Mining speech corpora for differences in gestural timing as a precursor to metathesis

Tyler Lau*

1 Introduction

Gestural timing has been adduced as a factor in various sound changes, including deletion, epenthesis, and assimilation. These changes may be triggered by the obscuring of an acoustic signal due to overlapping articulations, thus leading to ambiguity in the reconstruction of the underlying phonemes by the listener. If listeners misconstrue an intended signal, the phonemic representation of a phonetic sequence will change from speaker to listener, yielding a sound change. Metathesis, the reordering of phonological segments in a word (Grammont 1933; Hock 1985), has been shown through both diachronic and experimental data to be linked to perceptual biases in segmental order arising from phonetic properties of the phonemic segments. The role of production in triggering metathesis, however, has been less explored. While early works suggested that metathesis was primarily a result of sporadic speech errors, later works found structured synchronic and diachronic metatheses.

Metathesis between sibilants and stops is relatively common diachronically and is related to properties of both sibilants and stops. The difference in perceptibility and acoustic properties of sibilants and stops leads not only to obscuring of the stop gesture but also to a phenomenon known as *auditory decoupling*, in which sounds of highly differing frequencies are separated in the speech stream, leading to potential confusion of their temporal order Bregman and Campbell (1971). Blevins and Garrett (2004:140) note that Late West Saxon English and colloquial French show different directionality in metathesis of sibilant-stop (specifically *sk/ks*) clusters and suggest that the difference in the directionality of metathesis may be due to different stress patterns in the two languages. Such an explanation may be related to the degree of gestural overlap. A lower degree of gestural overlap allows more sibilant noise and subsequently a greater chance of segmental order confusion.

Metathesis has also been suggested to be linked to the degree of gestural variability by Yanagawa (2003), who conducted a production study on Hebrew. Yanagawa concluded

*Many thanks to Susan Lin and Susanne Gahl for guidance throughout this project and to Andrew Garrett, Keith Johnson, Brian Smith, my fellow classmates in Linguistics 201 (Advanced Graduate Proseminar) and the audience at the 42nd Penn Linguistics Conference for helpful comments. Special thanks as well to my undergraduate research apprentice, Anstonia Ma, and to Ronald Sprouse and Dmetri Hayes for help and suggestions in scripting the program to gather data from the Buckeye Corpus.

that greater variability in gestural timing occurs in Hebrew in positions in which actual metathesis occurs between sibilants and stops, suggesting that the degree of variability in gestural overlap is also a relevant production factor in metathesis.

The hypothesis that a lower magnitude of gestural overlap in sibilant-stop (henceforth ST/TS) sequences can increase likelihood of metathesis has yet to be tested. This study will explore the behavior of these clusters in an acoustic corpus in order to assess the relationship between likelihood of metathesis and magnitude of gestural overlap and explore Yanagawa's hypothesis about the effect of gestural variability on metathesis. I employ the Buckeye Corpus, consisting of recordings of interviews with 40 speakers from Columbus, Ohio, in order to assess the degree of and variability in gestural overlap in ST/TS sequences in various contexts in naturalistic speech. The goals of the study are twofold. The first goal is methodological: to determine whether findings in previous perceptual and articulatory experimentation work can be reproduced in naturalistic acoustic data. The second goal is to observe how magnitude and variability of sibilant duration in these clusters vary by different properties of the cluster, including position in the word, whether it contains a morpheme boundary, sibilant-stop order, voicing, and place of articulation, and to explore the implications for metathesis. I find that greater and more variable sibilant duration is robustly exhibited in word-medial TS clusters as compared to ST ones, a position in which diachronic metathesis has been well-documented, corroborating hypotheses that link metathesis to sibilant noise and gestural variability. I also find that in cases in which metathesis would lead to a perceptually suboptimal sequence, there is attenuation of sibilant noise, an effect that is potentially the result of phonetic properties being altered to improve perceptibility. The findings also contribute to explorations of the acoustic correlates of gestural overlap, providing a fruitful ground for comparative work with articulatory data.

2 Background

Early literature describes metathesis as an irregular and sporadic sound change (Osthoff and Brugmann 1878:xiv: n. 1; Grammont 1933; Hockett 1958; Bloomfield 1962; Lehmann 1962), often manifesting in speech errors in the form of spoonerisms, such as ***k**eep a **t**ape* → ***t**eep a **c**ape* (Fromkin 1971:30; Spencer 1996:68; Crystal 1997:240). Because of its irregular appearance in these cases, metathesis has been described as a sporadic change that is explained by performance factors and not by phonological systematicity (Powell 1985:106; Montreuil 1981:67). However, a closer look reveals that metathesis is phonologically structured and that its directionality is often predictable. Metathesis can occur either synchronically as a variant process that tends to be particularly noticeable in children's speech errors (1a), a morphophonological process (1b), or diachronically (1c). It may also occur between adjacent segments, as in (1a) and (1b), or at a distance, as in (1c).

(1) a. Dutch *asteri**sk*** ∼ *asteri**ks*** 'asterisk' (Harkema 1999)

b. Lithuanian *blo**ʃk**-e* 'toss (3sg.past)' ∼ *blo**kʃ**-k* 'toss (imper.sg.)' (Hume and Seo 2004:37)

c. Latin *parabola* > Spanish *palabra* 'word' (Penny 2002:36)

Various factors have been cited to explain the occurrence of metathesis. Blevins and Garrett (2004:120) lay out a typology of metathesis, categorizing each type as deriving from different sources. Table 1 provides examples of each type of metathesis. Perceptual metathesis involves segments with long phonetic cues, such as liquids and pharyngeals, and is argued to occur because of the perceptual difficulty of locating the position of segments with long cues, particularly as coarticulatory effects of these cues carry over for several syllables (Kelly and Local 1986; Heid and Hawkins 2000). Compensatory metathesis occurs between a prosodically weak vowel and a neighboring consonant, bringing the vowel into a prosodically strong position (Blevins and Garrett 1998). Coarticulatory metathesis involves segments with significant gestural overlap—a common example is the metathesis of labial-velar sequences into velar-labial sequences. This directionality is argued by Blevins and Garrett (2004) to be due to the tendency for velar closures and releases to precede labial ones (Connell 1994). Finally, auditory metathesis is described as involving confusion of segmental order due to the separation of sibilant noise from the rest of the speech stream in an auditory signal and so is argued by Blevins and Garrett to occur primarily, if not solely, in ST/TS sequences. Auditory metathesis is argued to have its source in misperception, as it is the nature of the high frequency of sibilant noise that is argued to lead to confusion for the listener. Since both sibilant noise and segments with long phonetic cues can cause perceptual difficulty in placing the linear order of the signal, auditory and perceptual metathesis may be argued to be the same subtype of metathesis, both driven by difficulty in locating a phonetic segment due to the nature of the signal. The difference between the two lies primarily in that perceptual metathesis may occur distantly.

