* (MOP; (Hooper, 1972; Kahn, 1976)), which holds that there is a preference for intervocalic consonants to be associated with the onset of a subsequent syllable as opposed to the coda of a preceding one. In the absence of other factors, the MOP dictates that VCV and VCCV sequences will be syllabified as V.CV and V.CCV, respectively.

A third factor relevant for English and some other languages with metrical stress is the avoidance of open stressed syllables with a lax or short vowel (Hammond, 1997). This accounts for why there are few monosyllabic content words ending with lax vowels like /æ/ or /a/, with exceptions tending to be discourse markers (*yeah*, *nah*) or morphological truncations (*pa*, *ma*). This pattern can also be analyzed as a

constraint against monomoraic stressed syllables; for purposes of brevity I refer to it as a *weight constraint* in this manuscript.

The interplay of the above factors is shown in Table 1, where the rightmost column indicates the relevant factors. For an intervocalic singleton, where no SSP or language-specific phonotactic constraints are applicable, the only relevant factors are the weight constraint (WEIGHT) and maximum onset principle (MOP). The contrast in Table 1 between (a) and (b) shows that in an iambic (ws) metrical context (e.g. *cadet*), where WEIGHT does not apply to the initial syllable, the MOP favors onset syllabification (V.CV), but in a trochaic (sw) context (e.g. *model*), the weight constraint applies and prefers the initial syllable to have a coda. Ambisyllabic structure in this environment (represented V <C> V) satisfies both the MOP and WEIGHT. The trochaic VCV environment is also one in which the stops /t, d/ exhibit flapping, voiceless stops lack aspiration, and liquids may have reduced or absent anterior lingual constriction (e.g. /l/ → [ɬ]).

Table 1. Examples illustrating factors involved in phonological analyses of syllabification

a.	intervocalic singleton (ws)	*cadet*	/k a . d ɛ t/	cf. */k a d . ɛ t/	MOP
b.	intervocalic singleton (sw)	*model*	/m a <d> ɛ l/	cf. */m a d . ɛ l/	WEIGHT & MOP
c.	rising sonority (ws)	*patrol*	/p a . t ɹ o l/	cf. */p a t . ɹ o l/	MOP
d.	falling/plateau sonority (ws)	*advance*	/æ d . v æ n s/	cf. */æ . d v æ n s/	SSP > MOP
e.	language-specific phonotactics (sw)	*meatless*	/m i t . l ɛ s/	cf. */m i . t l ɛ s/	LSP > MOP
f.	rising sonority (sw)	*quadrant*	/k w a <d> ɹ ɛ n t/	cf. */k w a d . ɹ ɛ n t/	WEIGHT & MOP
g.	morpheme integrity	*week+ly*	/w i k . l i/	cf. ʔ/w i . k l i/	INT > MOP

The effects of sonority-related and language-specific phonotactic constraints for syllabification in intervocalic clusters are exemplified in examples (c)-(e). When the cluster constitutes a sonority rise and is not prohibited by a language-specific constraint, onset syllabification is expected (e.g. *pa.trol*), but for clusters with flat/falling sonority or clusters subject to a language-specific constraint, the initial member of the cluster is syllabified as a coda (*ad.vance* and *meat.less*). Note that WEIGHT does not apply in these examples because the initial syllable is either unstressed or contains a tense vowel. These cases show that the SSP and language-specific phonotactics are more influential than the MOP. Example (f) is a cluster in a trochaic context where WEIGHT is a factor (e.g. *quadrant*); here many analyses also posit ambisyllabicity, in order to satisfy both WEIGHT and MOP.

Another factor that may play a role in syllabification is *morpheme integrity*—the preference for segments associated with a morpheme to be associated with the same syllable. This allows for a syllabification of words such as *weekly* as /wik.li/ rather than /wi.kli/ (example (g)), despite the fact that no other factor prohibits onset syllabification. Orthographic doubling is generally not considered to be a relevant factor, although some studies reviewed below have found orthographic doubling to influence syllabification intuitions. In the current study, stimuli with orthographic doubling were categorically excluded and morphologically complex stimuli were avoided as much as possible.

Many formal phonological analyses have explored the interplay of the above factors—i.e. the sonority sequencing principle, the maximal onset principle, language-specific phonotactic constraints, and the stressed syllable weight constraint—often employing ranked or weighted constraints to generate syllabifications that are consistent with intuitions and/or occurrence of alternations that are hypothesized to depend upon syllable association (such as flapping/tapping, deaspiration, and lateral velarization). It is worth emphasizing that despite the apparent certainty in syllabification that may be offered by some phonological analyses, there is no way to infer a ground truth syllabification of segments from what we observe in the acoustic signal. Furthermore, putting aside the desire for parsimony, all phonological patterns that may be described with syllable structure can alternatively be described with purely linear rules. Syllable structure undoubtedly simplifies phonological analyses, in a conceptual sense, but it is not a self-evident property of what is physically observed in speech.

Another point to note here is that modern syllabification analyses typically incorporate syllable subconstituent structure, either as (a) an onset/rime division and a nucleus/coda division within the rime, or (b) moraic subconstituent structure. In both cases, segments in structural representations are linked to subconstituents rather than directly to syllable nodes. This theoretical distinction is mostly orthogonal to the analyses and interpretations that we are concerned with in the current study, and so for the sake of neutrality and simplicity, I omit subconstituent structure from schematic representations throughout this manuscript.

Given the challenges that the results of the current experiment pose to a segment/syllable-based ontology of phonological organization, it is crucial to highlight some of the fundamental suppositions of such analyses. First, it is often presupposed that syllabification is *not* represented in the lexicon (i.e. in long-term memories of word forms). The absence of lexical syllabification is a desideratum of generative phonology in the sense that it requires less information in memory and relocates surface patterning to the output of a grammar. Post-lexical syllabification is also empirically reasonable given that lexical contrasts solely involving syllabification are unattested (Blevins, 1995). However, the absence of syllabification in the lexicon is not in line with modern theories of lexical memory, which allow for phonetic and paralinguistic/contextual detail to be associated with word forms: the prevalence of word-specific phonetic variation (Coleman, 2002; Pierrehumbert et al., 2002) tells us that memory can and does incorporate information well beyond abstract phonemic sequences. This leads to an apparent paradox: if our memories of word forms can be otherwise so rich, why do those memories appear not to encode syllabification?

Second, most theories hold that all word-internal segments are syllabified, i.e. are associated with some syllable (in other words, segments are *exhaustively parsed* into syllables). There are indeed proposals that word-initial or word-final consonants may be unsyllabified (a.k.a. *extraprosodic*; (Liberman & Prince, 1977)); these are often motivated by attempts to develop rules accounting for metrical patterns in languages where metrical strength and/or foot structure appear to depend on the segmental composition of syllables. However, to my knowledge there are no theoretical proposals that word-internal consonants may be unassociated with a syllable.

Third, and almost so general that they go unnoticed, are interrelated presuppositions of the unitariness and discreteness of segments, which hold that entities such as segments are coherent wholes (units) and distinguishable from each other on the basis of their component features. These notions entail that segments are never divided into smaller units that contain subsets of features. In other words, there are no conceptions of syllabification that allow for segments to be split into parts that obtain different relations with different syllables. The same logic applies to timing units in autosegmental theories.

In contrast to the generally accepted propositions discussed above (i.e., absence of lexical syllabification, exhaustive parsing, and unitariness/discreteness), the premise that segments are *uniquely* associated with syllables is contested. Uniqueness of segment-syllable association holds that each segment can be associated with at most one syllable. Ambisyllabic representations violate this premise,

and thus theoretical disagreement regarding ambisyllabicity can be framed as dispute over whether unique association is a necessary property of phonological structure. However, in my opinion, there is a more useful way to interpret the notion of unique association in segmental phonological frameworks, and we can see this through careful examination of the conceptual metaphors / image schemas that underlie such frameworks.

Modern segment-syllabification analyses make an existence presupposition regarding the entities that we are concerned with—segments and syllables. In other words, such analyses take it for granted that (i) speech is cognitively represented as a sequence of segments and (ii) that in the process of producing speech, cognitive representations arise in which segmental units are associated with syllable units. Note that these presuppositions are about entities hypothesized to be present in *cognitive* representations, or perhaps, entities that are useful for descriptions of cognitive states, but not about entities in the physical manifestation of speech. Many authors have distinguished between "phonological" (i.e. cognitive) and "phonetic" (i.e. physically manifested) syllables (Schettino et al., 2015). We should be reluctant to embrace such distinctions because there is no way to identify syllables in physical signals that is non-circular: only by defining *a priori* what a syllable is in acoustic or articulatory signals can "syllables" be "identified"; in the absence of a ground-truth discovery procedure that finds syllables without presupposing their properties, it is a tautology to argue that they exist in physical signals. The same point applies to segments: these entities have never been discovered in physical signals, but they are commonly held to be physical manifestations of cognitive entities. Nonetheless, the distinction between what is cognitive and what is physical is sometimes blurred in analyses of relevant phenomena, and must be kept in mind in interpreting the claims of formal analyses.

It is important to recognize that the cognitive existence presuppositions of modern formal analyses are not logical necessities, but rather, historical contingencies. This is obvious when one compares the conceptual metaphors implied by pre-modern discussions of syllables and segments with those implied by modern formal theories. Careful analysis along these lines (see Lakoff, 1993; Lakoff & Johnson, 2008) requires that we consider exactly which image-schematic entailments an author appears to adopt in using expressions like *syllable* and *segment* (or related terms like *consonant* and *vowel*).

The most relevant image-schematic distinction to consider here is that of *containment* vs. *overlap*, or alternatively, object containment vs. parallel spatial overlap, which crucially relates to whether segmental and syllabic units are image schematically conceptualized as objects/containers located in a single spatial/temporal coordinate, or as objects/events located in parallel spatial/temporal coordinates. In one of the earliest discussions of syllable/segment structure, Pike & Pike (1947) conceptualize syllables as events rather than containers and likewise think of segments as events rather than objects contained in syllables; most importantly, they envision these two sorts of events to be located in distinct, parallel temporal coordinates. To wit, in prefacing their analysis of Mazateco syllables, Pike & Pike (1947) state that syllables and their constituents are "rather like an overlapping series of layers of bricks," and subsequently that "the first consonant of a cluster at the beginning of one syllable may pass partially to the preceding of a beginning syllable" and "the foreigner in speaking the language…[may] err by transferring too much of rather than too little of the consonant to the preceding syllable". These quotations make it clear that the authors conceptualized syllables and segments as events that are associated with periods of time (which is metaphorically a linear space), and so expressions like "pass partially to" and "transferring" were understood as entailing overlap of regions between parallel temporal coordinates (see Fig. 1a).

It is unclear whether Pike & Pike (1947) were describing a phonetic/physical or phonological/cognitive conception of syllables; regardless, they do not appear to imply that segments are object-like units within container-like syllables. Similar sentiments are expressed by other authors, with similar ambiguity in whether the description is cognitive or physical. For example, in discussing words like *being* and *booing*, Trager & Smith (1941:233) say, "...in cases like these, the intersyllabic glide is ambisyllabic (i.e., forms phonetically the end of the first and the beginning of the second syllable), so that these words exhibit a

syllabic structure exactly parallel to that of such words as bidding...", and Smalley (1968:154) points out that it is easy to identify the "crests" of syllables but notes that "it is not always possible to determine an exact syllable boundary. A consonant between two syllables may belong phonetically to both."

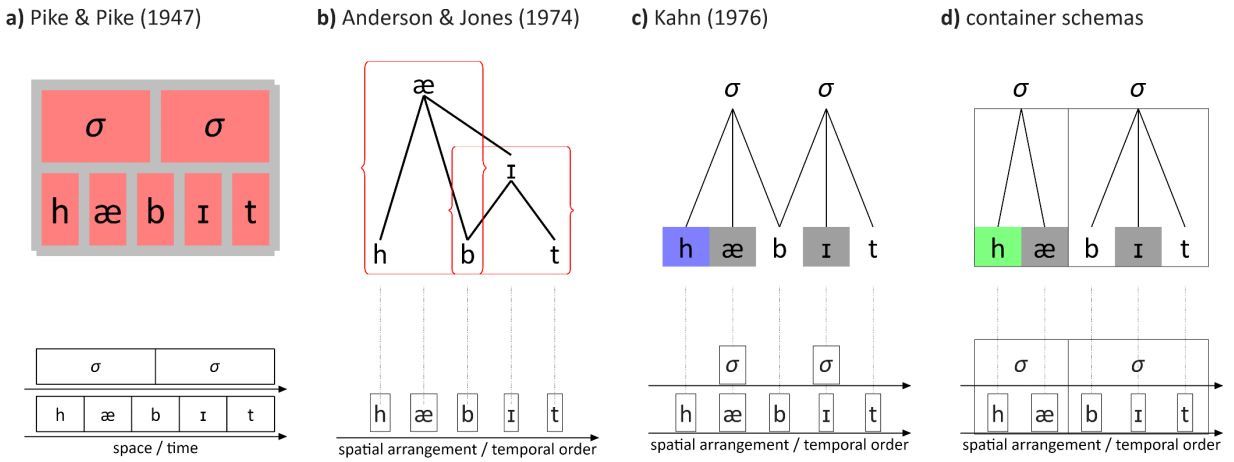

Fig. 1. Image-schematic conceptions of syllabification in segmental phonological theories.

In contrast to the temporal overlap conceptualization, Stampe (1973) employed an object containment schema to conceptualize syllabification. He used metaphoric expressions in which syllables may “carry” segments, segments may be “forced into” a syllable, and various phonological processes occur “within” syllables. On top of this, Stampe never mentioned the possibility of ambisyllabicity, despite almost certainly being aware of it, implying that he rejected the idea on the basis of a strict object-containment schema. This conceptualization derives from a standard entailment of the container image schema: a single object cannot simultaneously be contained within two distinct containers, which follows from our everyday experience with objects and containers.

Despite his use of containment metaphors and implicit rejection of ambisyllabicity, it is noteworthy that Stampe (1973) also frequently used the word *attachment* to describe relations between segments and syllables. This term evokes a different sort of schema—a linkage schema—that has become dominant in modern theories of syllabification. On the basis of this, it is reasonable to infer that Stampe had a schematic blend in mind along the lines of Fig. 1d, in which segments are linked (i.e. *attached*) to syllables, and in which the projection of these entities to a temporal coordinate does not allow for the period of time associated with a segment to be located in the periods associated with two distinct syllables. In the more modern analysis of Jensen (2000), containment entailments are explicitly the motivation for prohibiting ambisyllabic representations. Jensen maintains that each unit in the prosodic hierarchy is uniquely contained by a higher-level unit, and therefore schemas in which a given segment is linked to two syllables are ruled out. This conceptualization blends containment with linkage, such that linkage relations, vis-a-vis their vertical orientation, entail that lower-level objects are uniquely contained within the higher-level objects that they are linked to.

The earliest linkage-based formalization of ambisyllabicity that I am aware of, Anderson & Jones (1974), employed dependency graphs without containment as in Fig. 1b, where vertical position encodes directed graph structure. It is important to note that Anderson & Jones (1974) did not conceptualize syllables as the sorts of entities that could be nodes of the graph; instead, they viewed syllables as subgraphs, allowing for different subgraphs to share nodes. Hence in their formalization an ambisyllabic structure is one in which a consonant is a dependent of two different vowel nodes.

Kahn (1976) was the first theoretical analysis to adopt a linkage schema in which syllables are nodes dominating segments, shown in Fig. 1c. Kahn readily accepted structures in which the same segment could

be linked to two syllables, which is not entirely unlike the multiple dependency between consonants and vowels allowed by Anderson & Jones (1974). The crucial difference is that Kahn appeared to envision syllables to be projected to a distinct temporal/spatial coordinate from the segments, whereas Anderson and Jones did not conceptualize syllables as the sorts of entities that are subject to a temporal projection. The fact that Kahn rejected containment entailments is evident from his analogy between syllables and mountains: "one might consider mountain ranges; the claim that a given range consists of, say, five mountains [syllables] loses none of its validity on the basis of one's inability to say where one mountain [syllable] ends and the next [syllable] begins" (bracketed expressions added). Elaborating on this with a bit of poetic license, we can say that Khan did not insist that any particular segment (a part of a mountain, such as a foothill) belongs uniquely to a particular syllable (the mountain itself), since mountains within a range are not container-like—intervening valleys/foothills are not obviously "contained" solely within one particular mountain in the range. Selkirk (1982) is another example of the linkage-without-containment conceptualization advanced by Kahn.

It is noteworthy that Kahn erroneously argued against Anderson & Jones (1974) on the basis that their formalization allowed for temporal discontinuity in syllable-segment association, when in fact their projection constraints prohibited any such discontinuity. In my opinion, the more fundamental motivation behind Kahn's innovation was the notion that there is a separate but parallel temporal coordinate for the projection of syllable nodes: syllable and segment boundaries simply do not obtain on the same temporal coordinate. Interestingly, Kahn (1976) also expressed an intuition that ambisyllabic organization depends on speech rate/style. He suggested that in relatively slow/careful speech (which he described as "sci-fi movie robot speech"), syllabification of an intervocalic consonant is dictated by onset maximization (e.g. *ha.bit*), but in relatively fast/casual speech, the ambisyllabic (V<C>V) organization occurs.

What the reader can take away from the above analysis is that, in the context of segmental phonology, modern theoretical disagreement regarding ambisyllabicity can be reduced to a difference in image-schematic preferences for reasoning about syllabification, amounting to whether a theory does or does not conceptualize syllables as containers of segments. Theories of organization that impose containment (Jensen, 2000; Stampe, 1973) prohibit ambisyllabicity; theories that conceptualize syllabification purely as linkage (Anderson & Jones, 1974; Kahn, 1976; Selkirk, 1982) do not.

1.2 Syllabification and ambisyllabicity in gestural phonology

Gestural phonology refers to phonological frameworks in which articulatory gestures are hypothesized to be the fundamental entities of phonological organization. The concept of an articulatory gesture originates in the framework of Articulatory Phonology (Browman & Goldstein, 1992) and is based on the mathematical model of Task Dynamics (Kelso et al., 1986; Saltzman & Munhall, 1989), but to avoid confusion it is important to recognize that there are several related uses of the term *gesture*. In the most precise and constrained use, gestures are dynamical *systems* and their state variables are activations. When gestural systems have high levels of activation, they exert forces on other systems, *vocal tract systems*, which control geometric states of the vocal tract. The term *gesture* is often metonymically extended to refer to periods of time in which gestural systems are highly active (technically, these are *gestural activation intervals*), and even more generally *gesture* is used to refer to the movements that may occur as a consequence of high activation (see Tilsen (2018) for further discussion of this polysemy).

Gestural frameworks do not take for granted the foundational premises of syllabification, i.e., (i) that speech is comprised of segments and (ii) that segments are associated with syllables. There is indeed a diverse range of perspectives on role of these constructs from a gestural perspective, and perspectives have evolved over recent decades. Specifically regarding ambisyllabicity, Browman & Goldstein (1990), in addressing the phasing of gestures in VC(C(C))V sequences, state that "the leftmost consonantal gesture of a consonant sequence intervening between two vocalic gestures is associated with both vocalic

gestures". This same interpretation of ambisyllabicity was adopted in Gick (2009): "certain intervocalic gestures are crucially phased to both the preceding and following vowel gestures" (2009: 224).

A somewhat different take on ambisyllabicity appears in Browman & Goldstein (1992), in a discussion of the lack of aspiration in stops in words like *rapid*. They noted that some phonological analyses (e.g. Kahn 1976) attributed the lack of aspiration to ambisyllabicity, but they argued that, to the contrary, the aspiration pattern arises from a gradient reduction of the magnitude of a glottal opening gesture, due to stress and position in word. The same logic was used to explain flapping in this environment. Browman and Goldstein (1992) notably did not reference the dual phasing interpretation from their earlier analysis, but it is nonetheless fair to take the Browman & Goldstein (1990) dual-phasing interpretation as one possible gestural account of ambisyllabicity.

An important consideration in any gestural analysis relates to whether the release of a consonantal constriction is viewed as an indirect consequence of the deactivation of the relevant constriction formation gesture, or as a distinct release gesture, on par with the formation. In the former view, the change of vocal tract state that corresponds to constriction release occurs because the constriction formation gesture becomes inactive, and in the absence of its influence, the relevant tract variable is subject to influences of other gestures, if active, or otherwise to a default control system (called the *neutral attractor* (Saltzman & Munhall, 1989)). In the latter view, also known as the *split-gesture* hypothesis (Browman, 1994; Nam, 2007), the release movement is the consequence of an active release gesture. Nam (2007) maintained that for a singleton onset consonant, both closure and release gestures are in-phase coupled to the vocalic gesture, but anti-phase coupled to each other. Empirical support for the split gesture hypothesis and these phasing relations was found in Tilsen (2017).

Several alternative proposals for gestural organization in ambisyllabic environments are shown in Fig. 2, where coupling graphs and their associated gestural scores are presented. Proposal (a) corresponds to the dual-phasing account (Browman & Goldstein, 1990). (b) is a split-gesture analysis of ambisyllabicity from Bradley et al. (2023), where in addition to the coupling relations that constriction (clo) and release (op) gestures are hypothesized to obtain in prototypical onset consonants, both gestures are anti-phase coupled to the preceding vocalic gesture. This parallels the dual phasing interpretation but extends it to a split-gesture model. Proposal (c) is a simpler possibility akin to the analysis of geminate consonants in Burroni (2022), in which constriction is anti-phase coupled to the preceding vowel, release is in-phase coupled to the following vowel, and constriction and release are anti-phased.

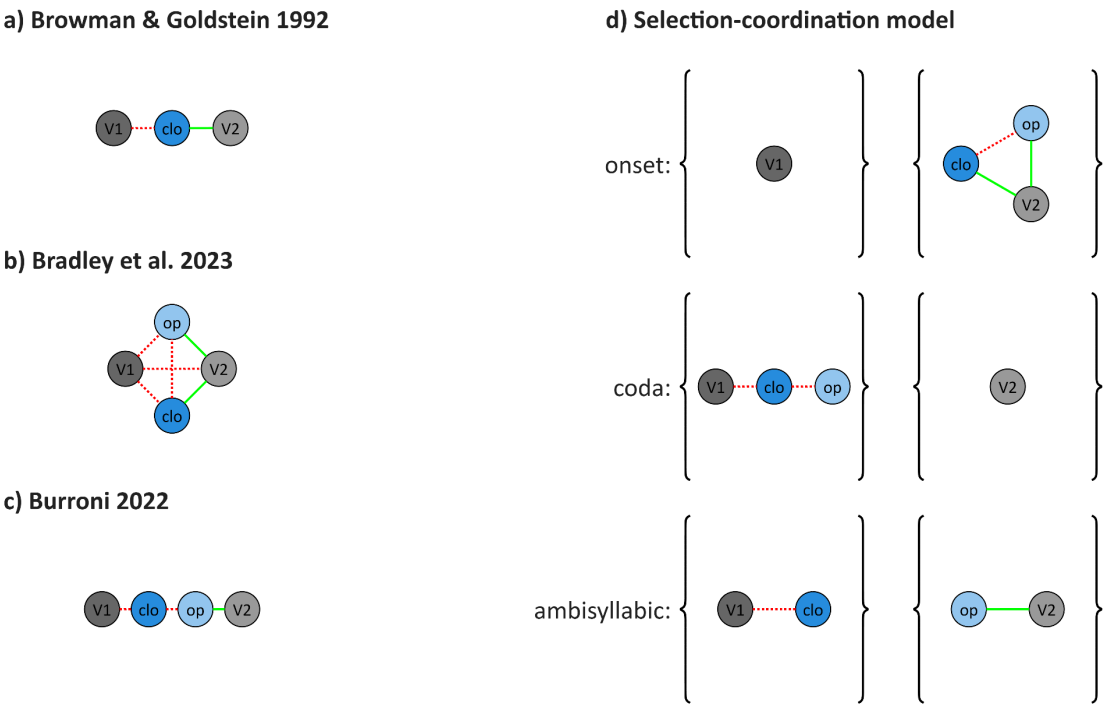

Fig. 2. Interpretations of gestural organization. Green lines: in-phase coupling; red lines: anti-phase coupling. (a)-(c) Gestural interpretations of ambisyllabicity. (d) Syllabification in the selection-coordination framework. Curly braces indicate sets of coordinated gestures, with sets themselves being competitively selected.

A substantially different theory of gestural organization in the Selection-coordination framework (Tilsen, 2013, 2016, 2022) maintains that phasing/coupling of gestures is restricted to relatively small sets, which are more prototypically the size of moras in early development but can expand to be syllable-sized in later development. A crucial restriction of this model is that gestures that are organized into different sets *cannot* be phased/coupled. Rather, intergestural timing between gestures in different sets is governed by feedback mechanisms that regulate the sequential selection of the sets themselves. These feedback mechanisms are held to rely more on internal predictive feedback for adults, but in early development they rely on external sensory feedback. Fig. 2d shows how the constriction formation and release gestures of an intervocalic singleton consonant would be organized in ways that reflect onset syllabification, coda syllabification, and ambisyllabicity. The crucial claim of this model is that ambisyllabicity is a pattern of gestural organization in which formation and release gestures are selected with different sets. The hallmark of ambisyllabicity is thus that the formation and release gestures that are typically selected together and coordinated are dissociated, in the sense that they are selected in separate sets and therefore their relative timing is determined by feedback processes, rather than by a coordinative phasing mechanism.

1.3 Experimental research on ambisyllabicity

Much of the theoretical work in the segmental phonology of ambisyllabicity has been based on the intuitions and auditory impressions of the researchers who conducted it. What do we know about ambisyllabicity from empirical studies of speakers without linguistic training? A useful way to classify such research revolves around whether or not the tasks performed by participants require metalinguistic reasoning, i.e. whether the task explicitly evokes the notions of segments and/or syllables for speakers (or,

more neutrally, the notion that the sounds of words can be broken into parts). First, let us consider acoustic and articulatory studies of production that are not metalinguistic in this way, i.e., studies that simply examine phonetic properties of sounds without bringing attention to their organization.

Ambisyllabicity studies using non-metalinguistic production tasks generally employ a common logic of interpretation that goes as follows: (i) a given consonant has acoustic/articulatory properties in a position in which it is assumed to be an onset, and some different properties in position in which it is assumed to be a coda; (ii) in an environment in which the syllabification of the consonant is ambiguous, (a) if the properties of the segment are similar to its properties in onset position, then it is syllabified as an onset, (b) if they are similar to its coda properties, then it is syllabified as a coda, and (c) if they are otherwise—i.e. a mixture of, or intermediate between, or merely different from the onset and coda properties—then the segment may be ambisyllabic, or alternatively, it may be an allophone that is specific to word-medial position. The logic is expressed, for example, in Turk (1994): "ambisyllabic consonants are predicted to share phonetic characteristics of both syllable-initial and syllable-final consonants".

An early example of this *property similarity logic* can be found in Krakow (1989), who investigated the articulatory characteristics of nasal consonants in word-initial (#_V), word-final (V_#), intervocalic (V_V), and cluster (Vl_V) environments in English. In comparing temporal and kinematic measures of velar and labial movements from two speakers between these environments (e.g. *pa made* (onset) vs. *palm aid* (coda) vs. *pomade* (ambiguous)), Krakow observed mostly cases of intervocalic segments that exhibited onset-like or coda-like properties; cases where the segment exhibited a combination of onset-like and coda-like properties were fewer. She acknowledged that it was not possible to draw strong inferences from her limited sample size, and did not conduct a statistical analysis of measurements. Another example of property similarity logic is from Turk (1994), who used speech imaged with x-ray microbeam to compare upper lip kinematics between environments (e.g. *captor* (coda) vs. *repair* (onset) vs. *leper* (ambiguous)). Turk notably considered constriction formation and release movements separately in her analysis, and concluded that the ambiguous segments patterned like codas.

Gick (2009) analyzed electromagnetic articulograph data for lingual and labial gestures of /l/, /w/, and /j/ from three English speakers in onset vs. coda vs. ambiguous environments (e.g. *hall hotter* (coda) vs. *ha lotter* (onset) vs. *hall otter* (ambiguous)). On the basis of statistical comparisons of various properties, he concluded that in the ambiguous environment, the tongue tip gesture of /l/ and the labial gesture of /w/ resyllabify as an onset, but the tongue body gestures of these segments remain coda-like; he argued that this dissociation of the more posterior (or vowel-like) and anterior (or consonant-like) gestures of a complex segment is a manifestation of ambisyllabicity.

Scobbie & Wrench (2003) used electropalatographic data and analyzed the rate of /l/ vocalization (i.e. velarization /l/ → [ɬ]), operationalized via a threshold degree of absence of linguopalatal contact. They found that vocalization rates in ambiguous environments differed between subjects, some being similar to rates in coda environments, others similar to rates in onset environments. Lee (2025) used ultrasound to image /ɹ/ articulation in onset, coda, and ambiguous environments. He observed that tongue contours temporally transitioned from an intermediate shape, between onset-like and coda-like, to an onset-like shape. From this he concluded that "ambisyllabicity lacks a stable articulatory correlate and functions primarily as a theoretical construct" and suggested that ambisyllabic retroflexes "may be regarded as constituting an independent allophone, characterized by intermediate tongue contours that are distinct from onset and coda allophones".

In addition to articulatory studies, there are a number of acoustic studies that employ the property similarity logic. On the basis of similarity to acoustic patterns associated with codas, specifically preceding vowel nasalization and consonant devoicing, Durvasula & Huang (2017) argued that ambiguous nasal consonants are codas rather than ambisyllabic. Likewise, based on similarity of effects on pitch and duration of preceding vowels, Nesbitt (2018) argued that the ambiguous stops are codas rather than multiply linked segments. A somewhat different conclusion was reached by Lee & Seo (2019), who found

no temporal or spectral differences between purportedly ambisyllabic and non-ambisyllabic intervocalic /l/, along with codas and onsets, and therefore argued that purportedly ambisyllabic /l/ were a third allophone of /l/. Strycharczuk & Kohlberger (2016) found that word-final prevocalic /s/ in Spanish exhibits neither onset-like nor coda-like duration and argued that these segments are partially resyllabified in a way that resembles ambisyllabic organization.

In sum, the collection of articulatory and acoustic studies of ambisyllabicity that have employed a property similarity logic have led to no consensus regarding their phonological organization. In my opinion, this is partly consequence of the property similarity logic itself, which has several shortcomings. First, the logic takes for granted that segments have position-specific properties, i.e. that there are properties of a given segment in onset position and different properties in coda-position. This assumption is questionable given findings from Sproat & Fujimura (1993), whose study of x-ray microbeam from five speakers focused specifically on English /l/. They elicited this segment in a wide variety of environments characterized by differing prosodic/morphosyntactic boundaries, including an intervocalic one with no boundary (*Mr. Beelik*). They compared the means of various durational and kinematic measures of tongue dorsum movement across these environments. They did not specifically draw inferences regarding ambisyllabicity; rather, they observed that articulatory properties varied gradiently across environments rather than falling neatly into onset-like and coda-like groups.

Second, with the exception of Turk (1994), articulatory studies have not examined both constriction formation and release kinematics—dissociating these may provide crucial insight to phonological organization. Indeed, as hypothesized in Fig. 2d, ambisyllabicity may correspond to a dissociation of constriction formation and release gestures.

Third, in many cases, for purposes of comparison, studies have assumed that segments in particular environments are unambiguously syllabified as onsets or codas. For example, Krakow (1989) assumed that /m/ is unambiguously a coda or onset, in *palm#aid* vs. *pa#made*, respectively. This assumption derives from the notion that word boundaries prevent resyllabification or ambisyllabicity, but the assumption cannot be verified due to the absence of a ground truth procedure for assessing syllabification, and it is contradicted by analyses that allow resyllabification across word boundaries (see Strycharczuk & Kohlberger, 2016). Indeed, it is plausible that in rapid casual speech, or in laboratory tasks where speakers repeat a small set of phrases, segments may be syllabified across word boundaries. Note that the intuition that ambisyllabicity only occurs in faster speech (Kahn, 1976) is another instance of this general idea, i.e. that syllabication may be variable.

Fourth, all experimental designs based on property-similarity logic necessarily introduce environmental confounds. For example, in Krakow (1989) it is assumed that the only difference between an ambiguous environment (such as *pomade*) and presupposed onset/coda environments (*palm aid* / *pa made*) is syllabification, when in fact the absence of a word boundary in *pomade* is a confound. Similarly, in Turk (1994), the comparison between *captor* (coda), *repair* (onset), and *leper* (ambiguous) assumes that the phonological SSP and iambic metrical environment enforce coda/onset syllabifications, but these factors are also confounds: the properties of a consonant in an ambiguous environment may be influenced by its metrical context and the absence of an adjacent consonant, rather than its syllabic organization *per se*. These sorts of confounds are unavoidable because morphosyntactic, metrical, and sonority-based factors are the hypothesized causes of differences in syllabification, and hence these additional factors necessary differ between the environments used in property-similarity analyses.

Finally, there is a shortcoming that relates specifically to drawing inferences about gestural organization from articulatory timing. In general, if a constriction formation gesture is timed more reliably (i.e. with less variance) in relation to events associated with a preceding/following vocalic gesture, it might be inferred to be organized as a coda/onset, respectively. However, detecting effects from this sort of variance comparison requires substantial statistical power, and ideally the comparison should be conducted with productions made across a range of speech rates. Prior articulatory studies of

ambisyllabicity have not pursued rate manipulations of this sort. In some ways, the metalinguistic manipulation of paused production employed in the current study can be viewed as an extreme form of a speech rate manipulation, where rate is locally slowed in the vicinity of the intervocalic environment.

Moving beyond analyses of speech elicited in a standard mode of production, there are other studies of ambisyllabicity that have made use of metalinguistic tasks, in which participants are confronted with the notion that words may be broken into parts. These studies have generally been interested in the effects on syllabification intuitions of factors such as orthographic doubling, metrical context, preceding vowel tense/lax status, and/or consonant type (often liquid vs. nasal vs. oral stop). The general pattern that emerges from such studies is that ambisyllabicity intuitions are more likely when there is orthographic doubling, when the preceding syllable is stressed, when the preceding vowel is lax/short, and when the consonant is a liquid or nasal, as opposed to a stop.

Perhaps the earliest example of a metalinguistic study of ambisyllabicity is the doubled-syllable production task of Fallows (1981), where English-speaking children from two different age groups heard a disyllabic word produced by the experimenter and repeated the word in two ways, one with the first part doubled (e.g. *habit* → *hahabit*), another with the second part doubled (e.g. *habit* → *habitbit*). When doubling of the target consonant occurred in both the first and second parts, ambisyllabicity was inferred. Preceding stress and preceding lax vowels were found to increase the likelihood of ambisyllabicity intuitions. Notably, Fallows (1981) also mentions conducting an intersyllable pause production task, in which the first part of the word was spoken, followed by a pause, followed by the second part (i.e. *hab…it*), but she did this only for an unspecified subset of the words, and only to resolve ambiguities arising from transcriptions of doubled-syllable productions.

A different method in Treiman & Danis (1988) used a syllable reversal task in which the experimenter produced a multisyllabic word and the participant responded by producing the non-initial syllables, followed by a pause, followed by the initial syllable (e.g. *habit* → *bit…ha*). Participants were trained to do this in examples with morphological breaks (e.g. *kidnap* → *nap…kid*), and syllables were never explicitly mentioned. Responses were categorized according to whether, in the auditory impression of the experimenter, the target phoneme was "placed" before the pause, after, or both, the latter of which was taken to reflect an ambisyllabic parsing. Ambisyllabic parsings were most frequent with preceding stress and orthographic doubling, and were influenced by segment class (liquid, nasal, or oral stop) and preceding vowel type (tense or lax). A notable limitation of the analysis is that the auditory impressions of the experimenter are segmental and therefore cannot distinguish constriction formation and release actions. Nonetheless, the findings do provide evidence that speaker syllabification intuitions are systematically biased by various orthographic and phonological factors.

A primarily perceptual task (called the *pause-break* task) was employed in Derwing (1992), where listeners selected the "most natural" "breaking" of disyllabic auditory stimuli with an intersyllabic pause, i.e. they heard *ha…bit* vs. *hab..it* vs. *hab…bit*. The task was applied to five different languages. The stimuli were presumably recorded by experimenters, but the authors provided no detail on the guidelines for such productions. The results for English were generally consistent with those of Treiman & Danis (1988), finding orthographic doubling, metrical context, preceding vowel tense/lax status, and consonant class to be influential for predicting ambisyllabic behavior. The authors did not analyze the acoustic properties of their pause-break stimuli, so we cannot know if purported onset/coda syllabified stimuli did in fact acoustically reflect the absence/presence of any consonantal information associated in the initial syllable, nor can we know whether the coda of the ambisyllabic stimuli (*hab…bit*) was released. Another issue is that the pause-break task places high demands on working memory: participants must retain three similar stimuli in mind before selecting the most natural one; these memories may interact and change over the period of time in which the decision is made.

There are more examples of alternative choice tasks like the one above. In addition to their syllable reversal task, Treiman & Danis (1988) conducted a binary-choice task in which participants were given two

orthographic choices for the syllabification of a written word, with the syllable break indicated by a slash, and were asked to circle the better one, i.e. circle either *ha/bit* (onset response) or *hab/it* (coda response). Notably there was no ambisyllabic option. Ishikawa (2002) applied a binary-choice paradigm to both English and Japanese speakers with English words and non-words. A similar approach (Elzinga & Eddington, 2014) involves analysis of word-part selection of the first and last parts of a word, e.g. select the first part of *habit*: *ha* or *hab*, and select the second part: *bit* or *it*. The results in all cases are generally similar, indicating systematic biases of orthographic and phonological factors.

Perhaps the closest analogue to the task in the current study is Ishikawa (2002), who, in addition to conducting a syllable reversal task, conducted a inter-syllable pause task in which participants listened to disyllabic stimuli and then produced them in two parts, by inserting a relatively long pause (approximately two seconds). The productions were not acoustically analyzed, but rather categorized as onset-syllabified, coda-syllabified, or ambisyllabic, based on the auditory impressions of the author. Both English- and Japanese-speaking participants were examined, and both words and non-words were used as stimuli. As with previous studies, orthographic and phonological factors biased the syllabifications inferred from auditory impression.

What can we conclude from the various metalinguistic studies above? The fact that there are systematic biases on metalinguistic behavior relating to syllabification and ambisyllabicity (i.e. influences of orthographic doubling, stress, preceding vowel tense/lax status, consonant identity, phonotactic constraints for clusters) tells us that there is some knowledge that speakers have that can result in consonants being "associated" (in an atheoretical sense) with both a preceding "part" of a word and a following "part". There are several limitations that arise, however, in interpreting the results of metalinguistic tasks. For studies that use orthographic stimuli, influences on behavior may arise from orthographic factors that are not strictly speaking structural in origin, such as graphemic frequency. For studies employing multiple or binary choices, participants are forced to discretely resolve any uncertainty they might have. For studies involving reversed, doubled, or paused production, analyses have relied solely on auditory transcriptions or impressions of the experimenters—no studies employing a metalinguistic production task have employed acoustic or articulatory measurements.