Type	Example
Perceptual	Prakrit /mahisa/ > Marathi /m^{h}ais/ 'buffalo'
Compensatory	Rotuman /seséva/ → [seséav] 'erroneous'
Coarticulatory	Cebuano /libgus/ ∼ Aklanon /ligbus/ 'mushroom'
Auditory	Old Dutch *wepse* > Dutch *wespe* 'wasp'

Table 1: Types of metathesis

I focus on ST/TS clusters in this study, labeled by Blevins and Garrett as auditory metathesis, and will explore the role of sibilant noise in metathesis. While auditory metathesis is primarily related to perceptual factors, I argue that production plays a role because magnitude of and variability in gestural overlap can lead to an ambiguous signal. This ambiguous signal can then trigger a change in the perception of segmental order and lead to metathesis. In this way, auditory metathesis shares with coarticulatory metathesis the property of being related to effects from gestural overlap.

2.1 Explanations for metathesis

Perceptual explanations for metathesis primarily focus on the confusion of the segmental order of consonants due to an ambiguous phonetic signal. ST/TS sequences are notable in this regard because of the large difference in perceptibility between fricatives (especially sibilants) and stops. The manner of articulation of both stops and fricatives can be mostly determined without contextual cues: a (voiceless) stop can be identified by a period of silence and is distinguished from a lack of a phone by a burst and the transitions into or out of a neighboring vowel. Fricatives are easily identified by the presence of noise, which is particularly high in frequency and amplitude in the case of sibilants. The identification of place of articulation, on the other hand, is heavily context-dependent for stops. The burst of the stop and F2 transition into or out of the stop is particularly informative. The place of articulation of a fricative, however, can be identified by both the amplitude and the center of gravity as the noise in the spectrum is characteristically different for each fricative (Wright 2001). Thus, while the perception of a stop's place of articulation is more reliant on surrounding phones for transitional cues than that of a fricative.

Listener- and speaker-oriented explanations have both been put forth for sound change. With respect to metathesis, explanations center around ambiguity in segmental order due to the auditory signal. The listener-driven approach attributes metathesis to misperception of the segmental order by listeners, an approach that builds off Ohala's listener-based model of sound change (Ohala 1981, 1993). Under this view, the phonemic realization of a word changes from speaker to listener because the speaker interprets the ambiguous phonetic signal with a different underlying representation (Blevins and Garrett 1998, 2004; Hume 2004). Hume's account, known as the *indeterminacy/attestation hypothesis*, claims that the acoustic signal of the sequence must not only be ambiguous, but also that the metathesized outcome must also be attested in the language and that phonotactic frequency will also affect directionality of metathesis. The underlying claim, then, is that metathesis is structure-preserving (whereas other processes such as deletion and epenthesis are not).

Hume's argument for attestation as a necessary prerequisite comes from the fact that listeners are biased to perceive impermissible consonant clusters in a form that is permissible in their language. For example, English speakers are more likely to perceive epenthetic schwas in impermissible clusters ([tlæ] → [təlæ]) than in permissible ones ([træ] → [təræ]) (Pitt and McQueen 1998). Both phonological context and frequency are also important in perception—Mielke (2001) studied the perception of [h] in different phonological contexts by speakers of Arabic, Turkish, English, and French. He found that /h/ was more perceptible to Arabic and Turkish speakers than to English and French speakers in both pre- and post-vocalic postion and that in pre-vocalic position, it was more perceptible to English than to French speakers. The salience of /h/ prevocalically for Arabic, Turkish, and English speakers over French ones and of /h/ postvocalically for Arabic and Turkish speakers over English ones reflects the general phonotactics of the languages—Arabic and Turkish have /h/ in both pre- and post-vocalic position, English only has /h/ prevocalically, and French has no /h/ in its phonemic inventory. The salience of /h/ *even* in pre-vocalic position for Arabic and Turkish speakers over English ones, on the other hand, reflects the higher frequency of /h/

in Arabic and Turkish. Hume (2004:23), drawing from a database of 34 cases of synchronic consonant/consonant metathesis, notes that there are no cases of metathesis into a sequence unattested in the language. Notably, however, Blevins and Garrett (1998) point out various examples in which metathesis is *not* structure-preserving. As an example, Rotuman *seséva* is metathesized to *seséav*, which not only changes the CVCV template, but also creates a coda.

A speaker-oriented approach to sound change argues that metathesis occurs in order to optimize the perception of weakened phonetic cues (Hume 1998; Steriade 2001). In looking at ST/TS sequences in which sibilants are significantly more perceptible than a neighboring stop, one possible consequence of the reduced perceptibility is simply deletion, which would be predicted by listener misperception. The deletion of stops that lack a neighboring vowel is a very common process both synchronically and diachronically. The examples in 2 demonstrate deletion of stops in different reduced-perceptibility environments.

(2) a. #TC > #C: Buchan Scots [gnjaːv] ~ [njaːv] 'gnaw' (Kökeritz 1945:79)
b. CT# > C#: Ibiza Catalan /pɔnt/ → [pɔn] 'bridge' (Wheeler 2005:73)
c. CTC > CC: Faroese /skarp-t/ → [skart] 'sharp, shrivelled (neut. sg.) (Hume and Seo 2004:37)

Deletion of stops is a likely consequence due to the lack of perceptibility of the stop in these positions, although in the case of a stop flanked by two consonants, it is possible that gestural overlap simply masks the auditory cues, a finding shown by articulatory studies to occur in phrases such as *perfe**ct** **m**emory* (Browman and Goldstein 1989:215–216). Both acoustic and articulatory explanations predict pressures that will lead to deletion over time.

Both the ambiguity and perceptual optimization hypotheses can also explain why metathesis would occur in such situations. There may be biases to perceive consonant sequences containing stops in an order in which the stop is more perceptible. Alternatively, speakers may pronounce words such that segments that are in suboptimal positions are produced in a more perceptually prominent position as an optimizing strategy. These segmental reorderings trigger metathesis with an adjacent consonant. ST/TS sequences are a particularly apt example that can be predicted to lead to deletion or to metathesis. The significant noise in sibilants is known to mask acoustics of a neighboring stop (Mielke 2001). The masking may lead to misperception of the sequence without the stop, which would cause deletion, or it could lead to the speaker's optimization of the sequence to improve the phonetic cues of the stop, resulting in metathesis. *Auditory decoupling*, a process in which high and low tones are perceived by listeners in separate streams, is also relevant. In music, this can lead to confusion in the order of notes. In speech, because sibilant noise contains high frequencies, it may also be decoupled from the rest of the speech stream, leading to a possibility of confusion of segmental order (Bregman and Campbell 1971; Bregman 1990; Ladefoged 2001). Thus, the specific nature of sibilant noise plays a key role in creating ambiguous conditions that catalyze metathesis.

Since the place of articulation of a stop is most clearly cued by vocalic transitions, stops should be predicted to metathesize into positions adjacent to a vowel. In the case where a cluster containing a stop is flanked by vowels, there is a bias for the stop to be prevocalic rather than postvocalic. Fujimura et al. (1978) demonstrate in an experiment that if the place cues of CV transitions into a vowel conflict with those of VC transitions out of a vowel, American English and Japanese speaking subjects are biased to perceive the consonant with the place identified by the CV transition.[1] Examples from Steriade (2001:231) of the predicted directions from these two generalizations are provided in (3)[2].