More generally, the analyses and interpretations of metalinguistic studies have without exception taken for granted the notion that speech is comprised of segments that are associated with syllables. This is problematic because these hypothetical constructs are effectively presupposed to govern behavior, resulting in a circular logic of interpretation. From an articulatory perspective, consonantal segments are not monoliths, but rather are associated with constriction formation and release actions that may be coordinated with one another and with preceding or following actions. Auditory transcriptions cannot reliably probe the temporal or kinematic characteristics of these actions. In a sense, the current study addresses these shortcomings by applying articulatory analysis to productions in a metalinguistic task.

2. Hypotheses and Predictions

The hypotheses considered in this study relate to syllabification, or in gestural terms, the organization of constriction formation and/or release gestures relative to vocalic gestures. The formation/release gestures examined here are associated with bilabial and alveolar voiced and voiceless stop consonants (p, b, t, d), and predictions focus on timing of these gestures in intersyllable pause production (*paused production*). Stops were elicited as singletons and as the first member of two-consonant clusters. For the clusters, environments with and without a sequencing constraint were examined (no such constraints apply for intervocalic singletons). The environments were crossed with two metrical contexts: trochaic (sw, ˈσσ) and iambic (ws, σˈσ), which differ in the applicability of the weight constraint (see section 1.1). This results in six structural environments, exemplified in Table 2 below.

Table 2. Phonological environments and predicted syllabification

		sequencing constraint	no sequencing constraint	
		*V.CCV	VCV	V.CCV
weight constraint	trochaic sw: ’σσ	1 *ab.sent*	3 *ha<b>it*	3 *go<b>let*
no weight constraint	iambic ws: σ’σ	1 *ob.serve*	2 *a.bout*	2 *a.broad*

Of the six environments represented in Table 2, phonological analyses (section 1.1) predict three syllabification patterns. Coda syllabification of the target consonant (1) is expected in clusters when a sequencing constraint applies (*V.CCV), regardless of whether a weight constraint applies (e.g. *ab.sent* and *ob.serve*). Onset syllabification (2) is predicted when no sequencing constraint or weight constraint applies, i.e. in iambic VCV and V.CCV environments (e.g. *a.bout* and *a.broad*). Ambisyllabicity (3) is predicted in the absence of a sequencing constraint but in the presence of a weight constraint, i.e., in trochaic forms without a sonority-based or language-specific phonotactic constraint (e.g. *habit* and *goblet*).

To infer whether constriction formation and release gestures are temporally organized in a coda-like way, an onset-like way, or are dissociated, we examine whether in paused production these gestures are more closely aligned in time to the acoustic offset of V_1 (coda coupling) or the acoustic onset of V_2 (onset coupling), which are proxies for vocalic gestural landmarks. Note that our focus on relative timing and use of acoustic reference points from vowels are consequences of stimulus design constraints discussed further in Section 3.2. Fig. 3 schematizes these gestural timing diagnostics. Unshaded intervals are acoustic segments; filled boxes are gestural activation intervals, i.e., periods of time in which gestural systems are highly active; below the activation intervals are time series of vocal tract system states (tract variables), i.e. changes in vocal tract geometry that are driven by activation of gestural systems. The schema only depicts patterns for a generic intervocalic (VCV) environment, but the diagnostics apply to the gestures associated with the target consonant in clusters as well.

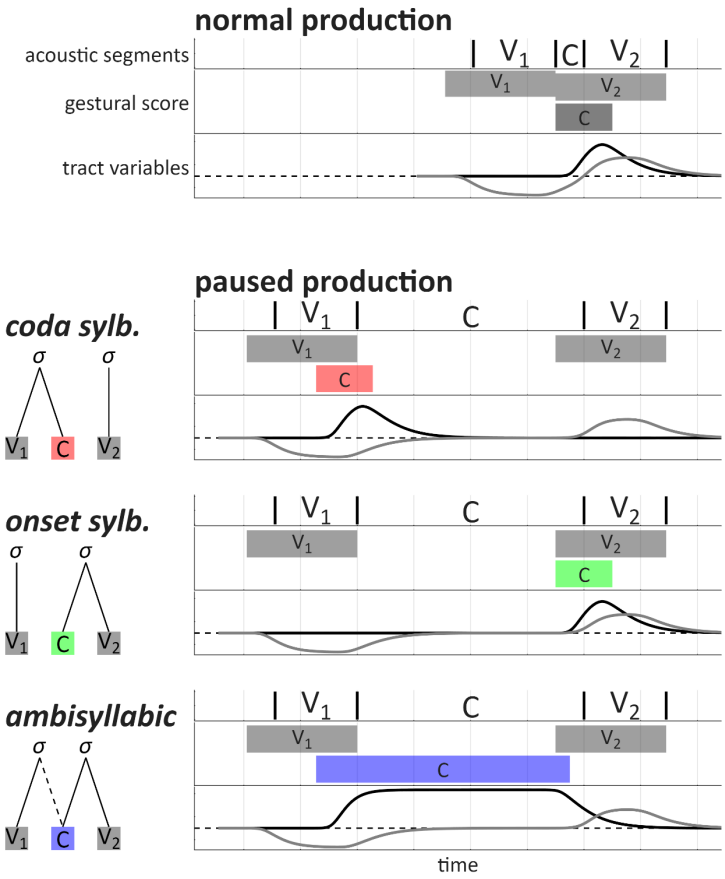

Fig. 3. Syllabification/coupling diagnostics for an intervocalic singleton. Topmost panel: normal production; other panels: coda-, onset-, and ambisyllabic-syllabification in paused production. Each panel depicts from top to bottom: acoustic segmental intervals, gestural activation intervals, and simulated tract variables (black lines: consonant-related tract variable; gray lines: vowel-related tract variable).

As shown in Fig. 3, in paused production coda-syllabification is expected to result in a temporal proximity of both constriction formation and release gestures to the preceding vowel. Conversely, onset-syllabification predicts temporal proximality of both gestures to the following vowel. The primary hypothesis of the study is the *formation-release decoupling* hypothesis, which holds that in ambisyllabic environments, constriction formation is coupled to the preceding vowel and release is coupled to the following vowel, effectively resulting in an extended period of closure. An alternative/null hypothesis is *gestural-segmental integrity*, which holds that formation and release gestures associated with the same segment are always coupled and therefore never associated with different vowels—in other words, no ambisyllabic dissociation patterns will be observed, and gestural organization will always appear onset-like or coda-like.

One important consideration in evaluating empirical patterns against these predictions is the possibility of systematic variation in temporal alignment beyond the variation that may arise from environmental effects. Variation in temporal patterns may occur between consonants and/or lexical items within speaker; it may also occur between speakers. Variation could even occur between repetitions, such that inferred coupling/syllabification patterns may differ for different tokens of the same word form produced by the same speaker. We make no prior hypotheses about such forms of variation, but do assess whether such variation is present.

An additional hypothesis considered here is that metalinguistic intuitions reflect knowledge of gestural organization. To test this, participants performed an orthography-based syllabification task after the production task. This hypothesis predicts that variation in articulatory timing patterns will correlate with variation in metalinguistic intuitions.

Before proceeding to describe the methods, let us address an objection that some readers may have to drawing general conclusions about phonological organization from the results of this study. This *unnaturalness objection* is as follows: one cannot draw inferences about phonological organization in typical speech production from paused production, because paused production is an unnatural form of speech. There are three reasons this objection is unwarranted. First, the unnaturalness objection presupposes that typical production and paused production are governed by different systems, when a far more parsimonious and realistic view recognizes that both modes of production make use of the same control system, perhaps differing in its parameterization or operating in different regimes. Second, the assertion that there is a category of {natural speech} distinct from {unnatural speech} is untenable: all speech varies according to many contextually related factors (i.e. tempo, rhythmicity, hyper-/hypo-articulation, formality, degree of attention to sensory feedback, paralinguistic goals, audience, etc.), and so any distinction between natural and unnatural speech is a necessarily arbitrary one, based on a false dichotomy. Third, paused production does occasionally occur in everyday contexts: a speaker might say: “my toe is bro…ken”, purposefully pausing between syllables as a form of emphasis or to communicate paralinguistic meaning. In addition, since formal phonologists have explicitly used intuitions from imagined/subvocally rehearsed paused production to infer syllabification patterns (e.g. Kahn, 1976), it seems fair to entertain the possibility that empirical study of paused production with non-linguistically-trained speakers may provide useful insights.

3. Method

3.1 Participants and tasks

Thirteen native speakers of English participated in a one-hour experimental session. One participant was excluded from all analyses because submental facial hair interfered with ultrasound imaging. Each participant performed 345 trials of a production task, resulting in a total of 4140 trials. During this task they were seated in front of a computer monitor and were recorded with midsagittal ultrasound (Telemed ArtUS; EchoWave II, 80-95 Hz) and a webcam (Logitech Brio, 90 Hz, 1280x720 pixels) positioned on top of the monitor (approximately 0.75 m from their face). The ultrasound probe was stabilized with ALPHUS probe holder (Chen et al., 2024). Ultrasound-synchronized audio was recorded at 44100 Hz with a shotgun microphone approximately 1 m from the participant; webcam-synchronized audio was recorded at 44100 Hz with the built-in webcam microphone.

The production task was organized into 15 blocks. Within each block all target words were elicited once, in random order, in the same mode of production: intersyllable pause production (*paused production*) or normal production. Block conditions were ordered in a repeating three-block cycle of paused, paused, normal. Thus two paused productions were elicited for every normal production; this was done to obtain more samples of paused productions, which in pilot testing were observed to have higher temporal variability. On each trial, the participant was cued with the orthographic form of the target word. They had 2.25 s to finish producing the word, after which time it disappeared. Before data collection began, participants practiced three trials in each production condition. For intersyllable pause production, participants were instructed as follows: “when each word appears, say it out loud, but pause between the two syllables”. Participants were shown all of the target words during the instructions to check that they were familiar with those words.

After the production task, participants performed a metalinguistic syllabification intuition task. This consisted of 23 trials, one for each target word. On each trial, the participant was shown an orthographic representation of the target word in which vertical lines were interposed between and across the center of all graphemes, as in Fig. 4 below. Participants were instructed: "click on the line indicating the break between the syllables. If the sound represented by a letter is part of both syllables, you can click on the letter". The response was coded according to which vertical line was closest to the position of the mouse when the participant clicked. Note that participants were unaware that they would be performing this syllabification intuition task during the preceding production task.

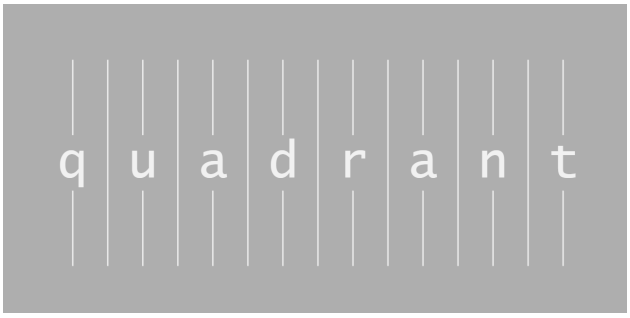

Fig. 4. Example syllabification intuition task stimulus.

3.2 Stimuli

Target words were selected as follows. First, candidate disyllabic word forms in the CMU pronunciation dictionary containing the pattern V1x|V0 Cx (C) V1|V0 were identified (V1x: /æ/ or /a/ with primary stress, V1: vowel with primary stress, V0: /ə/, Cx: {/p, b, t, d/}, C: any consonant). The target consonants (Cx) were restricted to labial and alveolar voiced/voiceless stops because the constriction actions associated with these sounds are relatively straightforward to analyze from processing of ultrasound/webcam videos; only four consonants were examined to augment statistical power. The stressed vowels in the first syllable were restricted to low vowels to promote larger lingual and labial movements in Cx constriction formation after stressed vowels, and because these vowels are considered lax in phonological analyses and therefore induce the weight constraint. Second, candidates with graphemic doubles for the target consonant (e.g. *attic*, *apple*) were excluded, since the doubled grapheme might bias production and complicates analysis of metalinguistic syllabification intuitions. Third, candidates with inflectional morphology (e.g. plural, past, progressive) were excluded.

The remaining candidates were coded according to stress pattern of the disyllable (i.e. strong-weak (trochaic, ’σσ) or weak-strong (iambic, σ’σ)), and presence/absence/non-applicability of a sequencing constraint, defined as: (i) sequencing constraint (*V.C_1C_2V), where the sonority of C_1 is greater than or equal to the sonority of C_2 or a language-specific constraint applies; (ii) non-applicable for intervocalic (VCV); and (iii) no sequencing constraint (V.C_1C_2V), where the sonority of C_1 is less than the sonority of C_2 and no language-specific constraint applies. The coding was used to create 24 lists (4 consonants x 3 sequencing constraint contexts x 2 stress patterns).

The 50 candidates with the highest log frequencies (obtained from the Google n-gram corpus) in each list were exported for manual inspection. For each list, the candidates were reviewed and selected with the following desiderata: avoiding morphologically complex words, avoiding words with multiple pronunciation variants (e.g. *data*: [deirə] ~ [dærə]), avoiding words with initial consonant clusters, avoiding words that may be unfamiliar to participants. This procedure resulted in the stimuli shown in Fig. 5 below.

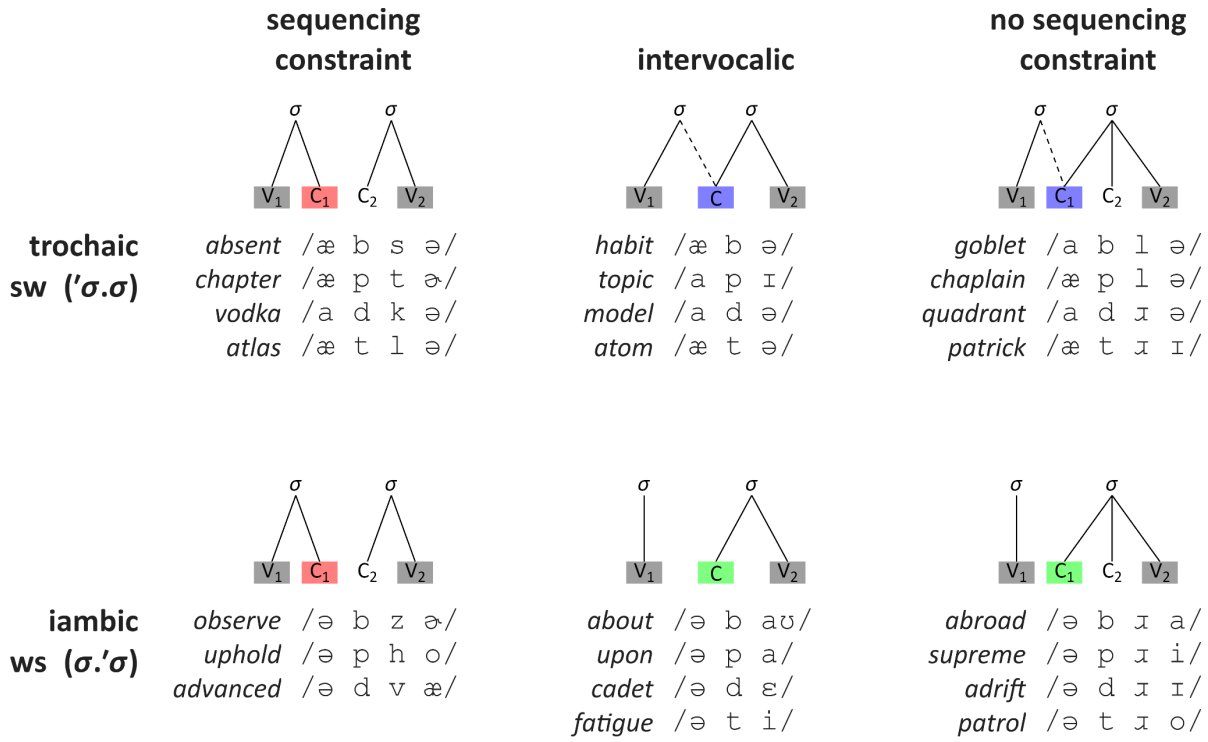

Fig. 5. Target words by environment. Environments are characterized by stress pattern (strong-weak or weak-strong) and structural context (coda-biased (VC.CV), intervocalic (VCV), or onset-biased (V.CCV).

It was not possible to satisfy all of the desiderata in target word selection, given the lexicon of English. For one, there were a number of selected forms that participants might view as morphologically complex: *uphold*, *upon*, *abroad*, *adrift*, *goblet*, and *quadrant*. Effects of morphological boundaries in these forms are likely mitigated by the circumstances that (i) the morphological complexity is derivational and rather than inflectional, (ii) it does not have a high degree of productivity, and (iii) in some of cases requires positing bound roots (*gob-* in *goblet* and *-rant* in *quadrant*). Second, the candidates available for the target consonant /t/ in clusters in the trochaic context without a sequencing constraint (e.g. *fatwa*, *hatrick*) were judged to be too low in familiarity, and hence the name *patrick* was used in this condition. Third, there were no suitable forms for /t/ in the iambic cluster environment with a sequencing constraint, and hence this condition was excluded; thus only 23 unique target words were employed in the study, rather than 24. Fourth, control over V_2 and non-target consonants in clusters was not possible.

3.3 Data processing

Tongue surfaces were extracted from ultrasound videos using the DeepEdge AI Tool for Ultrasound (v2.8.3) with model DpEdg_General_128_25MAR2023 (Chen et al., 2020). Subsequently the surface contours were resampled with a 2000x2000 grid using 2D parametric interpolation. One of the challenges in analysis of ultrasound tongue surfaces is that in general the surfaces are not single-valued functions: for each horizontal coordinate, there may be more than one vertical coordinate along the surface—this circumstance occasionally occurs near the tongue root or tongue apex, depending on lingual posture and participant-specific anatomical variation. To address this, surfaces were converted to surface envelopes,

in which the envelope represents the longest contiguous stretch of surface pixel coordinates that is a single-valued function of the horizontal coordinate, with coordinates rounded to the nearest pixel. This strategy is effective because the horizontal coordinates where the surface is multi-valued are almost always associated with the most posterior/anterior portions of the tongue. Note that polar coordinate representations of surface envelopes were also extracted, but these did not lead to qualitative differences in analyses and hence are not reported here.

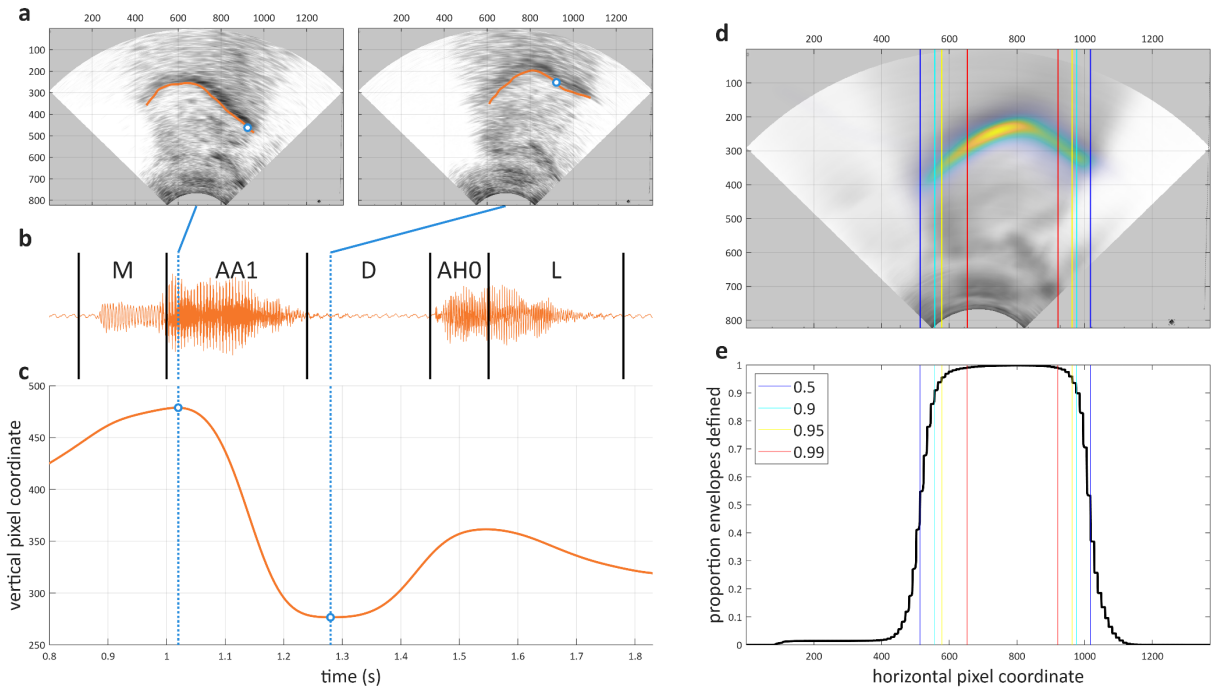

Fig. 6. Tongue blade position extraction. (a) example ultrasound image frames and tongue surfaces. (b) acoustic segmentation from forced alignment. (c) timeseries of vertical position of tongue anterior. (d) density of tongue surface envelope points from an experimental session. (e) proportion of frames with a defined surface envelope point, as a function horizontal pixel coordinate.

Scalar articulatory trajectories representing temporal variation in spatially distinct portions of the tongue were extracted from each session as follows. First, across all image frames from the session, the proportion of video frames in which there was a defined surface value was calculated (Fig. 6e). This function represents how often a tongue surface envelope point was detected by the edge tracking algorithm at each horizontal image coordinate. As long as the ultrasound probe is positioned so that both mandibular and hyoid shadows are visible, it is reasonable to assume that the endpoints of the range of coordinate values over which this proportion is relatively high (≥ 0.95) corresponds to the most anterior and posterior portions of the tongue that could be reliably detected in the images. Time series representing the height of the back, middle, and front of the tongue were obtained by averaging over three adjacent coordinates in the reliable contour range that were most posterior, most central, and most anterior. Missing values in the timeseries were linearly interpolated, then the timeseries were smoothed with zero-phase filtering using a 4^{th} order Butterworth low-pass filter (cutoff 10 Hz). Lastly the smoothed timeseries were resampled to 200 Hz with piecewise cubic Hermite polynomial interpolation (PCHIP).

Face video images were processed with Google Mediapipe (model: ceclm, 10 optimization iterations, refine_landmarks=True). This provided 128 points for each image frame, representing various facial landmarks (including points on the lips and jaw). The following Mediapipe feature indices were used to

obtain timeseries of facial features in both vertical and horizontal dimensions: UL 12; LL 17; NOSE 5; RC (right lip corner) 63; LC (left lip corner) 292; JAW 153. Note that OpenFace was also tested for processing face video images, but was observed to have noisier and less robust labial point tracking. Because video frames from the webcam and hence the corresponding feature timeseries are not uniformly sampled, the time series were interpolated with the sgolay method (window = 0.333), which corrects for non-uniform sampling. Then zero-phase filtering was applied for smoothing (4^{th} order Butterworth, 10 Hz lowpass cutoff). Lastly, the timeseries were resampled to 200 Hz with PCHIP. Vertical lip aperture was defined as a the distance between the vertical values of the LL and UL at each time point. An example of processing outputs is shown in Fig. 7.

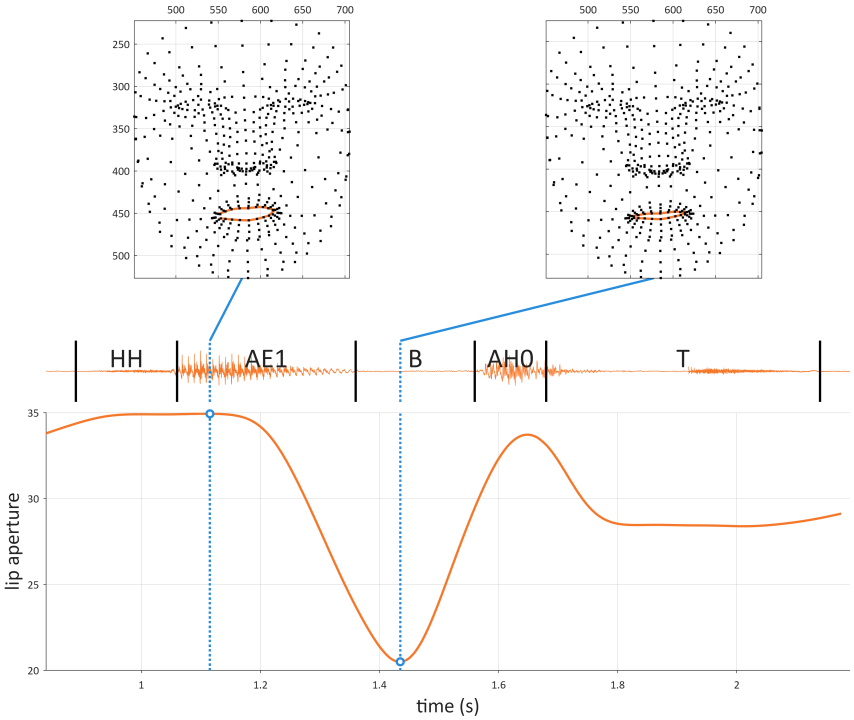

Fig. 7. Face video processing example. Top: example frames. Middle: acoustic segmentation. Bottom: lip aperture time series.

To synchronize tongue surface- and face tracking-derived time series, audio signals synchronized to the ultrasound and the webcam, respectively, were used. For each trial, the cross-correlation function between the ultrasound and webcam audio was calculated, and the lag associated with maximum cross-correlation was used to infer the time delay between acoustic and articulatory signals obtained from the two modalities. For each trial, the time coordinate of labial trajectories was shifted to compensate for this delay. Hence all articulatory and acoustic signals are represented in a time coordinate that corresponds to that of the ultrasound video.

Acoustic segmentation was obtained with the Montreal Forced Aligner (McAuliffe et al., 2017) using the english_us_arpa model. Because in paused production there is an anomalously long hesitation/period of low acoustic energy in the vicinity of the target consonant, the canonical pronunciation variants of the target words do not align well. To address this, the pronunciation dictionary was modified by adding a new dictionary entry for each target word, in which the intervocalic consonant(s) were replaced by a silent interval. The word-level transcriptions that guide the aligner were replaced with these modified entries for trials in the paused production condition; visual inspection of alignments with this modification indicated that this strategy resulted in satisfactory segmentation. With pronunciations modified in this

way, the location of acoustic C1 offset/C2 onset is not available; this is unproblematic because none of the analyses conducted here hinge on identification of this segmental boundary. Accordingly, in visualizations presented below, C1 and C2 segments of clusters are either not distinguished.

Due to various errors in acquisition and processing, 2.1% (85/4140) of trials were excluded in the above procedures (67 due to system errors in retrieving ultrasound audio; 2 failures in forced alignment; 1 failure in face tracking; 17 errors in tongue surface tracking). Thus 4055 trials remained for subsequent analysis.

Articulatory landmarks of the formation and release of constriction associated with the target consonant were extracted from all trials, using lip aperture time series for bilabial stops (/p, b/), and vertical position of tongue anterior for alveolar stops (/t, d/). Landmarks were selected by identifying candidates from a given trial and using a weighted cost function to select the optimal candidate. Landmarks from an example trial are shown in Fig. 8 below. To obtain landmarks, first the velocity extremum associated with the constriction formation (C-ons) was identified from the gradient of the relevant articulatory trajectory using equally weighted costs of (i) z-scored distance of the candidate from a reference time (the acoustic offset of V1) and (ii) z-scored magnitude of the velocity associated with the candidate. Next the preceding positional extremum (C-min) was identified using a similar cost regime applied to position, with the velocity extremum as the reference time and with candidates restricted to precede the velocity extremum. The following positional extremum (C-max) was identified in the same way, but restricting candidates to occur after the velocity extremum and giving twice as much weight to distance from reference—this weight adjustment was necessary because in some cases there are multiple positional extrema associated with the constriction plateau, especially during the intersyllable pause. Lastly, the velocity extremum associated with the constriction release (C-rel) was identified by using the positional extremum as a reference and restricting the time range of candidates to the period from the preceding position extremum to the midpoint of the acoustic segment of V2. The above procedure failed to identify all four landmarks on 7.2% (293/4055) of trials, which were excluded from subsequent analyses.

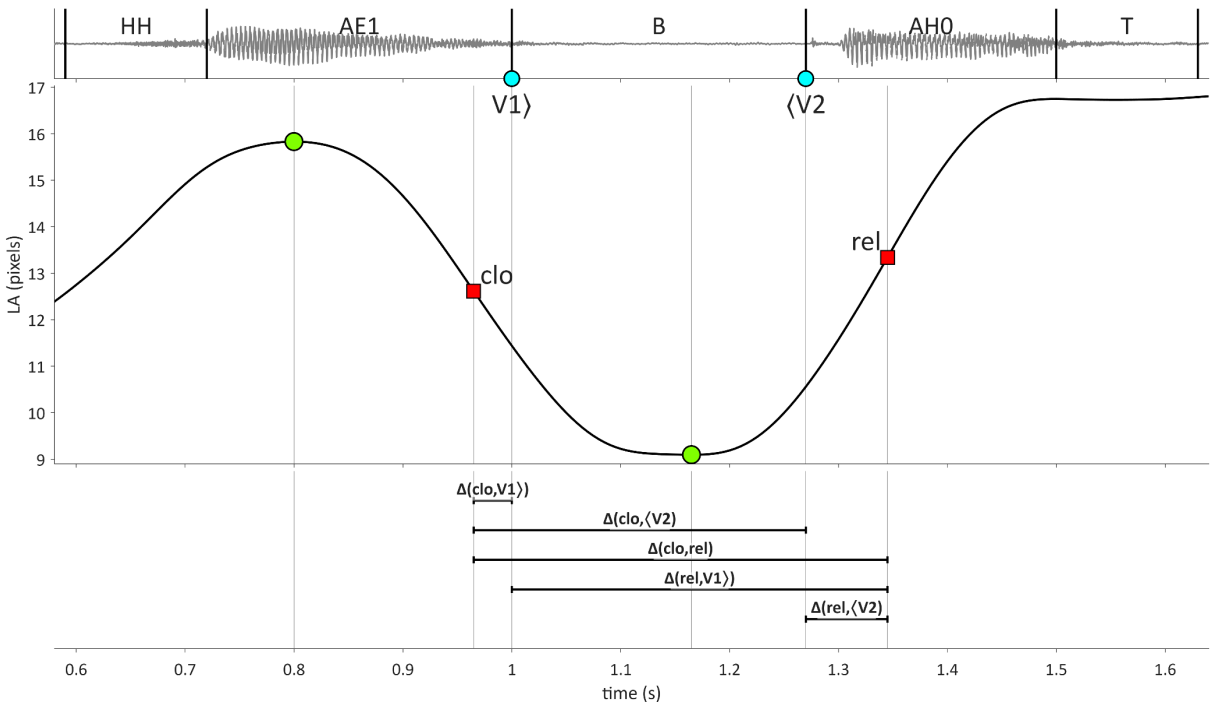

Fig. 8. Example of articulatory and acoustic landmarks for a paused production of the word *habit*. Top: acoustic segmentation. Middle: lip aperture timeseries. Bottom: timing intervals of interest for analyses. Landmarks shown: acoustic offset/onset of V1/V2 (blue circles); velocity extrema of closure and release movements (red squares); positional extrema preceding/following closure extremum (green circles).

For timing analyses, the velocity extrema associated with the formation and release of the consonantal closure (Fig. 8: clo, rel) are the landmarks of primary interest. It is important to note that these are not estimates of when constriction formation/release movements begin, but rather, estimates of when movements attain maximum speed. They are thus biased to occur a bit later in time than estimates of movement onset. We use them in this context for several reasons. First, movement velocity extrema can be identified more robustly than movement onsets, which can sometimes mistakenly identify onsets from non-active influences on vocal tract states. Second, in typical circumstances, movement onset estimates require arbitrary criteria (e.g. it is common to determine the onset as the time at which the trajectory gradient exceeds 20% of its maximum value). Third, variation in environments that precede/follow the target consonants adds additional noise and bias to estimates of constriction formation/release onset. Fourth, in relation to the experimental hypotheses, the velocity extrema that we use provide more conservative heuristics for assessing whether constriction formation is temporally coincident with the preceding vowel, since the velocity extremum necessarily occurs later in time than movement onset.

The temporal intervals examined in the analyses (Fig. 8, bottom) are defined with the articulatory landmarks (clo, rel) and the acoustic offset/onset of the preceding/following vowel, here notated as V1⟩ and ⟨V2. The acoustic segmental landmarks of vowels are used for this purpose, as opposed to articulatory signal-based landmarks, for the following reasons. First, word forms were elicited in isolation rather than in a carrier phrase, and this allows participants to anticipate vocalic tongue posture to varying extents before response initiation, rendering estimates of V1 gestural onsets more variable. Second, variation in V1 and V2 vowel quality induces stimulus-related bias in estimation of vocalic gestural landmarks. Third, coarticulatory interactions between vocalic and consonantal gestures (associated with consonants before V1, C1, and also C2 in stimuli with clusters) makes robust across-speaker detection of vocalic gestural landmarks impractical. In effect, V1⟩ (acoustic offset of preceding vowel) can be considered a proxy for

vocalic gesture target achievement, one which is biased to occur later in time than true target achievement, and likewise ⟨V2 (acoustic onset of following vowel) is a proxy for vocalic gesture onset, biased to occur later in time than true vocal gestural onset. Because the analyses rely primarily on comparisons between timing patterns in normal and paused production, these biases are not problematic.

Lastly, a standard mixed effects regression was conducted to identify trials in which participants failed to follow task instructions, produced an errorful response, or exhibited abnormal timing patterns. To this end, Δ[clo, ⟨V2] was regressed with fixed effects of production mode and environment and random slopes for production mode by subject. Trials whose z-scored residuals fell outside of central 98% of a normal distribution (i.e. |z|>2.32) were excluded from subsequent analyses (3.3% (126/3762) of trials). The timing patterns excluded in this step are sometimes associated with accidental pausing in normal production, forgetting to pause in paused production, or simply cases where a speaker paused substantially longer or shorter than usual.

3.4 Data analysis

Two different analyses of the production data were conducted. First, for all landmark-based dependent variables, acoustic segmental durations, and kinematic measures of closure/release (maximal speed and movement range) a Bayesian multilevel regression was conducted. The regression model included all interactions of production mode (2 levels: normal vs. paused), environment (6 levels: 2 metrical contexts x 3 structural contexts (cluster with sequencing constraint, cluster without sequencing constraint, intervocalic singleton), and consonant identity (p, b, t, d), along with a by-subject random intercept and by-subject random slopes for effects of production mode, environment, and their interaction. It also included a three-way interaction for fixed effects of production mode, environment, and consonant identity on the standard deviation model parameter, allowing for condition-related differences in variance to be estimated. R-hat values for all relevant model parameters were close to 1, indicating that the model sampled the posterior efficiently for all conditions. The model parameter associated with the combination of predictors missing from the target word set (i.e. /t/ in the coda-biased weak-strong environment) could not be sampled efficiently because no data were present in this condition. In sections 4.1, 4.2, and 4.4, we present distributions of posterior predicted means and standard deviations from the models, along with distributions of hypothesis-related predicted differences where relevant.

Second, in order to investigate more precisely when in time the influences of intersyllable pause production are manifested, along with the degree to which those influences vary by environment, a local relative rate analysis was conducted. Local relative rate is measure of the rate of temporal progression of one instance of a process relative to another instance of the same process. It is obtained by estimating the slope of the warping curve obtained from dynamic time warping (see Tilsen & Tiede (2022) for further detail). This slope is estimated in sliding windows, and hence the resulting local relative rate (LRR) timeseries represents an approximation of the instantaneous rate of progression a given instance of a process relative to a reference instance. The following steps were conducted in this analysis. First, for each subject/environment/production mode/consonant place, the relevant articulatory trajectories (LA for bilabial stops, tongue anterior height for alveolar stops) were collected and the portions of those trajectories spanning the [⟨V1, ⟨V2] interval were linearly time-normalized to their median duration. The normalized trajectories in each combination of factors were then fit with a GAM model with a smooth of time. Next, for each subject/environment/consonant place, a dynamical time warping (DTW) was conducted between multidimensional signals from normal and paused production modes, with the signal dimensions being the GAM-predicted mean trajectories and their gradients, both of which were re-scaled to unit variance in order to give them equal weight in the DTW algorithm. A weak local slope constraint was imposed on the DTW (symmetric_S10; see Tilsen & Tiede (2022)) in order to avoid infinite/zero slope estimates. The LLR was then estimated from the DTW warping curve. Section 4.3 shows average LLR

timeseries from each environment, after they have been time-normalized across subjects/places to the median length in that environment.

Metalinguistic syllabification intuition responses were analyzed as follows. First, participants provided four unique responses in the task. These are coded as .C1 (boundary before C1, implying onset syllabification), C1 (boundary on C1, implying ambisyllabicity), C1. (boundary after C1, implying coda syllabification), and C2 (boundary on C2). Of these, C2 (which is unexpected for any of the stimuli) occurred only twice (2/207; 0.97% of responses); hence these responses were excluded from all analyses. One participant (AG01) did not perform the syllabification intuition task due to time constraints. The responses were fit with a maximal Bayesian multinomial model with fixed effects of environment, target consonant place, target consonant voicing, and preceding vowel quality, along with reduced models excluding fixed effects. Models were compared using leave-one-out cross-validation estimates, and the simplest model that was statistically indistinguishable from the best model was selected for analyses—in this case, a model with only a fixed effect of environment. Section 4.6 presents distributions of posterior predicted response proportions from this model.

4. Results

4.1 Gestural timing patterns

Two distinct timing patterns were observed across all environments, corresponding to the predictions for coda-syllabification and ambisyllabicity. Fig. 9 shows for each environment, the hypothesized gestural score in paused production, along with posterior distributions of the mean timing of constriction formation (i.e. closure) and release landmarks (marginalized over consonant identity), for normal (blue) and paused (pink) production. Note that the time coordinate is defined relative to following vowel onset. Posterior mean segmental intervals for each environment/production mode are shown above the landmark distributions.

First, observe that in the presence of a sequencing constraint (*V.CCV environments), the timing patterns were consistent with coda-syllabification, as predicted. Specifically, both closure and release landmarks were displaced substantially earlier in time in paused production compared to normal production. Second, in both VCV environments and in the iambic V.CCV environment, the timing patterns in paused production were consistent with ambisyllabicity: closure was displaced with V1⟩, but release remained proximal to ⟨V2. In other words, closure and release were temporally dissociated. This pattern was predicted for trochaic VCV environments, but not for the iambic VCV and V.CCV environments. Below we will see that the unexpected occurrence of the dissociation pattern in the iambic context is a consequence of a methodological limitation: unstressed syllables become stressed in paused production. Lastly, in the trochaic VCCV environment, where the dissociation pattern was predicted, the results are somewhat ambiguous: release was displaced earlier in paused production, but not quite as early as in the *V.CCV environments where coda-syllabification was observed. Further inspection of model predictions below shows that the ambiguity in the trochaic VCCV environment is due to interspeaker and consonant-specific variation.