(3) a. #TSV → #STV: Greek *p*^{h}*syk*^{h}*e:* > *sp*^{h}*yk*^{h}*e:*
b. ST# → TS#: Southern American English *wasp* → *waps*
c. VSTC → VTSC: Lithuanian *dresk-ti:* → *dreks-ti*
d. VTSV → VSTV: Rural Latin *ipse* → *ispe*

While these asymmetries are predicted, Hume (2004) demonstrates that frequency can produce metathesis in an unexpected direction. In Mutsun, the locative suffix has two allomorphs: [-tak] following consonants and [-tka] following vowels. The nominal thematic plural suffix also has two allomorphs: [-mak] following consonants and [-kma] following vowels. Examples from Hume (2004:27) are provided in (4).

(4)	a.	lo:t	lo:t-tak	‘mud’
		si:	si:-tka	‘water’
	b.	ru:k	ru:k-mak	‘string’
		sinni	sinni-kma	‘child’

Following the locative suffix, the nominal thematic plural suffix should be expected to be *[-mka] following vowels, both due to preserving segmental order and also to the increased perceptibility of [k] in prevocalic position. However, while sequences of [m] + obstruent are attested in the language, [mk] is either rare or nonexistent in Mutsun. This (near-)zero frequency of the sequence leads to a strong pressure for metathesis, even when the result yields a sequence with reduced perceptual cues.

Two other counterexamples to metathesis directionality are noted by Blevins and Garrett (2004:139–140). In the change from Old English to the Late West Saxon dialect, VSTV is unexpectedly metathesized to VTSV. They also cite the unexpected direction of VTS# > VST# in colloquial French in an example from Grammont (1923:73). Examples are provided in (5).

(5) a. Old English *waskan* > Late West Saxon *waksan* ‘to wash’
b. Standard French *fiks* > Colloquial French *fisk* ‘fixed’

The Late West Saxon metathesis leaves the stop in a postvocalic rather than a prevocalic position, whereas the colloquial French metathesis leaves the stop wedged between a sibilant and silence. Steriade (2001:231) suggests that the French metathesis does not violate

[1] Interestingly, American English speakers have a greater tendency to identify the place with the VC transition if the accent pattern is high-low, however, whereas Japanese speakers are unaffected by the accent pattern.

[2] Steriade does not provide glosses for these words.

directionality because of released codas in French (thus, the phonetic realization of /fisk/ would be [fisk]); this explanation does not account for the VSTV > VTSV change, however. Blevins and Garrett instead suggest that stress may be at work in both cases. The ST sequences in all the Old English examples immediately follow tonic stress, whereas French has (weak) final stress. The post-tonic position of the ST sequence in Old English and the word-final position of the sibilant in TS sequences in French means that the sibilant in both cases could be lengthened. The lengthened sibilant could lead to confusion in segmental order due to auditory decoupling and thus catalyze metathesis. While Blevins and Garrett's stress hypothesis has not yet been tested experimentally, Lunden and Renoll (2015) demonstrate that in an experiment involving speakers reading 5-syllable nonce words, long-distance metathesis did not target the onset or any stressed syllables. Thus, stress appears to be a relevant factor in metathesis.

Perceptual biases in segmental order match up with attested directions of metathesis. Particularly for ST/TS sequences, in which the cues for the sibilant are particularly salient in comparison to those of the stop, even greater diminishment of the stop's cues can be expected to lead to misperception of the stop placement or to perceptual enhancement via metathesis by the speaker.

2.2 Previous experimental work

Experimental work on metathesis corroborates diachronic evidence by demonstrating that listeners are biased to perceive consonant sequences in perceptually optimal orders. Makashay (2001) tests perception of obstruent order in English via a lexical decision task. Words containing intervocalic obstruent clusters were altered such that the clusters were reversed to create non-word stimuli (such as *whiksy* from *whisky* and *taski* from *taksi*). Participants were asked to judge whether the stimuli were English words or not. The results show that participants were more quick to judge words that were less frequent and that had poorer perceptual cues as English words. For example, a stimulus such as *whiksy* would be more quickly identified as an English word than would a stimulus such as *taski.* Because the postvocalic [k] in *whiksy* has poorer transitional cues than the prevocalic [k] in *taski*, it is more ambiguous, and thus more likely to be "unmetathesized" by the listener.

Graff and Scontras (2012) conduct an experiment to test perception of aSTa/aTSa (as well as aNTa/aTNa[3]) sequences via a forced choice task asking subjects if the target consonant (either the non-stop or the stop) came first. They find that participants were more likely to mishear [aksa] as [aska] (and [atna] as [anta]) than vice versa, showing a bias toward perceiving stops in prevocalic position, the same bias toward CV over VC transitions found by Fujimura et al. (1978). Furthermore, if the stop burst (a crucial cue) is removed, the metathesized percept becomes even *more* likely when the non-stop is a sibilant, but *not* if it is a nasal, suggesting that sibilants hinder the perception of the stop more than nasals do. Jones (2016) explores the perceived order of consonant clusters in real Hebrew words by English speakers, concluding that fricatives and sibilants lead to significantly higher reaction

[3] Where N stands for a nasal.

times in subjects' determination of the segmental order of the cluster. These experiments show that misperceptions of order occur in the predicted direction of metathesis and also that sibilants (and perhaps fricatives in general) play a special role in the perception of segmental order.

In contrast to Hume's 2004 findings in Mutsun, Graff and Scontras (2012) demonstrate that while intervocalic /Tʃ/ is more than twice as frequent as /ʃT/, listeners are still biased to perceive the nonce-word [akʃa] as [aʃka]. Jones (2016) also finds that despite the higher phonotactic frequency of [dz] over [zd] in English, English speakers are more likely to hear Hebrew [dz] as [zd] than vice-versa. Thus, perceptual cues take precedence over frequency in these contexts. While metathesis may repair unattested or extremely infrequent obstruent clusters into attested or very frequent ones, even if the cues are weaker, it is possible that there is a ceiling effect on the infrequency of the consonant sequence—once an ordering is sufficiently frequent, perceptibility may take precedence over attestation.

While experimental work on production factors in metathesis has been limited, there is notable work on gestural overlap of consonant sequences. The degree and variability of gestural overlap are both shown to be affected by both positional and morphological factors.

Early articulatory studies show that onset clusters and coda clusters have different patterns of gestural overlap. Browman and Goldstein (1988:149) demonstrate that the c-center (the mean of all the midpoints of the gestures in a given consonant sequence) of onset clusters is timed in-phase with the peak of the following vowel gesture whereas for coda consonants, it is the left edge that is timed with the preceding vowel. However, (Byrd 1995:302) provides conflicting evidence that some speakers time the c-center of coda clusters with the preceding vowel instead. It has also been shown that there is less temporal overlap of consonant gestures in word onsets than in other positions. Chitoran et al. (2002:422) suggest that the reason may be first the fact that since word onsets can be utterance onsets, they may not have any acoustic information (such as VC formant transitions) preceding them. Thus, the overlap of an onset consonant cluster may be restricted as to maximize acoustic information. The pressure to maximize information via minimization of overlap may also come from the fact that word onsets are privileged in lexical access (Marslen-Wilson 1987). Medial and coda clusters have been shown to overlap more than onset ones in Georgian (Chitoran et al. 2002) and English (Byrd 1996), respectively. Greater overlap of clusters suggests less sibilant noise in the case of clusters containing sibilants, which would then predict less metathesis in word-onset position. Mielke and Hume (2001) indeed find in their typological study on metathesis (although not specifically ST/TS metathesis) that it avoids initial syllables.