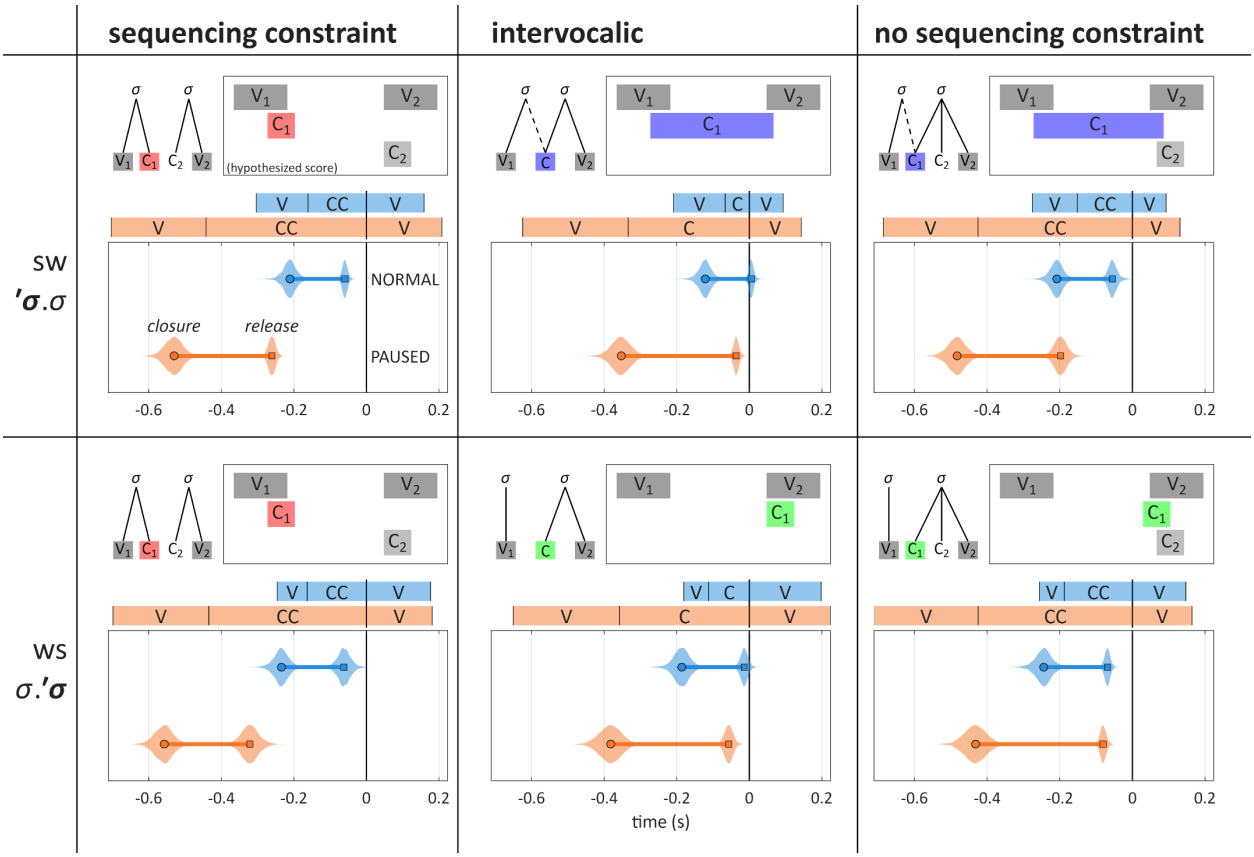

Fig. 9. Constriction formation timing by production mode and environment. Distributions are posterior predicted densities of mean constriction formation (circles) and release (squares), defined relative to V2 acoustic onset (black vertical lines). Distributions from normal (blue) and paused production (pink) are shown. Hypothesized syllabification and gestural scores are included for reference.

The occurrence of ambisyllabic dissociation in iambic VCV and V.CCV environments, where no weight constraint applies, has a straightforward explanation: in paused production, the initial unstressed syllable of an iamb is effectively produced in isolation, and this results in the vowel becoming stressed. Hence the intended effect of the iambic context—to remove the weight constraint—is unlikely to have manifested. Indeed, notice from the segmental intervals in Fig. 9 that V1 duration is not substantially different between iambic and trochaic contexts in paused production. We examine additional evidence for this metrical strengthening effect in section 4.2.

Another detail worth mentioning here relates to more subtle effects on the duration of the closure-release interval. Although releases are clearly displaced substantially earlier in time in the coda-syllabification pattern compared to the dissociation pattern, the coda-syllabification closure-release interval is nonetheless longer (by about 50-75 ms) in paused production than it is normal production. This effect is likely due to a slowing of speech rate that occurs in paused production, which we examine more closely in section 4.3.

Third, the ambiguity of the timing pattern in the trochaic V.CCV environment is evidently a consequence of interspeaker and consonant-specific variation. Inspection of by-speaker timing patterns in this environment (Appendix: Fig. A1) indicate that some speakers employed the coda-syllabification pattern, while others employed the dissociation pattern. Furthermore, inspection of by-consonant timing patterns (Appendix: Fig. A2) indicate that the coda-syllabification pattern was predominantly associated with alveolar stops in this environment, while the bilabials tended to be produced with the dissociation

pattern. Note that caution is warranted in drawing inferences about effects of consonant place and/or voicing *per se*, since stimulus design desiderata necessarily confounded these variables with lexical identity and with differences in segmental environments, which in turn can manifest as bias in articulatory landmarks, due to coarticulation.

The important conclusion from the above analyses is that there are two distinct timing patterns, and variance results provide further support of this. Each panel of Fig. 10 shows posterior distributions of the standard deviation parameters of relevant temporal intervals. From left to right these are: timing of closure relative to V1⟩, timing of closure relative to ⟨V2, timing of closure to release, timing of release to V1⟩, and timing of release to ⟨V2 (see Fig. 8 for depiction of intervals). When comparing the coda-syllabification environments (*V.CCV) to the ones in which the dissociation was observed (VCV and V.CCV), we see a clear difference between the variability of intervals defined relative to V1⟩ vs. intervals defined relative to ⟨V2. For the coda-syllabification pattern, both closure and release are less variably timed relative to V1⟩ and to each other than to ⟨V2 (examine the distributions connected with lines in the panels). This is exactly the what we would expect if closure is coordinated with the preceding vocalic gesture and release is coordinated with closure. In contrast, for the dissociation environments, whereas closure is less variably timed to V1⟩ than it is to ⟨V2, the reverse holds for release: release is less variably timed to ⟨V2 than to V1⟩.

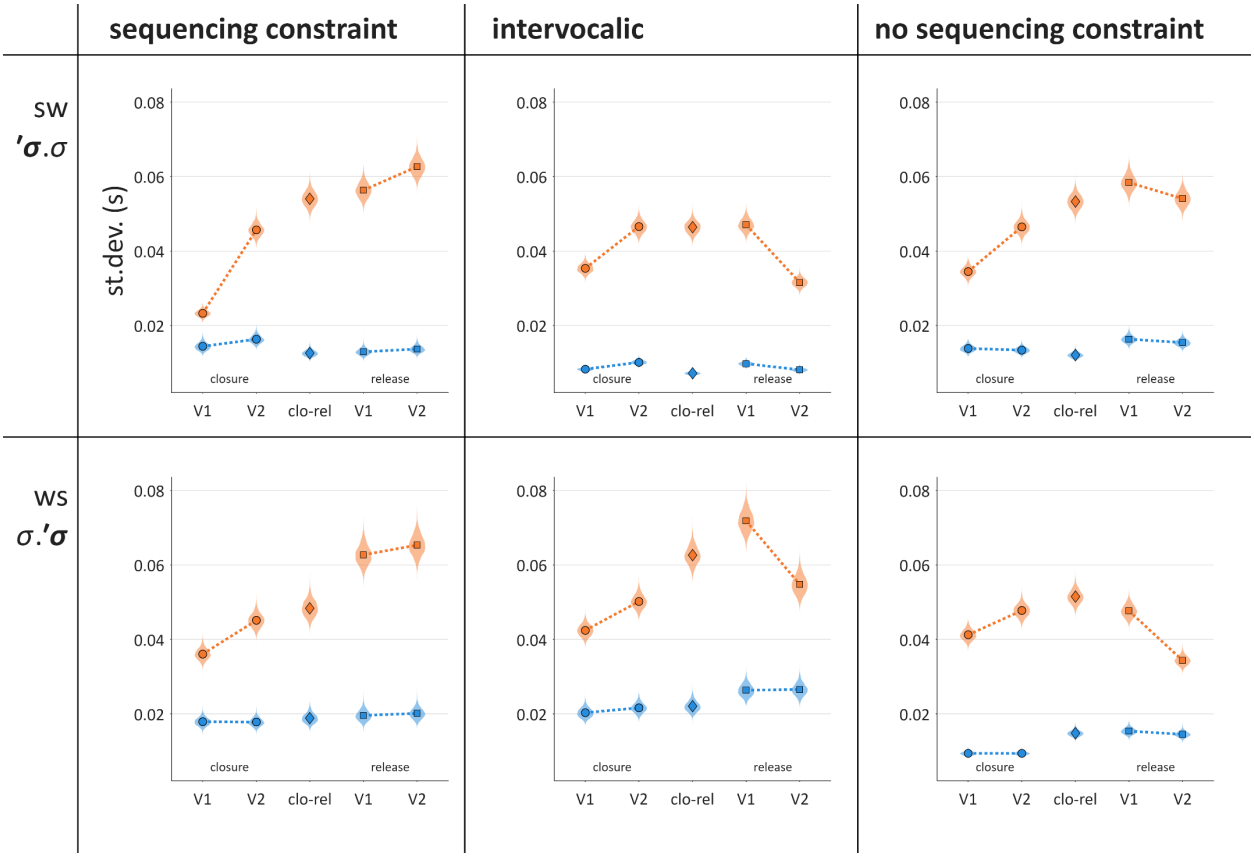

Fig. 10. Analysis of timing variability. Each panel shows posterior distributions of standard deviation parameters for relevant timing intervals. Lines indicate relevant comparisons between intervals defined relative to preceding vs. following vowels.

Notably, in the trochaic V.CCV environment, the differences in posterior standard deviations for intervals from release to V1⟩, closure, and ⟨V2 are consistent with dissociation but are relatively small in

magnitude. This is likely a consequence of the aforementioned speaker- and consonant-related variation. Hence the variance analysis reinforces the inference that two timing patterns occurred in paused production: either coda-syllabification or ambisyllabic dissociation.

4.2 Metrical context effects

The unexpected occurrence of the dissociation pattern in iambic VCV and V.CCV environments is likely due to a metrical reorganization in paused production: the initial unstressed syllable of the iamb, because it is produced in isolation in paused production, in effect becomes a stressed syllable, which in turn induces a weight constraint effect. This accounts for why dissociation occurs in these environments, but without further qualification it would predict that there are no timing differences between metrical contexts whatsoever.

To the contrary, there are some apparent differences between metrical contexts in paused production, and these depend on the structural condition. Fig. 11 shows posterior predicted differences (sw – ws) for both production modes, for several relevant variables. In panel (a) we see that, when the sequencing constraint is present, the closure (which typically precedes V1⟩) occurs closer to V1⟩ in trochaic contexts compared to iambic ones, but the reverse obtains in clusters without a sequencing constraint. A similar pattern is shown in panel (b) for the timing of release and the ⟨V2: with the sequencing constraint, release (which also typically precedes ⟨V2) occurs closer to ⟨V2 in the trochaic context than the iambic one, but the reverse holds without the sequencing constraint. It is also manifested in panel (c), which shows the closure-release interval.

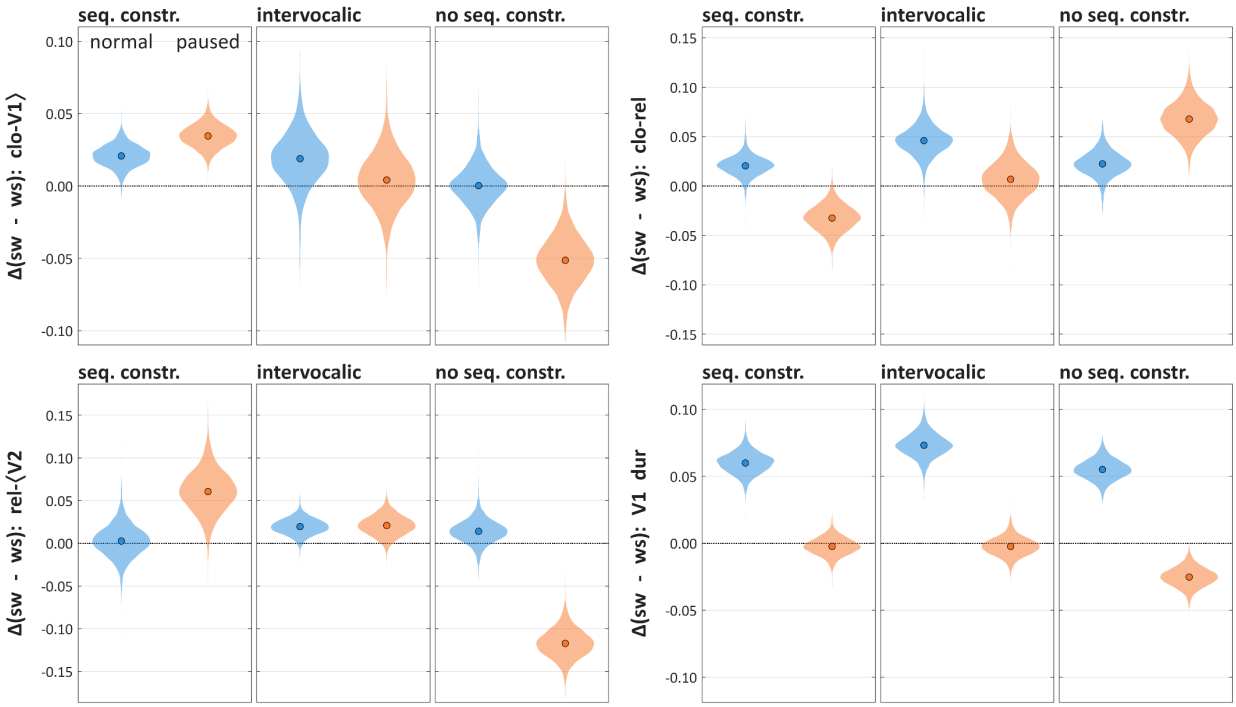

Fig. 11. Effects of stress on temporal intervals. Distributions are densities of posterior predicted differences between the trochaic and iambic contexts (sw – ws). Blue: normal production; pink: paused production.

Finally, the metrically conditioned durational contrast of the initial vowel (panel d), which is as expected in normal production for all three environments, is neutralized in the *V.CCV and VCV environments, but not in the V.CCV environment: here in paused production the vowel duration is actually

a bit *shorter* in the trochaic context than the iambic one. The crucial point to take away from these patterns is that, although the unstressed vowels of iambs patently become stressed in paused production, this does not result in temporal patterns that are indistinguishable from those in trochaic forms. This does not negate the proposed explanation for dissociation in iambic contexts—i.e., metrical strengthening of unstressed syllables in paused production—but it does indicate that some remnant of the metrical contrast in normal production persists in paused production.

4.3 Slow-down dynamics

Local relative rate analyses in Fig. 12 show that production rate is slowed throughout the ⟨V1-⟨V2 interval in paused production compared to normal production. The black lines in the figure show the average local relative rate, where values below 1.0 indicate a slow-down of paused production relative to normal production (see section 3.4 for methodological details).

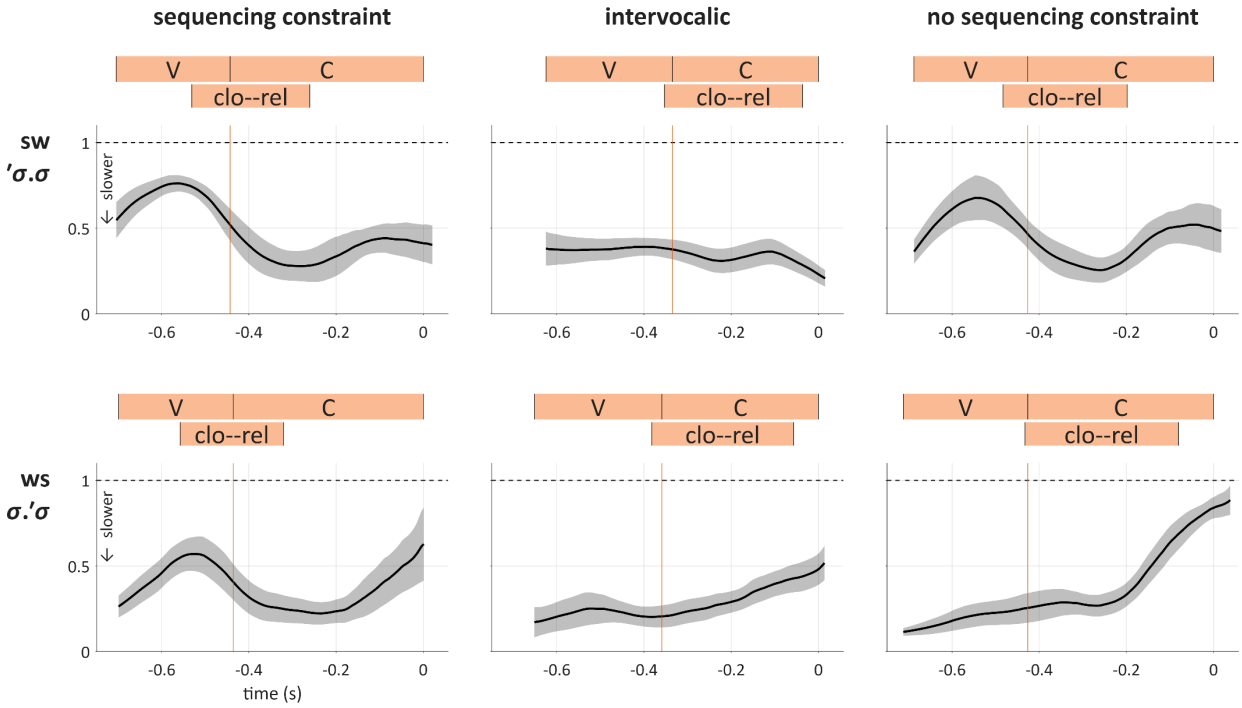

Fig. 12. Dynamic time-warping local relative rate timeseries by environment. Black lines: average local relative rate timeseries (slow down of pause production relative to normal production); gray bands: ± 1.0 standard error. Lower values indicate a greater degree of slowing.

Notably, in both trochaic and iambic *V.CCV environments, as well as the trochaic V.CCV environment, the slow-down was relatively less pronounced during the V1 acoustic interval and reaches a maximum degree of slowing in the vicinity of the consonantal release: this suggests that the production mode effect manifested most strongly in a delay of the release in these environments. In the other environments (VCV and iambic V.CCV), the slowing was relatively constant throughout most of the ⟨V1-⟨V2 interval, but exhibited a reduction in magnitude toward the end of the interval in iambic contexts. As with the temporal interval analysis, interspeaker and consonant-specific variation in syllabification in the trochaic V.CCV environment accounts for its similarity to the *V.CCV environments.

4.4 Constriction kinematics

Several aspects of the theoretical interpretation of articulatory control in paused production can be informed by analysis of movement kinematics. In particular, it is relevant to assess whether the constriction movement occurring after the initial vowel exhibits kinematic properties that are consistent with active gestural control, as opposed to a return to a neutral vocal tract state (see section 1.2). This is important because it is logically possible that the increase in constriction degree subsequent to the first vowel is initially driven by neutral control rather than an active closure gesture, in which case the effects of subsequent activation of the closure gesture would be diminished, because the neutral control target would have been approximated prior to that activation. A possible consequence of this scenario would be that estimated constriction formation landmarks reflect neutral control rather than active gestural control.