Hoole et al. (2013:86) explore onset cluster overlap in German and French and find that clusters ending in /r/ display less overlap than those ending in /l/; under the c-center hypothesis, the lower degree of overlap in /r/ should lead to greater overlap with the following vowel, which could lead to metathesis. Metathesis of rhotics with following vowels is well-attested (Icelandic *hross* ∼ rhotic English *horse*). Hoole et al. also demonstrate that German conforms to the c-center hypothesis, with the right edge of /r/ in /traːt/ shifted to the right and the left edge of the /t/ shifted left with respect to the coda as the anchor point, in comparison to the simplex onset in /taːt/. However, French interestingly shifts the initial

/b/ in the onset of /brak/ significantly left and the /r/ only slightly right (once again, with the coda as the anchor point), compared to /bak/, preventing much overlap with the following vowel, and therefore violating the expected outcome of the c-center hypothesis. Hoole et al. suggest that languages may differ in their overlap settings for consonant clusters and that in some language, like Slovak, a low overlap setting may lead to the emergence of syllabic consonants (Hoole et al. 2013:94). One might also expect that metathesis could emerge as a consequence of the low overlap, as there would be greater overlap with the subsequent vowel. Magnitude of gestural overlap is cited as an explanation for metathesis of stop-clusters. Blevins and Garrett (2004:136–137) cite /pk/ > /kp/, /tp/ > /pt/, /tk/ > /kt/ as examples of stop-stop metatheses that are unidirectional. The coarticulation and near-simultaneous closure of the two stops is often accompanied by a later release of labials after velars and coronals after noncoronals. If the release of the latter closure is taken as the crucial signal, the order of the stops may be reanalyzed to treat the latter release as the second consonant sequentially, thus leading to metathesis.

A number of studies looking at duration of English morphemes in both controlled experiments and in corpora have shown that morphological status affects the duration of a segment. In one study, Plag et al. (2017) demonstrate that in 600 tokens taken from the Buckeye Corpus, there are significant differences in the durations of different *-s* morphemes (or lack thereof). Plag et al. find that voicing and morpheme status interact, such that non-morphemic /s/ is significantly longer than all morphemic /s/ if they are voiceless, but that the plural marker *-s* and genitive plural marker *-s'* are significantly longer than all other /s/ (both morphemic and non-morphemic) if they are voiced.

Seyfarth et al. (2018) employ a reading task with 40 subjects to test the difference of final [s, z] suffixes from final [s, z] in uninflected words. They found that both stems *and* suffixes have significantly longer durations than segmentally identical uninflected words (for example, *free-s* vs. *freeze*), reproducing results from Walsh and Parker (1983). This finding, however, runs counter to the findings from Plag et al. (2017), a difference that they suggest may be an artifact of unbalanced data in Plag et al.'s study. However, they do not find a significant difference between words with [t, d] suffixes and uninflected words ending in [t, d] (echoing a result from Losiewicz (1995)), which they surmise may be a result of the fact that nouns tend to be longer than verbs and the fact that their fricative-final pairs were balanced for part of speech, but that their stop-final pairs were mainly verbs in inflected words but nouns in uninflected ones, leading to a greater likelihood for a lack of difference. This finding for final fricatives is also compatible with the view that there is lower gestural cohesion across a morpheme boundary.

These studies show contradictory findings as to the effect of morphemic status on the duration of a phonological segment. One question is whether the difference in the duration is due to a different degree of overlap between adjacent gestures across morpheme boundaries as opposed to within the same morpheme. Other studies directly explore the effect of morphology on gestural timing. Tapio (2008:23) demonstrates that segments before morpheme boundaries are longer than those not before boundaries and that the duration appears to correlate with lexical frequency, such that less frequent words show longer durations before boundaries. This finding suggests that gestural overlap may be decreased at morpheme boundaries.

Cho (2001) demonstrates that gestural timing in Korean is more variable across morpheme boundaries (heteromorphemic) than within the same morpheme (tautomorphemic), challenging the idea that morphology does not interact with phonetics (Kiparsky 1982; Levelt et al. 1999). Yanagawa (2003) replicates this finding in Hebrew and also shows that variability is also greater in medial position than in initial position. Yanagawa argues that the difference in gestural timing by word position and morpheme status are potential sources for metathesis. In Hebrew, metathesis is seen in one morphophonological environment: at a morpheme boundary in the *hitpa'el* binyan, which expresses the reflexive, reciprocal, and inchoative. This metathesis is also limited to roots beginning with sibilants /s, z, ʃ, ts/. Examples of *hitpa'el* verbs are provided in (6).

(6) a. No metathesis: /hit-labeʃ/ → [hitlabeʃ] *[hiltabeʃ] 'he got dressed'
b. Metathesis: /hit-sarek/ → [histarek] *[hitsarek] 'he combed his hair'

Yanagawa's findings apply to non-ST/TS sequences, but metathesis operates specifically on heteromorphemic medial TS sequences. She argues that the canonical perceptual account of TS > ST to improve perceptual cues is insufficient for the Hebrew case as TS sequences that are either in word onset position or that are morpheme-internal do *not* metathesize, as shown in (7), even though they should also be expected to do so.

(7) a. Onset position: /t-saper/ → [tsaper] *[staper] 'you (m.sg.) will tell'
b. Morpheme-internal: /hi-tsis/ → [hitsis] *[histis] 'it fermented'

Yanagawa takes an Optimality-Theoretic approach to explaining why metathesis occurs specifically in TS sequences by citing previous work showing that sibilants resist coarticulation with following vowels (Stoel-Gammon 1985; Recasens 1989; Recasens et al. 1992). The looser gestural cohesion (defined by greater gestural variability) in an environment that is the combination of three different cases—(1) between sibilants and a following vowel, (2) in word-medial position, (3) in heteremorphemic clusters—leads to the metathesis that is seen in Hebrew. In order to account for why only sibilants undergo metathesis, Yanagawa argues that given loose enough gestural cohesion, metathesis will be triggered to shift a sibilant away from a following vowel, contrasting with views in which metathesis occurs to shift a stop to a perceptually optimal position.

This study will respond to these past acoustic and articulatory findings through exploration of ST/TS clusters in the Buckeye Corpus, which was also employed by Plag et al. (2017). As the Buckeye Corpus consists of naturalistic data, several variables will be controlled for in an attempt at comparability with experimental data. The mechanisms by which factors in gestural timing may lead to metathesis will be explored in greater detail in this study.

Perceptual experiments have shown that TS sequences flanked by vowels are more likely to be perceived in ST order than vice-versa, a finding that accords with the increased perceptibility of a stop in prevocalic position as compared to postvocalic position. This finding also conforms to diachronic tendencies for stops to metathesize into prevocalic position. One open question is whether sibilant duration may also play a role in these asymmetrical findings. If Blevins and Garrett (2004) are correct in attributing sibilant-stop metathesis to greater sibilant noise duration leading to confusion of segmental order due to auditory decoupling, we may expect to see greater sibilant duration in clusters with postvocalic stops than in clusters with prevocalic stops. A longer sibilant gesture should lead to a greater likelihood for metathesis to occur. Differences in ST/TS clusters may also be mediated by position in the word. Diachronically, TS > ST is more frequent in onset and medial position, but the opposite pattern, ST > TS is more frequent in final position. Thus, it is possible that the effect is reversed in word-final position.

Gestural variability can be expected to contribute as well due to the difficulty of a secure underlying representation for a phonetic sequence if it is highly variable combined with the fact that sibilant noise already leads to the difficulty in localizing the cues. Similarly, if Yanagawa (2003) is correct in greater gestural variability being an important factor in metathesis, then greater variability may be expected in clusters with postvocalic stops than in clusters with prevocalic stops. It is this increased ambiguity caused by magnitude and variability of sibilant noise that could catalyze the process of metathesis.

2.3 Measures of Gestural Overlap in ST/TS Clusters

Gestural overlap cannot be directly measured when only acoustic information is available. Thus, a simplifying assumption is made that the ST ratio corresponds well to the degree of gestural overlap. Because a stop involves a complete closure of the oral cavity, airflow is blocked and sibilant noise cannot occur during the stop closure. Thus, the duration of sibilant noise should be relatively longer if there is less gestural overlap and shorter if there is greater gestural overlap. This difference is schematized in Figures 1 and 2, with each curve representing the movement of one gesture towards its intended target and then away from it.

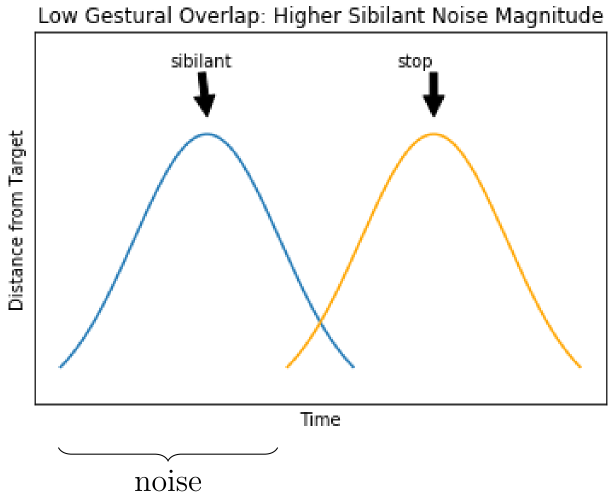

Figure 1: Greater sibilant noise

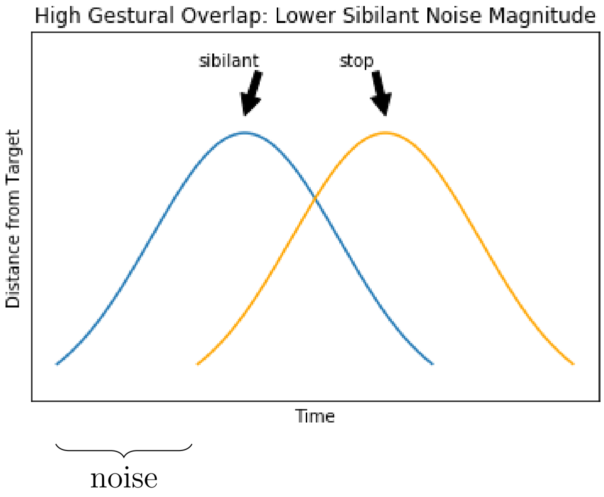

Figure 2: Lesser sibilant noise

In Figure 1, the stop gesture begins late, so sibilant noise carries through a longer duration than in Figure 2, in which the earlier onset of the stop gesture cuts off the sibilant noise earlier, resulting in a relatively shorter duration of sibilant noise and thus a lower ratio of sibilant to stop duration. Differences in gestural variability are schematized in Figures 3 and 4. The orange curves represent different instantiations of the stop gesture, some of which begin earlier and some of which begin later.

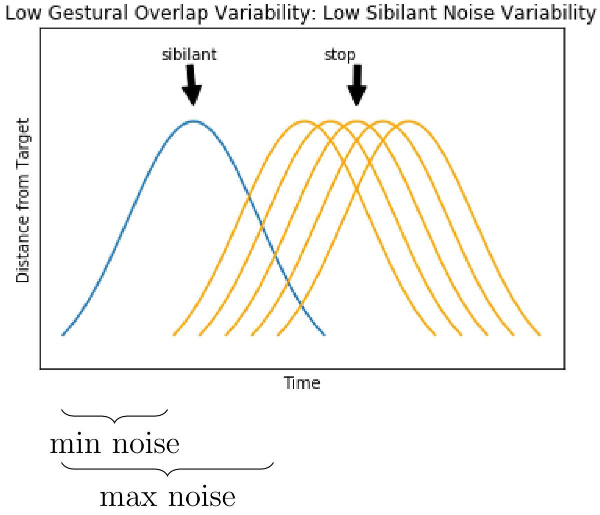
Low Gestural Overlap Variability: Low Sibilant Noise Variability
sibilant
stop
Distance from Target
Time
min noise
max noise

Figure 3: Lower sibilant noise variability

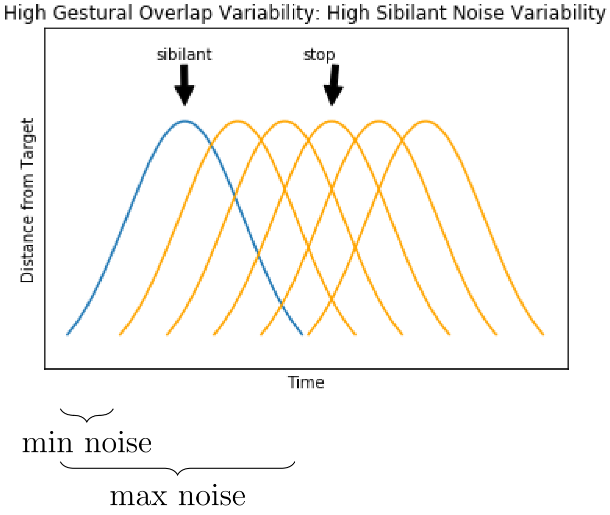
High Gestural Overlap Variability: High Sibilant Noise Variability
sibilant
stop
Distance from Target
Time
min noise
max noise

Figure 4: Greater sibilant noise variability

If there is low variability in gestural overlap as in Figure 3, we should expect to see less variability in sibilant noise as the minimum and maximum duration of sibilant noise in a given cluster will be more constrained. Conversely, high variability in gestural overlap as in Figure 4 should be correlated with high variability in sibilant noise.

3 Methodology

The Buckeye Corpus of Conversational Speech (Pitt et al. 2007) consists of approximately 300,000 words from recordings of interviews with 40 native speakers of American English from Columbus, Ohio. All speakers were middle-class Caucasians but were balanced for age (under thirty vs. over forty), gender (female vs. male), and gender of the interviewer (female vs. male). The interviews were structured to be “friendly conversations” aimed at eliciting naturalistic speech and lasted between 30 and 60 minutes. All the interviews were phonetically time-aligned, allowing for large-scale analysis of naturalistic data. The boundaries between stops, sibilants, and vowels is quite clear: the first is marked by a period of silence potentially accompanied by a burst, the second by high frequency noise, and the third by a periodic waveform. The transition between any of these two segments is rather noticeable and so the annotations that have been made for the corpus will be relied upon for durational measures.