Although movement speed and range measures do exhibit production mode and environment-conditioned differences, the magnitudes of speed differences are not consistent with a neutral control interpretation and the direction of range effects is reversed from what neutral control would predict. This is evident from Fig. 13, which shows posterior distributions of mean maximal speed and movement range by production mode and environment. Note that both measures are rescaled by place, such that 1 corresponds to the maximum speed/range observed across the experiment for a given place of articulation, and 0 indicates a velocity/range of 0.0.

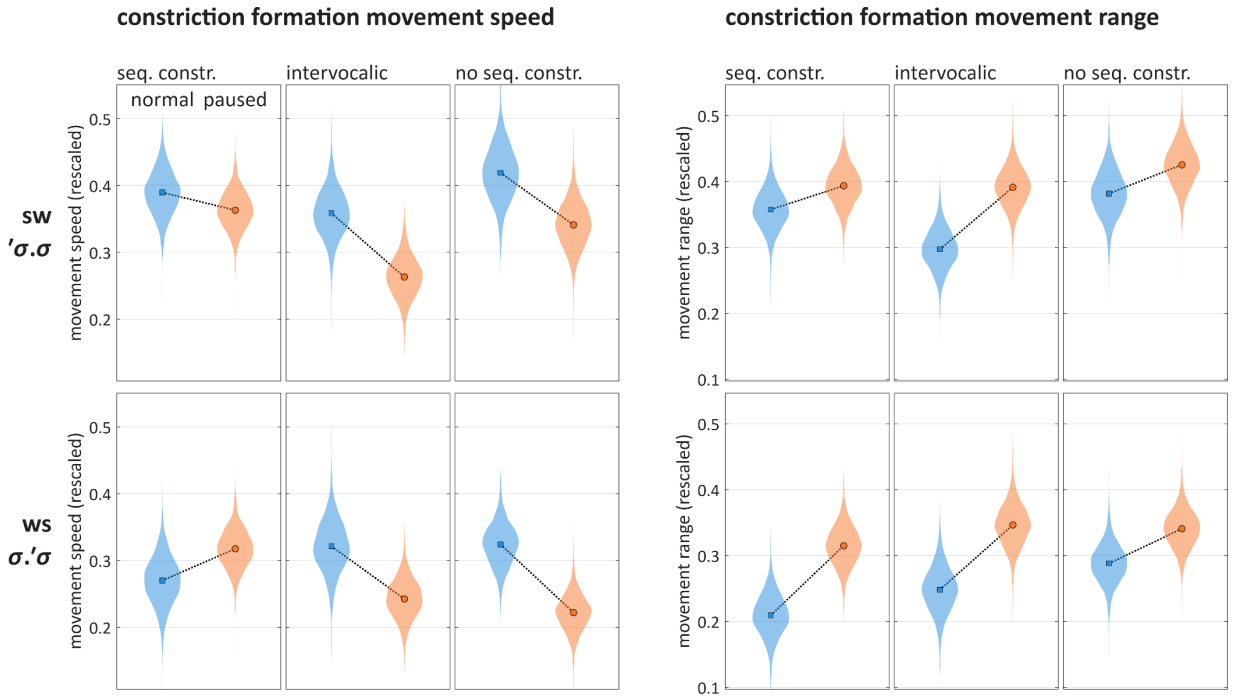

Fig. 13. Kinematic effects of production mode and environment on closure. Speed and range measures are rescaled by consonant place. Blue: normal production; pink: paused production.

The kinematic measures show that constriction formation movement speed is reduced to some extent in paused production compared to normal production, while movement range is increased. The largest speed differences between conditions are approximately 10% of the maximum (or 25% of the normal mode mean). The critical question is whether a proportional reduction of this magnitude suggests that the corresponding velocity extremum landmark is not associated with active control. On that point, consider that a similar effect magnitude obtains between trochaic vs. iambic contexts in normal production. If we interpret the 10% reduction of speed in paused vs. normal production as evidence of non-active control,

then we could also interpret the 10% reduction between iambic vs. trochaic contexts in this way. However, this conclusion is counter-intuitive, since there is no reason to believe that in normal production, non-active control influences would manifest in the iambic environments but not trochaic ones.

Furthermore, the fact that movement range *increases* in paused production is contrary to what non-active control would predict, since the tongue tip constriction degree and lip aperture targets of neutral control systems should have less extreme target postures than the targets of active closure gestures. Readers familiar with the Task Dynamic model may notice that the co-occurrence of reduced speed in combination with increased movement range appears to be paradoxical: we generally expect larger movement ranges to be associated with higher maximum speeds, a relation which derives from the assumptions that gestural stiffnesses and targets are fixed parameters. However, this relation only obtains when empirically observed movement range closely reflects the achievement of gestural target; to the contrary, if there is any target undershoot, the relation no longer holds. Target undershoot is quite likely in normal production in the current context. Furthermore, the relation can be easily obscured by coarticulatory effects, and empirical support for the assumptions that gestural targets and stiffnesses are invariant across utterances is weak. Hence, given the inconsistency of kinematic results with neutral control, it is reasonable to infer that the constriction movements after V1 in all environments result from active closure gestures.

4.5 Multiple constriction formation

In post-experiment inspection of articulatory trajectories, it was observed that, in some cases, a constriction formation movement occurred both before and after the pause. In other words, a closure movement occurred, followed by a partial relaxation of the closure, and subsequently a reformation of the closure. In order to assess how often these occurred, articulatory trajectories (lip aperture for bilabials, anterior tongue height for alveolars) were extracted from VCV environments, over the timespan defined between the closure and release velocity extrema. For each place of articulation, trajectories were time-normalized and rescaled to the range [0,1]. Then k-means clusterings of the time-normalized trajectory velocities were obtained for k=2 to 6 (Euclidean distance, 50 replicates). Note that by clustering only on trajectory velocities, cluster affiliation is based on trajectory shape rather than positional values. Cluster centroids and cluster assignment proportions within production mode from are shown in Fig. 14 for k=5, which was the minimum number of clusters necessary to capture the relevant qualitatively distinct articulatory patterns.

The cluster centroids for both bilabials and alveolars show three generic patterns. By far the most common (clusters 1-3, dotted lines) are trajectories with a single positional extremum flanked by unique positive and negative velocity extrema. The differences between these are likely a consequence of factors such as speaker, consonant, and stimulus-specific phonological environments. More interesting are the centroids of clusters 4 and 5 (solid lines), which can be interpreted as indicating multiple occurrence of closure. One of these (cluster 4/red) exhibits an early positional extremum, reflecting a fully formed closure; for bilabials, this was followed by a slight relaxation of closure and a subtle movement to reform closure; for alveolars, it was followed by a gradual relaxation of closure. The other pattern (cluster 5/blue) was somewhat less common: it exhibits a late positional extremum preceded by multiple closure formation velocity extrema.

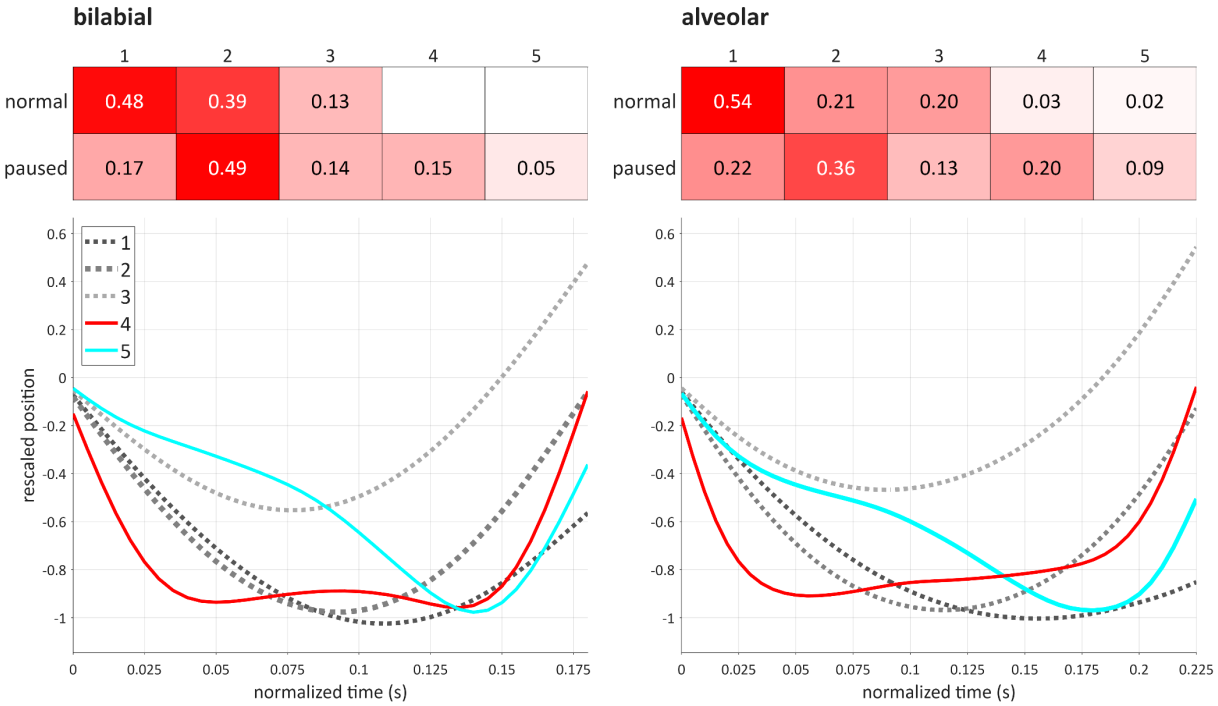

Fig. 14. Clustering-based analysis of multiple constriction formation occurrence. Tables show the proportions of cluster assignments within production mode. Trajectories are cluster centroids. Solid lines are centroid trajectories of clusters in which tokens are likely to exhibit multiple constriction formation movements.

A crucial point here is that in normal production, the multiple occurrence patterns never occurred with bilabials and were infrequent for alveolars; however, in paused production, multiple occurrence patterns occurred more frequently: about 20% and 29% of VCV environment trials for bilabials and alveolars, respectively. The infrequent cases for alveolars in normal production may result from measurement noise, and are unlikely to be of significance; but the frequency of cases in paused production is meaningful. Tokens grouped with the multiple occurrence clusters were observed for a majority of participants and for both consonant places; this indicates that the multiple constriction formation pattern is not attributable to a small subset of participants or to instrumental effects. Although multiple occurrence of closure was not predicted, Section 5.1 argues that it has a straightforward interpretation in Selection-coordination framework: closure is selected with the first vocalic gesture, its activation may then decay during the pause, and subsequently it is reselected with the second vocalic gesture.

4.6 Syllabification intuitions

Syllabification intuitions were generally consistent with gestural timing patterns, with the caveat that some participants appeared to avoid ambisyllabic responses where they would otherwise be expected. This avoidance likely arises from a dispreference for splitting a grapheme between syllables. Fig. 15 shows posterior-predicted distributions of coda-syllabification responses (C.), ambisyllabic responses (<C>), and onset-syllabification responses (.C), for each environment. These distributions represent the model-predicted proportions of each of the three syllabifications, without taking into account participant- and item-specific factors.

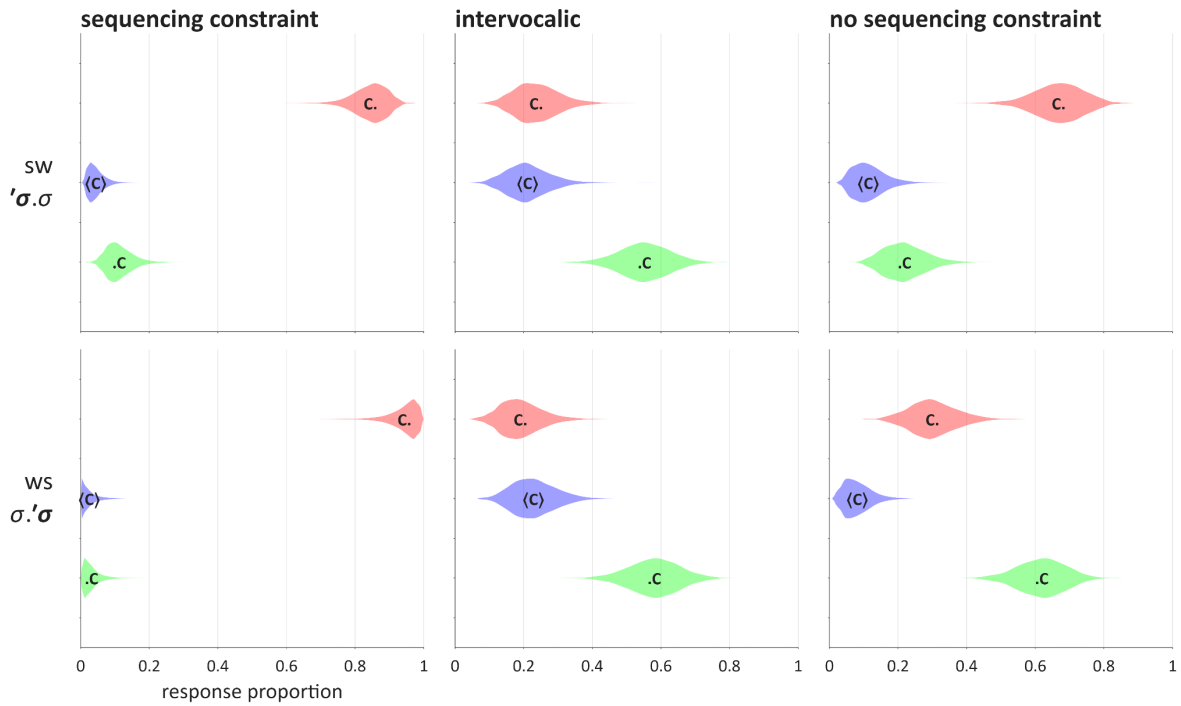

Fig. 15. Posterior predicted proportions of syllabification intuitions by environment. Panels show posterior distributions of proportions for coda-syllabification of the target consonant (C./red), ambisyllabicity (<C>/ blue), and onset-syllabification (C./green).

As one might expect, the proportion of coda-syllabification responses is near ceiling in *V.CCV environments. In contrast, in intervocalic environments (VCV), coda-syllabification responses were fairly low, while the combined proportions of onset-syllabification and ambisyllabic intuitions were relatively high. It is notable that onset syllabification responses were more prevalent than ambisyllabic ones, since these environments exhibited the formation-release gestural dissociation pattern; this discrepancy likely reflects the grapheme-splitting bias mentioned previously. Also noteworthy for VCV environments is the absence of a substantial difference in responses between trochaic and iambic contexts: phonological analyses would typically not predict ambisyllabic organization in an iambic VCV, due to the absence of a weight constraint. To the contrary, ambisyllabic responses were no less prevalent in iambic VCV than in trochaic VCV. This might result from a transfer of behavior for the preceding production task: speakers became accustomed to paused production speech, where initial unstressed syllables of iambs are strengthened (see section 4.2). If their intuitions were based on a subvocal rehearsal that similarly adopts paused production, this would explain the similarity of intuitions for trochaic and iambic VCV forms. Response proportions in the iambic V.CCV environment were quite similar to those in the VCV environments, but with moderately increased coda-syllabification and moderately decreased ambisyllabicity intuitions.

The most anomalous intuition results were observed in the trochaic V.CCV environment, where coda syllabification intuitions were about as common as the sum of onset-syllabification and ambisyllabic intuitions. This is consistent with the observed interspeaker variation in timing pattern: some speakers—those with coda-syllabification intuitions—may coordinate release with the preceding vowel, while other speakers—those with ambisyllabic or onset-syllabification intuitions—coordinate release with the following vowel. A post-hoc analysis of the correlation (in the trochaic V.CCV environment only) between the proportion of coda responses produced by participants and the estimated by-participant random effects for the duration of release-⟨V2 interval is relatively high in magnitude (r = -0.79), indicating that

participants who produced the release earlier relative to ⟨V2 were more likely to provide coda-syllabification intuitions. This finding should be interpreted with caution, because each participant only provides four syllabification responses in each environment, and because a more compelling analysis would make use of random slopes of consonant identity (which are not available). Nonetheless, the correlation reinforces the inference that variation in articulatory behavior and variation in metalinguistic intuitions arise from a common source.

5. Discussion

The main finding of the experiment is that constriction formation and release gestures can temporally dissociate in paused production: formation appears to be coordinated with the preceding vowel, while release is coordinated with the following vowel. Specifically, this occurs for words in which no sequencing constraint prohibits association of the release with the subsequent vowel, i.e. VCV and V.CCV environments. In contrast, in *V.CCV environments, a coda-syllabification timing pattern is employed: release is coordinated with the preceding formation gesture. A notable post-hoc finding is that constriction formation gestures can re-occur: speakers occasionally allow the initial closure to relax during the pause and subsequently form the closure again.

5.1 Interpretation of timing patterns

Two distinct patterns of constriction formation and release timing were observed in paused production: (i) in environments with a sequencing constraint (*V.CCV), a coda-syllabification pattern occurred, in which both formation and release gestures were temporally displaced in paused production, such that formation occurred in proximity to the acoustic offset of the preceding vowel, and release occurred shortly after formation; (ii) in environments without a sequencing constraint (VCV and V.CCV), formation and release were dissociated, such that release occurred in proximity to the acoustic onset of the following vowel, but formation occurred in proximity to the preceding vowel. Interval variance analyses (Fig. 10) reinforced this conclusion: release was less variably timed to the preceding vowel than the following vowel in the coda-syllabification pattern, but the converse held for the dissociation pattern.

One qualification regarding the above conclusion involves the timing pattern in the trochaic V.CCV environment, where the posterior fixed effect predictions (marginalized over target consonant identity) locate the release earlier in time than in other dissociation patterns, but not as early as in the coda-syllabification pattern. This equivocal patterning of the marginal fixed effect is due to variation associated with consonant identity and interspeaker variation. Fixed effect posterior predictions separated by consonant (Fig. A2) showed that in trochaic V.CCV, alveolars tended to exhibit the coda-syllabification pattern, while bilabials tended to exhibit the dissociation pattern. Note that, due to stimulus design limitations, place is confounded with target word (in each environment there are only two words for a given consonant place), and so we cannot strongly infer that the variation in trochaic V.CCV is due to place as opposed to word identity. Furthermore, in this same environment, some speakers appear to exhibit a stronger tendency toward coda-syllabification, while others exhibit the dissociation pattern. This is apparent from by-speaker posterior predictions for temporal intervals (Fig. A1). It is also supported by the observation that by-speaker variation in the displacement of the release-⟨V2 interval was strongly correlated with syllabification intuitions (section 4.6): speakers with the coda-syllabification pattern tended to produce coda-syllabification intuitions, while those with the dissociation pattern tended to production onset or ambisyllabic intuitions. Hence the experiment indicates that there is individual variation in the organization of control in this particular environment.

Contrary to what was hypothesized, formation-release dissociation was observed in iambic VCV and V.CCV environments, where phonological analyses predict onset syllabification due to the non-applicability

of a weight constraint. This result is very plausibly interpreted as the consequence of an unavoidable methodological limitation: the vowels of unstressed syllables in iambs effectively adopt the properties of stressed vowels in paused production, because they become monosyllabic utterances. In English, monosyllabic word forms that might otherwise have an unstressed vowel, when produced in isolation, exhibit a strong tendency for the vowel be stressed, or in other terms, the utterance becomes a unary metrical foot. For example, in response to the question: *what is your favorite determiner?*, speakers are far more likely to respond with [ei] than [ə] (for the indefinite article *a*). This exemplifies a more general constraint against monosyllabic utterances without a stressed vowel. Hence the temporal patterns observed in the iambic context, when examined in paused production, are similar to those in trochaic context, because in both cases, the preceding vowel is stressed. It follows that the weight constraint applied to paused production in all environments, not just the ones with trochaic word forms. This is an unavoidable confound of paused production, and consequently in this mode the experiment lacks environments in which onset-syllabification would be expected to occur.

Despite the neutralization of stress contrast associated with intersyllable pausing, some differences were indeed observed between iambic and trochaic environments in paused production (Fig. 11). Specifically, for *V.CCV, closure was earlier in relation to V1⟩ in the iambic context than in the trochaic one, while the reverse was true in V.CCV environments. The same pattern obtained for release in relation to ⟨V2. The closure-release interval was longer in the iambic context for *V.CCV and shorter for V.CCV, where ⟨V1⟩ duration was also shorter. These observations indicate that, even though paused production induced a partial metrical neutralization that resulted in qualitative similarity of timing patterns, some remnants of the iambic vs. trochaic contrast remained. Future studies with more carefully controlled phonological environments may be better suited to drawing inferences about the nature of such effects, such as whether they are due to interspeaker variation and/or related to consonant identity.

Overall, the coexistence of coda-syllabification and ambisyllabic dissociation patterns is not readily explained by conceptualizations of speech in which segments are associated with syllables in the course of utterance production. To elaborate this point, we need to speculate regarding what segmental theories actually predict about formation-release timing. Let us consider two possibilities: one, an endogenous-release version, holds that release movements are intrinsically associated with consonantal segments; the other, an exogenous-release version, holds that release movements are attributable to a following vowel or consonant. Both versions fail to account for the empirical patterns, but for different reasons. The endogenous-release version cannot predict the dissociation pattern, unless segments are allowed to break into two parts, one affiliated with constriction formation, the other with constriction release. Obviously, the notion that a segment can split in this way is both ad hoc and antithetical to the notion of a segment as a discrete unit or container of features. The exogenous-release version fails for a different reason: in the coda-syllabification pattern, the release remains temporally associated with the constriction formation/preceding vowel—this should not be the case if release is associated with a following segment. For these reasons, the segmental conception of speech is unable to provide a coherent basis for interpreting articulatory control in paused production, at least not without additional stipulations.

Of course, one can always argue that segmental organization is a form of cognitive representation that exists prior to the implementation of speech, with an articulatory implementation module translating that representation to instructions for gestural timing. In that case, it becomes necessary to explain why the module implements formation and release together in some environments, but dissociates them in others—this rhetorical maneuver of modularizing segmental representations simply displaces the burden of accounting for empirical patterns to a indeterminate implementation module, and ultimately begs the theoretical question of why formation and release would dissociate in the first place. While segments and syllables may be useful for describing phonological patterns on large spatial and temporal scales, their utility for analysis of utterance-scale patterns is questionable.

To adequately understand the temporal patterns observed in this experiment, I argue that we need a conceptual vocabulary in which actions themselves—i.e. gestures—are the basic units of phonological organization, as well as a model of how gestures are organized and the role that feedback plays in this organization. The gestural framework of Selection-coordination theory (Tilsen, 2014, 2016) provides a suitable vocabulary. In Selection-coordination theory, the activation states of gestural systems are constrained by their organization into groups, called *co-selection sets*. The gestural systems within a set are selected together and the relative timing of when they become active is controlled by a coordination mechanism, which makes use of in-phase or anti-phase coupling. The sets themselves are competitively selected, via a feedback-governed queuing mechanism. Importantly, the feedback involved in this queuing mechanism may be external sensory feedback or internal (predictive) feedback. The theory accounts for a variety of developmental and typological patterns with the hypothesis that in the course of language acquisition, children tend to acquire coordinative control of gestural systems that were previously competitively controlled, i.e. they learn to co-select gestural systems in larger sets. For example, in early development, a post-vocalic consonantal gesture is held to be competitively selected, i.e. {CV}{C}, but through internalization of feedback control, the gesture may become co-selected with the preceding vowel, i.e. {CV~C}. A similar progression is held to occur for word-initial clusters, i.e. {C}{CV} → {C~CV}. This analysis accounts for a variety of developmental patterns, and by allowing for developmental reorganization of control to vary across languages, the theory accounts for a number of cross-linguistic typological differences in phonetic and phonological patterning (see Tilsen (2016) for details).

For current purposes, a crucial claim of Selection-coordination theory is that the organization of gestural systems into co-selection sets is flexible for adults, such that factors like speech rate, style, and attention to self-feedback can be associated with differences in organization. An important consequence of this is that canonical structural units—segments, moras, and syllables—do not necessarily correspond to how gestures are organized for selection in any particular utterance; rather they correspond to different ways in which the gestures can be organized. Fig. 16 (a) and (b) depicts hypothesized selectional organization associated with the paused production coda-syllabification pattern (*V.CCV environments) and dissociation pattern (VCV environment). For coda syllabification, the release gesture is selected with the first co-selection set and is anti-phase coupled to the closure. As a consequence of this, the closure-release interval is expected to be relatively proximal to the preceding vocalic gesture and to have relatively low variance—both of which were observed in the empirical data. Feedback control in the competitive queuing mechanism governs when the second set is selected. Here the relevant sensory feedback state corresponds to no gestural target and minimal acoustic energy, i.e. a pause. As discussed in (Tilsen, 2022), control of the temporal delay between selection events can be accomplished either by elevating a feedback threshold parameter for this state or decreasing a feedback gain parameter. In either case, the result is a brief pause before the selection of the second set, which includes the constriction formation and release gestures associated with the post-target consonant and the following vocalic gesture.

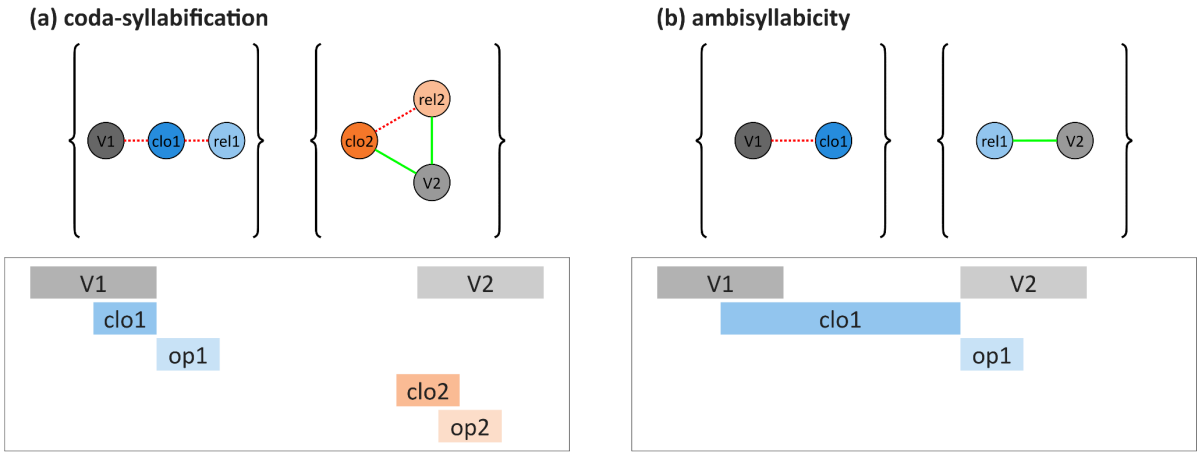

Fig. 16. Selection-coordination theory interpretations of paused production patterns. (a) coda-syllabification in *V.CCV environment: release is selected with and anti-phase coupled to closure. (b) ambisyllabic dissociation in VCV environment: release is selected with the following vocalic gesture.

For the ambisyllabic dissociation pattern (b), the primary difference in organization compared to coda-syllabification is simply that release is organized in the second set of gestures, rather than the first. An additional difference is that the feedback state used to regulate the pause includes the constriction target of the formation gesture, which is selected in the first set. Note that (b) exemplifies organization for a VCV environment; in the case of the V.CCV environment (not shown), an additional pair of formation and release gestures is selected in the second set.

This interpretation also offers a plausible way of understanding why multiple constriction movements occasionally occurred (section 4.5). First, it is important to emphasize that gestural activation is a continuous state variable. In early formulations of the Task Dynamic model (Saltzman & Munhall, 1989), gestural activation states were heuristically assumed to exhibit step function transitions between values of 0 and 1, and rectangular activation intervals in gestural scores implicitly reflected this assumption. However, step activation is not a necessary condition of the model, and subsequent work has suggested that gestural activation dynamics are more complicated (Byrd & Saltzman, 1998; Kirkham, 2025; Sorensen & Gafos, 2016; Tilsen, 2019, 2020). To account for a vocal tract state trajectories in which multiple constrictions occur, one plausible approach involves (i) allowing the activation of the constriction formation gesture to decay subsequent to its selection in the first set, and (ii) allowing the constriction formation gesture to be reselected in the second set. In this reselection account, the amount of decay of activation of the constriction formation gesture determines whether multiple constriction movements are observed in measurements of vocal tract states, along with the magnitude of the relaxation of the constriction that intervenes.

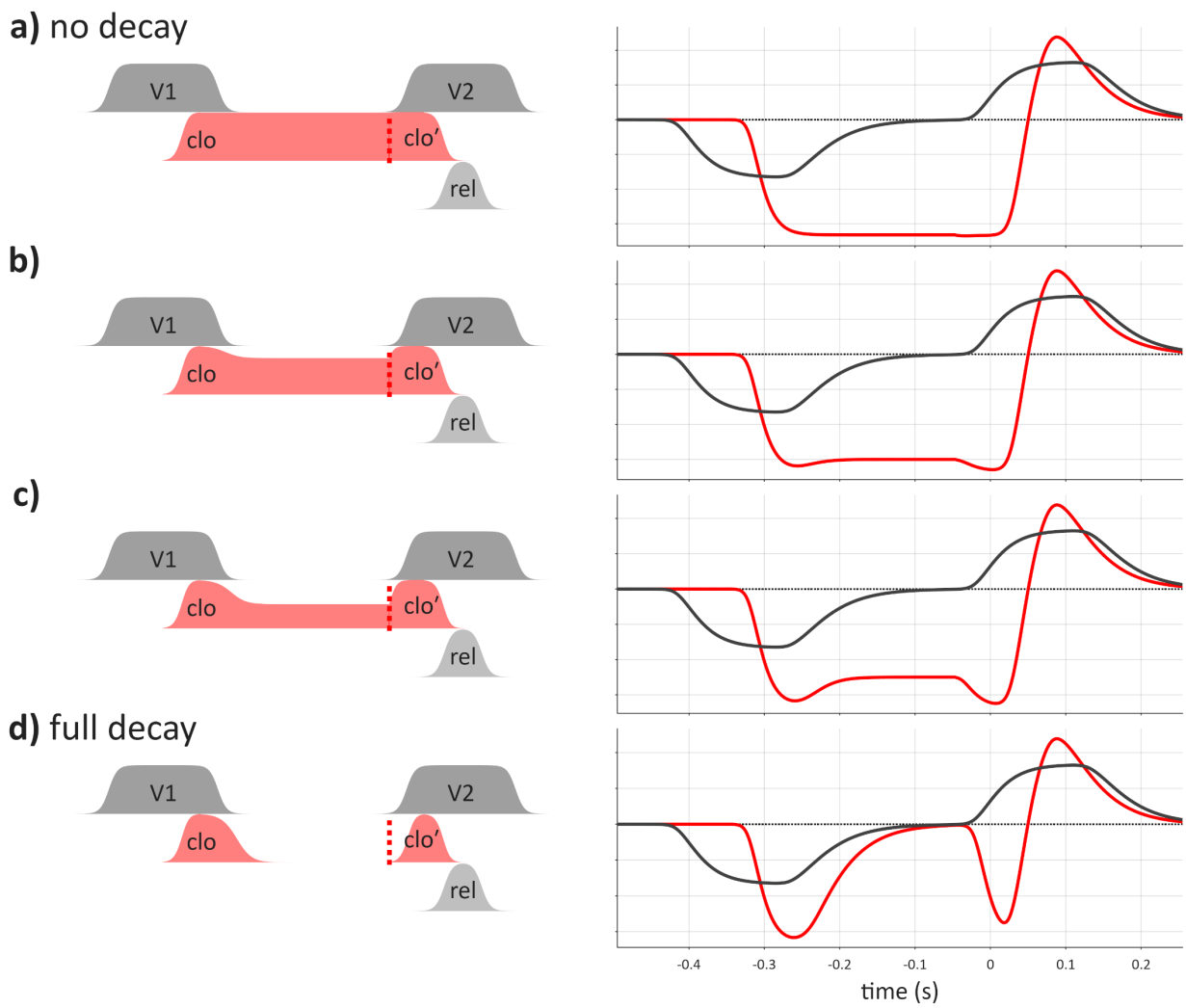

Fig. 17. Schematic illustration of multiple constriction patterns in a reselection model with decay. Left: gestural scores with continuous activation values and reselection of closure gesture (dashed red line). Right: tract variables associated with vocalic gestures (black) and consonantal gesture (red).

Fig. 17 illustrates simulations of tract variable timeseries based on the Task Dynamic model (Saltzman & Munhall, 1989), allowing for activation of the initial closure gesture to decay to a parameterized degree. Panels on the left show gestural scores with continuous activation states; corresponding tract variables are shown on the right. Panel (a) shows a model with no activation decay; this corresponds to the canonical dissociation pattern observed in VCV environments in the experiment. The vocal tract state trajectories under this parameterization are effectively indistinguishable from what Fig. 17b would generate, because the closure gesture is already in a state of maximal activation when it is reselected. In contrast, (d) shows a complete decay of activation, resulting in two extremely distinct constriction movements—this sort of pattern was not observed in the experiment. Panels (b) and (c) show simulations with intermediate degrees of decay, and generate vocal tract state trajectories that more closely resemble the multiple constriction formation patterns that occurred sporadically in paused production.

5.2 Generalizability and limitations

There are several questions to ask regarding whether inferences about phonological organization derived from paused production generalize to other contexts, specifically: (i) does the organization of formation and release gestures in the dissociation pattern generalize to more typical modes of production?, (ii) do the results for alveolar and bilabial stops extend to other consonants?, and (iii) do the specific words/stimuli used in the experiment restrict the generalizability of results?

Generalization from paused production to more typical modes of production entails that in ambisyllabic environments, the constriction formation gesture is selected and coordinated with the preceding vocalic gesture, while the release is selected and coordinated with the following vocalic gesture. The unnaturalness objection holds is that paused production is "unnatural" speech, and that we should not generalize from "unnatural" modes of speech to "natural" ones. On this point, consider that in spontaneous conversational speech, speakers sometimes do employ inter-syllable pauses to convey attitude or stance. For example, in the utterance *let me be per-fect-ly clear,* a speaker may pause between syllables of *perfectly* to convey emphasis. Likewise, in the utterance *that might be a mis-take*, the speaker employs an intersyllable pause in *mistake* to convey uncertainty. It is dubious to assert that paused production is "unnatural" when speakers appear to occasionally employ it in spontaneous speech.

More importantly, the dichotomy between so-called "natural" and "unnatural" speech is a false one. To my knowledge, no one has attempted to define the properties of these purported categories in a way that would allow any instance of speech to be categorized as {natural} or {unnatural}, and this is probably because the dichotomy itself is highly arbitrary. The category of {speech} is better conceptualized as a category with family resemblance structure, where any instance of the category is associated with a multitude of properties that apply to the speech itself or the context in which speech occurs. These properties include whether the speech is spontaneously generated or not, whether it occurs in the context of a conversation, what style it associated with (e.g. formal/casual), rate of speech, the attention of the speaker to their own external sensory feedback, environmental acoustic conditions, etc. The family resemblance structure of {speech} entails that any two instances of the category will share some properties, but there is no set of necessary and sufficient properties that determine membership in the category. Accordingly, there can be no set of properties that reliably distinguishes between subcategories of natural and unnatural speech.

From this perspective, it is dubious to maintain that empirical patterns observed in paused production cannot inform our understanding of phonological organization in more typical modes of production. For this argument to be plausible, one would have to assert that an entirely distinct cognitive system is responsible for phonological organization in paused production than the system involved in more typical modes of production. This assertion is largely untenable, given our modern understanding of how sensorimotor control is implemented in the nervous system (e.g., there is no variety of speech that does not share a large degree of overlap with other varieties, with regard to activation of neural populations). A more reasonable view is that the same control system is involved in both paused and normal production, but certain control parameters that determine the dynamics of the system differ between modes. As proposed above, the occurrence of the pause is determined by adjustment of a feedback threshold or gain parameter that regulates when the first set of gestures is suppressed and the second set of gestures is selected. This view—a single control system, differing parameterization—offers a more parsimonious understanding of speech across contexts than a view in which two distinct systems govern control in paused vs. normal modes. To reinforce this point, consider that interpretations of disordered speech generally do not maintain that speakers with neurological disorders make use of a different control system than healthy speakers; rather, they hold that it is the same system involved in normal and disordered speech, yet the system operates differently depending on the nature of the disorder.

The challenge that arises in testing inferences about gestural organization in normal production is primarily that the predicted empirical consequences of dissociated organization of formation and release are more subtle in normal production than paused production, and therefore to test these predictions

requires more statistical power than is offered by the current design. Specifically, in a VCV environment, dissociated organization makes the prediction that the variance of the temporal interval between the constriction formation and preceding vocalic gesture will be lower than that of the interval between formation and following vocalic gesture, while conversely, release to following vocalic gesture will be less variable than release to preceding vocalic gesture. This pattern was indeed indirectly observed in normal production of trochaic VCV in the current experiment (section 4.1, Fig. 10). However, there are reasons to be cautious in drawing strong inferences from this. First, the formation and release landmarks used to define intervals analyzed in this experiment were based on velocity extrema rather than estimates of movement onsets, which more closely correspond (in a theoretical sense) to when gestural systems become active. Second the vocalic landmarks used to define the intervals were based on acoustic segmentation, whereas theoretical models make predictions about the initiation of vocalic gestures. The methodological heuristics employed in the current study were a necessary consequence of stimuli selection and design limitations, which prevented precise control over segmental environments. Yet one can readily envision an alternative design using more tightly controlled non-word stimuli in a fixed carrier phase. This would allow for estimation of the theoretically relevant gestural landmarks and provide greater statistical power for estimation of interval variances. In my opinion, the results of the current study can be taken to provide evidence in favor of generalizing the dissociation model of ambisyllabicity to more typical speech, but the evidence is not conclusive, and further investigation is warranted.

Another issue of generalization relates to whether the dissociation patterns for alveolar and bilabial stops extend to other consonants not considered in this study. To my knowledge, prior articulatory studies have not examined constriction formation and release timing in the way that they were analyzed here, so the prior literature does not directly bear on this question. It is notable, however, that the study of complex segments (/l/ and /w/) in Gick (2009) observed somewhat analogous patterns. Specifically, the posterior/dorsal and anterior/coronal gestures of /l/ appeared to be more closely associated with the preceding and following vocalic gestures, respectively. Likewise for the dorsal and labial gestures of /w/. Exactly how similar these pairs of gestures are to stop closure and release gestures examined in the current study is an open question, but the pattern itself is nonetheless partly analogous. In the absence of more detailed analyses of articulatory timing, our prior allocation of credibility to the hypothesis that formation and release are dissociated in consonants not examined in the study may be slightly more than our allocation of credibility to the null hypothesis that they are not dissociated, merely on the basis of parsimony: the system dynamics are simpler if all consonantal constrictions obtain the same pattern in a given environment.