ST/TS sequences appearing in the *phonetic* transcription[4] were extracted from the corpus. The sibilants of English are /s z ʃ ʒ/ while the stops are /p t k b d g/. Of these tokens, only sequences in which the stop was adjacent to a vowel were considered as the hypotheses about magnitude or variability of sibilant noise when the stop in the cluster is not adjacent to any vowel are unclear. Following this removal, 11,536 tokens remained for analysis.

3.1 Assessing Differences in Gestural Magnitude

The variable of interest in this study was the relative duration of the sibilant to the stop in ST/TS clusters. To estimate the relative duration while controlling for speech rate, the ratio of the sibilant duration to stop duration was taken as a measure of normalized duration instead of the raw duration (measured in seconds). As distributions of duration tend to be right-skewed, the sibilant and stop durations were first logged (after shifting the values to the right by 1 second, in order to prevent negative log values). The ratio of the logs was then taken as the dependent measure. This variable stood in as a proxy for the relative duration of the sibilant to the stop and will be referred to as ST Ratio (sibilant to stop ratio).

The predictor variables are as follows:

Position in the Word. The cluster may be *initial*, *medial*, or *final*. Given previous literature showing greater gestural overlap of consonant clusters in non-onset positions, the ST ratio should accordingly be lower in medial and final position than in onset position. Greater gestural variability in non-initial position has also been demonstrated and so the ST ratio should also be more variable in non-initial position.

Morpheme. Morpheme boundaries are not marked in the Buckeye Corpus. Thus, a script was written to automatically tag clusters ending in phonetic [s] or [z] and orthographic <s> as containing a morpheme boundary and those ending in phonetic [t] or [d] and orthographic <ed> as containing a morpheme boundary. All other ST/TS clusters were tagged as non-morphemic. Due to the limited information in the Buckeye corpus labels, the different types of morphemes have not yet been further subdivided. Both morphemes have several meanings. *-s* may be the pluralizer, genitive marker, third person singular marker, citicized version of *is*, or the citicized version of *has*. *-ed* may be the past tense marker or the perfect marker. Longer and more variable sibilant noise should be expected heteromorphemically as opposed to tautomorphemically.

Cluster Order. Clusters are coded as STV (the stop is prevocalic) or VTS (the stop is postvocalic). Theoretically, the difference between these sequences can only be compared if the cluster itself is in word-medial position; however, to prevent this limitation, a cluster in which a word boundary exists between the stop and the vowel is still coded as if there is no word boundary (i.e. ST#V = STV and V#TS = VTS), such that effects can be observed in onset and coda position as well.[5] Because the Buckeye Corpus involves running speech,

[4] There are many cases of a phonemic ST/TS cluster not appearing in the phonetic form: for example, word final /t/ is often dropped after an /s/, as in /wɛst/ → [wɛs].

[5] Since the model also takes position in the word into account, differences between ST#V and STV sequences as well as between V#TS and VTS sequences can be observed in the comparison of word-final STV sequences to non-word-final ones and word-initial VTS sequences to non-word-initial ones.

a following vowel, even if it is in a separate word, may still provide a stop with suitable transitional cues. The word boundary is only not ignored if there is a significant period of silence demarcating it, as annotated in the corpus. VTS sequences should show longer and more variable sibilant noise than STV ones.

Voicing Given the findings from Plag et al. (2017), an interaction between Voicing and Morpheme is expected. If the results of Plag et al. obtain here as well, homomorphemic clusters with voiceless sibilants should be longer than heteromorphemic ones, and voiced sibilants in heteromorphemic clusters may be longer than ones in heteromorphemic clusters.

Place of Articulation Different patterns of gestural overlap should be expected depending on the place of articulation of the stop adjacent to the sibilant. The stops are *labial*, *alveolar*, or *velar*. Because the tongue tip and/or blade is used in the articulation of all four English sibilants, gestural overlap should be constrained by whether the tongue is also involved in the articulation of the stop. The greatest overlap and thus lowest ST ratio is expected with labials as lip movement does not constrain the movement of the tongue. Alveolars also require usage of the tongue tip, so the least overlap should be expected with alveolars since the same part of the tongue must be used. Velars should be intermediate as the movement of the tongue body still constrains that of the tongue tip.

The following three factors are not expected to *directly* gestural overlap, but since gender is known to affect properties of sibilants, age can affect speech rate, and audience can alter one's speech patterns, these factors will be considered as potential influencers on gestural overlap.

Speaker Gender. The center of gravity of sibilant noise is known to be higher in females than in males (Flipsen et al. 1999; Jongman et al. 2000). Because of gender differences in the spectral energy of sibilants, it is possible that sibilant duration may also show gender differences. The Buckeye Corpus codes speakers as *female* or *male*.

Speaker Age. Younger speakers tend to speak faster than elderly speakers (Linville 2001). It is possible that this may affect the pattern of gestural overlap and thus lead to differences in sibilant duration. The Buckeye Corpus codes speakers as *old* if they are over 40 and *young* if they are under 40.

Interviewer Gender. Speakers are known to linguistically accommodate to their audience (Giles 1973; Giles et al. 1973; Trudgill 1981; Pardo 2006; Gallois and Giles 2015). If there are differences in sibilant duration between female and male speakers, it is also possible that accommodation can lead to differences. The Buckeye Corpus codes interviewer gender as *female* or *male*.

One potential confound comes from the assumption of ST ratio as a proxy for gestural overlap. In taking the ratio, the confound of speech rate is removed. However, there is the question of whether the *overall duration of the cluster* in different conditions is correlated with the ST Ratio measure. If they are not, then the cluster duration should also be controlled for. To account for speech rate, the cluster duration was first normalized by dividing by the local speech rate. The local speech rate was calculated as the number of syllables occurring over the minute surrounding the cluster (ignoring silences). A Pearson's product-moment correlation test was then carried out to preliminarily assess whether the logged normalized duration of the ST/TS clusters was related to the ST Ratio. The results revealed a significant correlation (r = .406, p < .001). Since the ST Ratio and normalized duration of the cluster were highly correlated, the normalized duration was not included in the analysis.

3.2 Assessing Differences in Gestural Variability

Meaningful differences in variance are difficult to explore with corpus data, particularly because unlike in experimental data, factors are not controlled for. Whereas linear mixed effects models can control for both fixed and random factors to determine group differences, they do not directly answer the question of whether group variances are significantly different from one another.

Differences in variability are much less commonly explored and tools for analyzing variance differences are more limited (Kuppens and Yzerbyt 2014). Kuppens and Yzerbyt provide a method to model variability of data as a function of predictor variables using multilevel modeling. Typically, diagnostics such as the Levene's Test for Equality of Variances are used to determine whether two distributions are equal in variance and thus to test the homoscedascity assumption. Heterogeneity is thus treated as a problem for statistical inference procedures, rather than an object of interest. Kuppens and Yzerbyt's procedure involves comparing a model of hetereogeneous variance that takes into account covariances between within-subject variables to a model assuming homogeneous variance. A likelihood ratio or deviance test based on the maximum likelihood deviance is then run on the two models. If a chi-squared test shows that the deviance score for the hetereogeneous model is significantly lower than for the homogeneous model, it can be inferred that the variances are significantly different. All the predictor variables in this study are nested within subjects, so multilevel modeling would be a suitable method to test whether the variances in different conditions are significantly different, while controlling for other variables. Given limitations, however, this analysis will be pursued in future work.

The current study employs a more rudimentary method of assessing differences in variance. In order to test the question of whether variances significantly differ while controlling for other factors, smaller datasets were constructed from the larger corpus by taking tokens minimally differing in condition. Only combinations of factors that had a sizeable number of tokens were considered due to the dependence of variance measures on sample size. Including Place of Articulation, however, leads to miniscule counts for each speaker, so counts were collapsed across Place of Articulation. Because voiced tokens only make up 6.68% of the data and are thus heavily underpresented as compared to voiceless ones, only voiceless tokens were considered. Only four conditions had a sizeable number of speakers with enough tokens: tautomorphemic STV clusters in onset position, tautomorphemic STV clusters in medial position, tautomorphemic VTS clusters in medial position, and heteromorphemic VTS clusters in final position. Thus, only two hypotheses could be tested:

1. Onset tautomorphemic STV clusters show less variance of ST Ratio than medial ones

2. Medial tautomorphemic STV clusters show less variance of ST Ratio than VTS ones

While the counts are also robust for heteromorphemic final VTS clusters, there is no comparison that can be made to answer any of our hypotheses: onset and medial VTS clusters are both tautomorphemic so hypotheses on word position cannot be tested and the counts for final tautomorphemic VTS clusters and final heteromorphemic STV clusters are too thin, so hypotheses on morphemic status or cluster type cannot be tested. Thus, this condition was not considered. Counts for each speaker by condition may be found in Appendix B. The analyses were limited to speakers who had at least 30 tokens[6] for the three remaining conditions, leaving 23 speakers.

To answer the question of whether differences in group variances are significant, two analyses were run. The first analysis compares variances of the ST ratio with the Brown-Forsythe version of the Levene Test for equality of variances. This implementation of the Levene Test measures variance using the median, a technique which is more robust to deviations from normality than using the mean or using the F-ratio or Bartlett tests (Brown and Forsythe 1974). The second analysis considers speaker differences by calculating the variance of each speaker's distributions of ST ratio by each condition and running a linear model to test whether the variances of the conditions are significantly different. Two measures of variance were compared: standard deviation (SD) and median absolute deviation (MAD). The former, which calculates variance based off the mean of a distribution, is used in Cho's 2001 study on gestural variability in Korean and the latter, which calculates variance based off the median of a distribution, is used in Yanagawa's 2003 study on gestural variability in Hebrew. The variance measures for each speaker by condition may be found in Appendix C.

The counts for onset and medial STV clusters, split by place of articulation, are provided in Table 2 and the counts for medial STV and VTS clusters, split by place of articulation, are provided in Table 3.

Word Position	Alveolar	Velar	Labial	Total
Onset	1118	532	268	1918
Medial	799	171	426	1396
Total	1917	703	694	3314

Table 2: Counts for Onset vs. Medial STV clusters

Word Position	Alveolar	Velar	Labial	Total
STV	799	171	426	1396
VTS	87	831	103	1021
Total	886	1002	529	2417

Table 3: Counts for STV vs. VTS medial clusters

[6] This number is rather arbitrary: reducing the minimum to 25 yields 30 speakers and increasing it to 35 yields 20 speakers. Either way, power is lost either by reducing the number of tokens, leading to less representative data per speaker, or by reducing the number of speakers, leading to less robust generalizations across speakers.

Whether the variances between the conditions significantly differ was tested with a linear model in R. Because each speaker only contributed two data points to each hypothesis (one for each condition), a random effects structure was not added so as to prevent overfitting. Following previous literature on gestural overlap by word position, medial clusters should show greater variance of ST ratio than onset ones and heteromorphemic clusters should show greater variance of ST ratio than onset ones.

The hypotheses to be tested are laid out in (4).

Variable	Magnitude	Variability	Note
Word Position	medial & final < onset	medial & final > onset	
Morpheme	hetero- ≠ tauto-	hetero- > tauto-	may depend on voicing
Cluster Order	VTS > STV	VTS > STV	may depend on word position
Voicing	voiceless > voiced	?	
POA	alveolar > velar > labial	?	

Table 4: Hypotheses about ST Ratio

Magnitude and variability of the ST ratio should both lead to confusion in segmental ordering. Longer duration can lead to difficulty in parsing segmental order due to auditory decoupling. Greater variability of sibilant noise duration leads to a wider distribution in the representation of a cluster and thus ambiguity in the exact ordering of the individual segments. The confusion in these highly noisy, highly variable environments may trigger metathesis by leading a listener to conclude the reversed linear representation of a consonant sequence as the underlying form. For example, if a listener hears the word /wasp/ as ambiguously [wasp] ~ [waps] and decides due to ambiguity from various factors that the more likely representation is /waps/, metathesis will have occurred.

3.3 Model Selection

A linear mixed effects analysis was performed using the `lme4` (Bates et al. 2015) and `lmerTest` (Kuznetsova et al. 2017) packages in R (R Core Team 2017) with the ratio of the logged sibilant duration to the logged stop duration (the proxy for sibilant duration and called STRatio here) as the response variable.

Model comparison was carried out by first fitting a maximal model with the aforementioned independent factors and relevant interactions as fixed effects and random intercepts for each speaker with by-speaker random slopes for those factors and interactions (except for Speaker Gender, Speaker Age, and Interviewer Gender), as speakers may differ in the extent to which these factors affect the ST ratio. Models with morpheme coded binarily were compared to ones with morpheme coded ternarily. For the expected interactions, a model with a two-way interaction between Order x Word Position and Morpheme x Voicing was compared to a model with with no interactions and models with each of the interactions removed.

A factor was kept in the model if inclusion lowered the Akaike information criterion (AIC) and if the Likelihood Ratio Test (LRT) comparing the model with the factor to one without it yielded a p-value below 0.05. All the aforementioned fixed effects were retained in the final model, with morpheme coded ternarily. A model with two separate interactions for Order x Word Position and Morpheme x Voicing was determined to be better than a model missing one or both interactions. As most of the models also did not converge with the maximal random effects structure, random effects were removed stepwise in increasing order of the amount of variance they explained. With interaction terms included in the random slopes, only Order x Word Position by Speaker (but with no simple factors) allowed the model to converge. Without interaction terms, the best model had Morpheme, Order, and Place of Articulation as random slopes by speaker. The model without interactions in the random effects yielded a better model than the one with only Order x Word Position by Speaker. Speaker Gender, Speaker Age, and Interviewer Gender were insignificant factors. The model is captured by the following equation:

$$\begin{gathered}\text{STRatio} \sim \text{Order x WordPosition} + \text{Morpheme x Voicing} + \text{POA} + \\ (1 + \text{Order} + \text{Morpheme} + \text{POA} \mid \text{Speaker})\end{gathered}$$

Because the residuals for this model were right-skewed, data points with residuals over 2.5 standard deviations from the mean were removed. These points comprised 268 tokens, 2.32% of the data, leaving 11,268 data points. The counts for each of the variables following outlier removal is provided in Table 5.