Yet another generalization question relates to the specific word forms employed for the experimental stimuli. Desiderata of familiarity, segmental homogeneity across environments, and monomorphemicity were applied to selection of stimuli, but these desiderata were not uniformly achievable across all phonologically defined conditions (see Section 3.2). Thus to some extent, the results could reflect imbalances that are attributable to the limited availability of relatively familiar, morphologically simplex words in the English lexicon. This is yet another reason why an experiment with nonword stimuli would be informative, or perhaps experiments designed to directly test for effects of familiarity or differences in segmental environments. Nonetheless, in the absence of having data from such experiments, it is reasonable to suppose that familiarity and morphological effects, if present, are secondary ones. It could be the case that only more familiar forms exhibit ambisyllabic dissociation, but this would not negate the theoretical interpretation itself; it would only indicate that the hypothesized patterns of organization can vary according to lexical familiarity.

One final point to mention relates to the correlation observed between metalinguistic syllabification intuitions and timing patterns, particularly in the trochaic V.CCV environment, where there is interspeaker and consonant place-related variation in timing patterns. Speakers with coda-syllabification judgements exhibited a strong tendency to produce timing patterns consistent with coda-syllabification, and

conversely, those with ambisyllabic intuitions exhibited patterns consistent with dissociation. The correlation indicates that phonological organization—which Selection-coordination theory interprets as a matter of grouping gestures into competitively selected sets—is a form of organization that not only determines articulatory behavior but also that speakers have access to in metalinguistic reasoning. A similar connection between articulatory behavior and metalinguistic intuitions was observed in an acoustic production study yoked with a syllable count judgment (Tilsen & Cohn, 2016), where rime durations of forms with tense vowels or diphthongs and liquid rimes (e.g. *feel*, *fire*) were correlated with syllable count intuitions. Such correlations reinforce the value of eliciting metaphonological judgments from naïve speakers.

6. Conclusion

Using an intersyllable pause production task, compelling evidence was found for the coexistence of two distinct consonantal articulatory timing patterns in environments with intervocalic singletons and consonant clusters. In environments where a phonotactic sequencing constraint is present, a coda-syllabification pattern was observed; otherwise, the formation and release of the post-vocalic consonantal constriction were typically dissociated (with some interspeaker/consonant place-related variation the in trochaic V.CCV environment). The two patterns can be readily interpreted in Selection-coordination theory as the consequence of a difference in gestural organization. Specifically, for coda-syllabification, constriction formation and release are co-selected and coordinated; for the dissociation pattern, formation is selected along with the preceding vocalic gesture, while release is selected with the following vocalic gesture. Furthermore, sporadic occurrence of articulatory trajectories in which a constriction movement occurs before the pause, partially relaxes, and then re-occurs after the pause, can be understood as the consequence of selecting the formation gesture in the first set, allowing its activation to subsequently decay, and then reselecting it in the second set.

The results ultimately call into question the utility of segment syllabification-based theories of phonological organization, at least for the purpose of understanding utterance-timescale articulatory control. Without ad-hoc revisions, such theories cannot account for the coexistence of both timing patterns. Either formation and release actions must both be associated with a consonantal segment, in which case one would expect them to always cohere in paused production; or, release must be attributable to following segments, in which case one would expect them to always dissociate. The results of the study lend credibility to the hypothesis that selectional dissociation of formation and release generalizes to typical modes of production, but future work should investigate temporal patterning with more carefully controlled non-word stimuli and with empirical measures from articulatory events that more directly index theoretically relevant events.

References

Anderson, J., & Jones, C. (1974). Three theses concerning phonological representations. *Journal of Linguistics*, *10*(1), 1–26.

Ballier, N., & Pouillon, V. (2013). The treatment of syllables and syllable boundaries in Thomas Sheridan's English pronouncing dictionary of 1780. *The Syllable, State of the Art and Perspectives*.

Blevins, J. (1995). The syllable in phonological theory. In *J. Goldsmith (ed.) The Handbook of Phonological Theory* (pp. 206–244).

Browman, C. (1994). Lip aperture and consonant releases. In P. Keating (Ed.), *Phonological Structure and Phonetic Form* (pp. 331–353). Cambridge University Press.

Browman, C., & Goldstein, L. (1990). Tiers in articulatory phonology, with some implications for casual speech. *Between the Grammar and Physics of Speech*, 341–376.

Browman, C., & Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica*, *49*(3–4), 155–180.

Burroni, F. (2022). A split-gesture, competitive, coupled oscillator model of syllable structure predicts the emergence of edge gemination and degemination. *Proceedings of the Society for Computation in Linguistics 2022*, 11–22.

Byrd, D., & Saltzman, E. (1998). Intragestural dynamics of multiple prosodic boundaries. *Journal of Phonetics*, *26*, 173–199.

Chen, W.-R., Stern, M. C., Whalen, D. H., Derrick, D., Carignan, C., Best, C. T., & Tiede, M. (2024). Assessing ultrasound probe stabilization for quantifying speech production contrasts using the Adjustable Laboratory Probe Holder for UltraSound (ALPHUS). *Journal of Phonetics*, *105*, 101339.

Chen, W.-R., Tiede, M., & Whalen, D. H. (2020). DeepEdge: Automatic ultrasound tongue contouring combining a deep neural network and an edge detection algorithm. *Proc. ISSP*, *2*.

Coleman, J. (2002). Phonetic Representations in The Mental Lexicon. *Phonetics, Phonology, and Cognition*, *3*, 96.

Durvasula, K., & Huang, H.-H. (2017). Word-internal "ambisyllabic" consonants are not multiply-linked in American English. *Language Sciences*, *62*, 17–36.

Elzinga, D., & Eddington, D. (2014). An experimental approach to ambisyllabicity in English. *Topics in Linguistics*, *14*(1), 34–47.

Fallows, D. (1981). Experimental evidence for English syllabification and syllable structure. *Journal of Linguistics*, *17*(2), 309–317.

Gick, B. (2009). Articulatory correlates of ambisyllabicity in English glides and liquids. *Papers in Laboratory Phonology VI: Constraints on Phonetic Interpretation*, 222–236.

Hammond, M. (1997). Vowel quantity and syllabification in English. *Language*, 1–17.

Hooper, J. B. (1972). The syllable in phonological theory. *Language*, 525–540.

Hooper, J. B. (1976). An introduction to natural generative phonology. *(No Title)*.

Ishikawa, K. (2002). Syllabification of Intervocalic Consonants by English and Japanese Speakers. *Language and Speech*, *45*(4), 355–385. https://doi.org/10.1177/00238309020450040301

Jensen, J. T. (2000). Against ambisyllabicity. *Phonology*, *17*(2), 187–235.

Kahn, D. (1976). *Syllable-based generalizations in English phonology* (Vol. 156). Indiana University Linguistics Club Bloomington.

Kelso, J. A., Saltzman, E. L., & Tuller, B. (1986). The dynamical perspective on speech production: Data and theory. *Journal of Phonetics*, *14*(1), 29–59.

Kirkham, S. (2025). Discovering Dynamical Laws for Speech Gestures. *Cognitive Science*, *49*(5), e70064. https://doi.org/10.1111/cogs.70064

Krakow, R. A. (1989). *The articulatory organization of syllables: A kinematic analysis of labial and velar gestures*. Doctoral Dissertation, Yale University, New Haven, CT.

Lakoff, G. (1993). *The contemporary theory of metaphor*.

Lakoff, G., & Johnson, M. (2008). *Metaphors we live by*. University of Chicago press.
Lee, J.-K. (2025). An ultrasound study of ambisyllabicity: The case of American English retroflex/ɹ. *Linguistic Research*, *42*(3).
Lee, J.-K., & Seo, Y. (2019). A phonetic examination of phonological ambisyllabicity: Focusing on temporal and spectral characteristics. *Linguistic Research*, *36*(1).
Liberman, M., & Prince, A. (1977). On stress and linguistic rhythm. *Linguistic Inquiry*, *8*(2), 249–336.
McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal forced aligner: Trainable text-speech alignment using kaldi. *Interspeech*, *2017*, 498–502.
Nam, H. (2007). Syllable-level intergestural timing model: Split-gesture dynamics focusing on positional asymmetry and moraic structure. In *In J. Cole & J. I. Hualde (Eds.), Laboratory phonology* (Vol. 9, pp. 483–506). Walter de Gruyter.
Nesbitt, M. (2018). *Acoustic Correlates to Ambisyllabic Representations in American English*.
Pierrehumbert, J., Gussenhoven, C., & Warner, N. (2002). Word-specific phonetics. *Laboratory Phonology*, *7*.
Pike, K. L., & Pike, E. V. (1947). Immediate Constituents of Mazateco Syllables. *International Journal of American Linguistics*, *13*(2), 78–91. https://doi.org/10.1086/463932
Saltzman, E., & Munhall, K. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, *1*(4), 333–382.
Schettino, V., Cutugno, F., & Origlia, A. (2015). From theory to applications, the syllable connection. In *The Notion of Syllable across History, Theories and Analysis* (pp. 388–416). Cambridge Scholars Publishing.
Scobbie, J. M., & Wrench, A. A. (2003). An articulatory investigation of word final/l/and/l/-sandhi in three dialects of English. *Proceedings of the 15th International Congress of Phonetic Sciences*.
Selkirk, E. (1982). The syllable. *The Structure of Phonological Representations*, *2*, 337–383.
Sorensen, T., & Gafos, A. (2016). The gesture as an autonomous nonlinear dynamical system. *Ecological Psychology*, *28*(4), 188–215.
Stampe, D. (1973). *A dissertation on natural phonology*. The University of Chicago.
Strycharczuk, P., & Kohlberger, M. (2016). Resyllabification reconsidered: On the durational properties of word-final/s/in Spanish. *Laboratory Phonology*, *7*(1).
Swadesh, M. (1936). Phonemic contrasts. *American Speech*, *11*(4), 298–301.
Tilsen, S. (2013). A Dynamical Model of Hierarchical Selection and Coordination in Speech Planning. *PloS One*, *8*(4), e62800.
Tilsen, S. (2014). Selection-coordination theory. *Cornell Working Papers in Phonetics and Phonology, 2014*, 24–72.
Tilsen, S. (2016). Selection and coordination: The articulatory basis for the emergence of phonological structure. *Journal of Phonetics*, *55*, 53–77.
Tilsen, S. (2017). Exertive modulation of speech and articulatory phasing. *Journal of Phonetics*, *64*, 34–50.
Tilsen, S. (2019). Motoric mechanisms for the emergence of non-local phonological patterns. *Frontiers in Psychology*, *10*, 2143.
Tilsen, S. (2020). *A different view of gestural activation: Learning gestural parameters and activation with an RNN* (Cornell Working Papers in Phonetics and Phonology 2020).
Tilsen, S. (2022). An informal logic of feedback-based temporal control. *Frontiers in Human Neuroscience*. https://doi.org/10.3389/fnhum.2022.851991
Tilsen, S., & Cohn, A. (2016). Shared representations underlie metaphonological judgments and speech motor control. *Laboratory Phonology*.
Tilsen, S., & Tiede, M. (2022). *Looking within events: Examining internal temporal structure with local relative rate*.

Treiman, R., & Danis, C. (1988). Syllabification of intervocalic consonants. *Journal of Memory and Language*, *27*(1), 87–104. https://doi.org/10.1016/0749-596X(88)90050-2

Turk, A. (1994). Articulatory phonetic cues to syllable affiliation: Gestural characteristics of bilabial stops. *Phonological Structure and Phonetic Form: Papers in Laboratory Phonology*, *3*, 107–135.

Zec, D. (1995). Sonority constraints on syllable structure. *Phonology*, *12*(1), 85–129.

Appendix

Fig. A1 shows by-participant posterior-prediction distributions of the difference in the release-⟨V2 interval scaled by mean of the posterior predicted difference in the closure-V1⟩ interval. In effect, the values provide an index of the extent to which the release is displaced earlier in time from V2 acoustic onset, relative to the duration of the pause. Larger values indicate a earlier displacement of the release. The critical observation is that in the trochaic V.CCV environment, the range of posterior predicted means for the measure spans the ranges that are associated with coda-syllabification and dissociation patterns, which are distinct in the other environments. Hence the results suggest that in this environment, there is interspeaker variation in timing pattern.

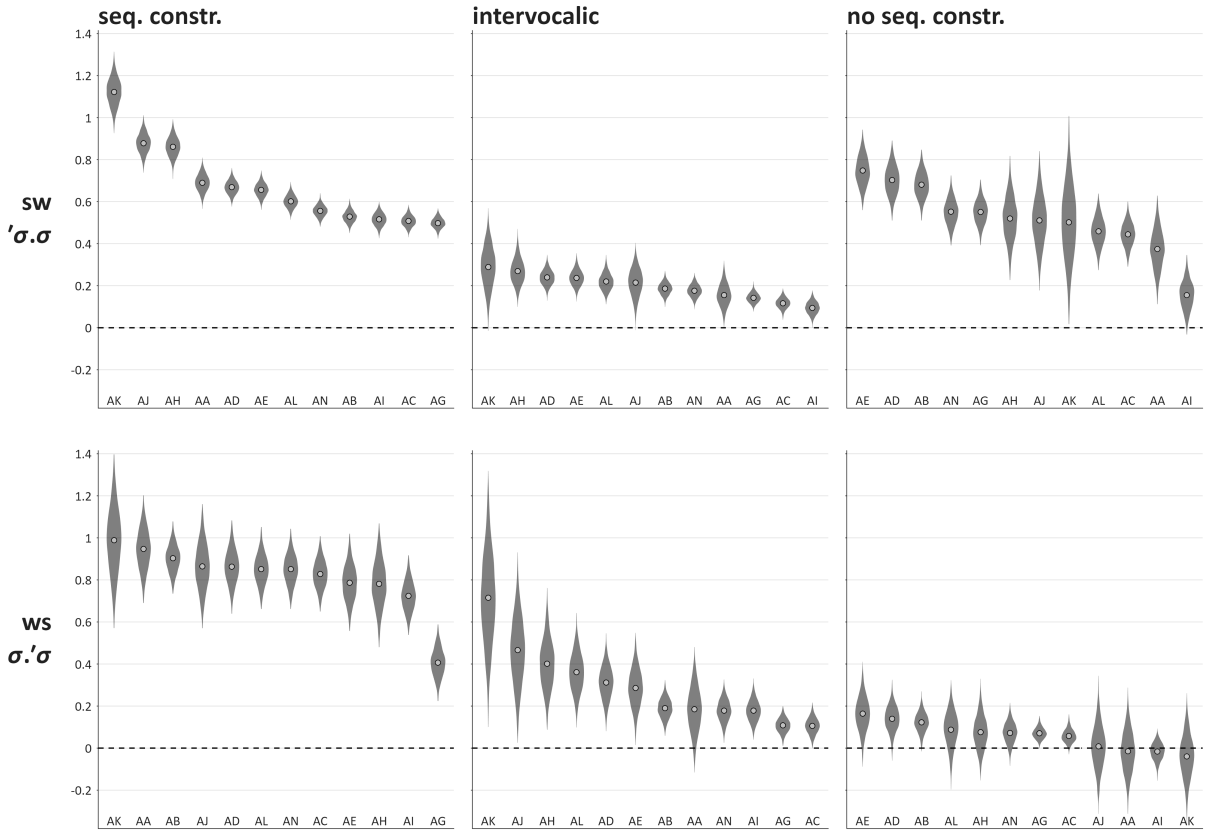

Fig. A1. By session timing patterns. Distributions represent the posterior predicted difference in the release - <V2 interval scaled by mean of the posterior predicted difference in the closure – V1> interval. Larger values indicate a greater displacement of the release toward the preceding value.

Fig. A2 shows by-consonant posterior predicted distributions of mean closure and release timing, relative to the acoustic onset of the following vowel. Posterior mean segmental intervals are shown as well. Of particular interest are the timing patterns in the trochaic V.CCV environment (no sequencing constraint), where it is evident that the timing of release of bilabial stops is more consistent with ambisyllabic dissociation, while timing of release of alveolar stops is more consistent with coda-syllabification.

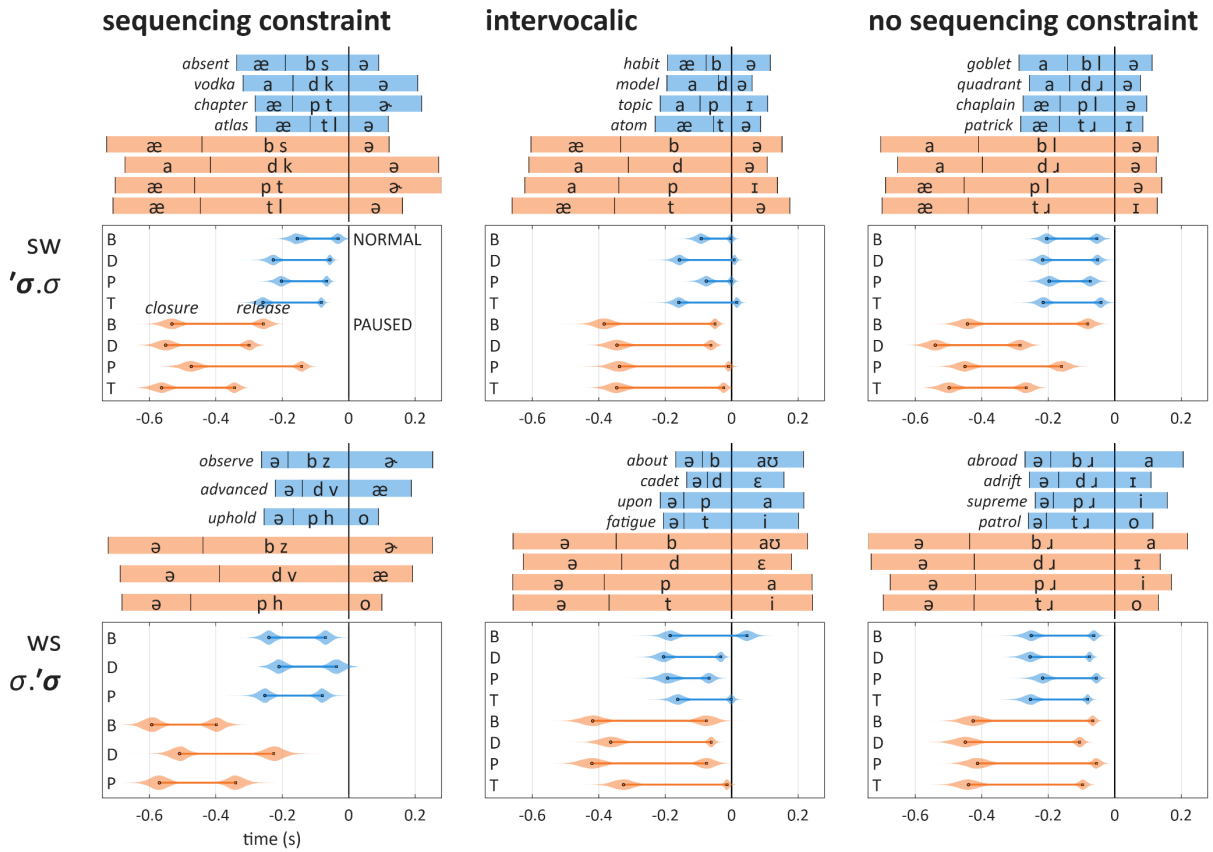

Fig. A2. Constriction formation and release timing by environment and consonant. Mean posterior segmental intervals are shown, along with posterior mean distributions of timing of constriction formation and release relative to acoustic onset of the following vowel.