Order x Word Position			
	Onset	Medial	Final
STV	2834	2004	602
VTS	36	1597	4195

Morpheme x Voicing			
	None	-s	-ed
Voiced	306	407	38
Voiceless	6966	3443	108

Place of Articulation	
Labial	1327
Alveolar	6727
Velar	3214

Table 5: Dependent variable counts for ST Ratio Analysis

The reference levels for each factor are as follows: Order = *STV*, Word Position = *Onset*, Morpheme = *None*, Voicing = *Voiced*, Place Of Articulation = *Alveolar*. The post-hoc significance between groups was determined with the `emmeans` packages (Lenth 2018), which employs Tukey's test to correct for family-wise error rate. Estimates for each of the effects in the fitted model were gathered and visualized with the aid of the `effects` package (Fox and Hong 2009).

4 Results

4.1 Magnitude of Sibilant-Stop Ratio

p-values for each of the fixed effects are provided in Table 6. All main effects except Voicing are significant and both interactions are significant. The estimates of the model and their *p*-values are provided in Table 7.

	Sum Sq.	Mean Sq.	NumDf	DenDF	F.value	Pr(> $\|F\|$)
Order	4.784	4.784	1	117.6	12.803	0.001
WordPosition	6.936	3.468	2	10,812.8	9.281	0.000
Morpheme	6.963	3.482	2	209.6	9.317	0.000
Voicing	0.039	0.039	1	4,770.1	0.104	0.747
POA	15.713	7.857	2	41.0	21.025	0.000
Order:WordPos	11.514	5.757	2	10,780.5	15.406	0.000
Morpheme:Voicing	5.439	2.719	2	6,499.8	7.278	0.001

Table 6: *p*-values of fixed effects

	Est.	Std. Error	df	*t*-value	Pr(> $\|t\|$)
(Intercept)	1.347	0.054	158.8	24.873	0.000
Order$_{\text{VTS}}$	-0.162	0.109	1,905	-1.495	0.135
WordPosition$_{\text{medial}}$	-0.128	0.019	10,950	-6.799	0.000
WordPosition$_{\text{end}}$	-0.166	0.031	11,140	-5.258	0.000
Morpheme$_{\text{s}}$	0.015	0.065	524.6	-0.231	0.818
Morpheme$_{\text{ed}}$	0.524	0.112	1,417	4.681	0.000
Voicing$_{\text{voiceless}}$	0.184	0.037	10,480	4.931	0.000
POA$_{\text{vel}}$	-0.163	0.030	38.500	-5.442	0.000
POA$_{\text{lab}}$	-0.187	0.033	40.98	-5.678	0.000
Order$_{\text{VTS}}$:WordPosition$_{\text{medial}}$	0.567	0.106	10,520	5.344	0.000
Order$_{\text{VTS}}$:WordPosition$_{\text{end}}$	0.471	0.113	10,610	4.164	0.000
Morpheme$_{\text{s}}$:Voicing$_{\text{voiceless}}$	-0.045	0.050	10,920	-0.898	0.369
Morpheme$_{\text{ed}}$:Voicing$_{\text{voiceless}}$	-0.467	0.122	4,824	-3.815	0.000

Table 7: Estimated coefficients of fixed effects on ST Ratio

The only factor not involved in an interaction is Place of Articulation. The results, visualized in Figure 5, show that the ST ratio in clusters with velar or labial stops is significantly less than in those with alveolar stops. While labial clusters do have a lower ST ratio than velar clusters, the difference is not significant (p = .468).

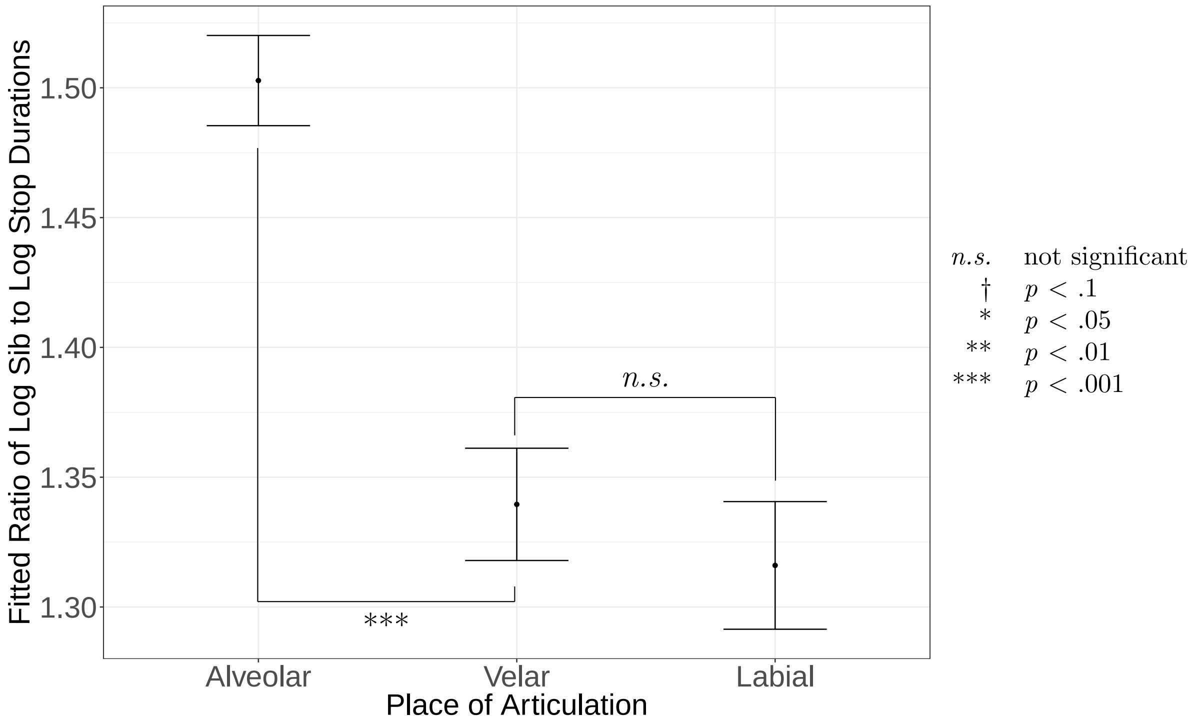

Figure 5: Model Estimates for Effect of Place of Articulation on ST Ratio

Figure 6 shows the estimated effects of the interaction between WORD POSITION and CLUSTER TYPE on the ST ratio. For STV clusters, the ST ratio is significantly higher in onset clusters than in medial and final ones, but there is no significant difference between medial and final clusters. VTS clusters share the same pattern with STV ones in medial and final position: non-initial clusters have significantly lower ST ratios than medial ones. However, initial VTS clusters have lower ST ratios than both their STV counterparts as well as medial and final VTS clusters.

The results for onset VTS sibilants, however, must be taken with caution. First, the difference between onset VTS and STV clusters is not significant (p = .279). Second, English does not have phonemic initial TS clusters—these clusters are almost all the result of unstressed vowel deletion in the initial syllable (ex. *decided* [dsaɪɾɪd]). Finally, there are only 36 such clusters (0.32% of the whole dataset).

Putting aside the onset VTS sibilant results, it can be seen that there is a general downward trend of sibilant noise as the position is later in the word. STV and VTS clusters pattern slightly differently: final STV clusters do not have significantly shorter ST ratios than medial ones (p = .464), but final VTS sibilants do (p = .002).

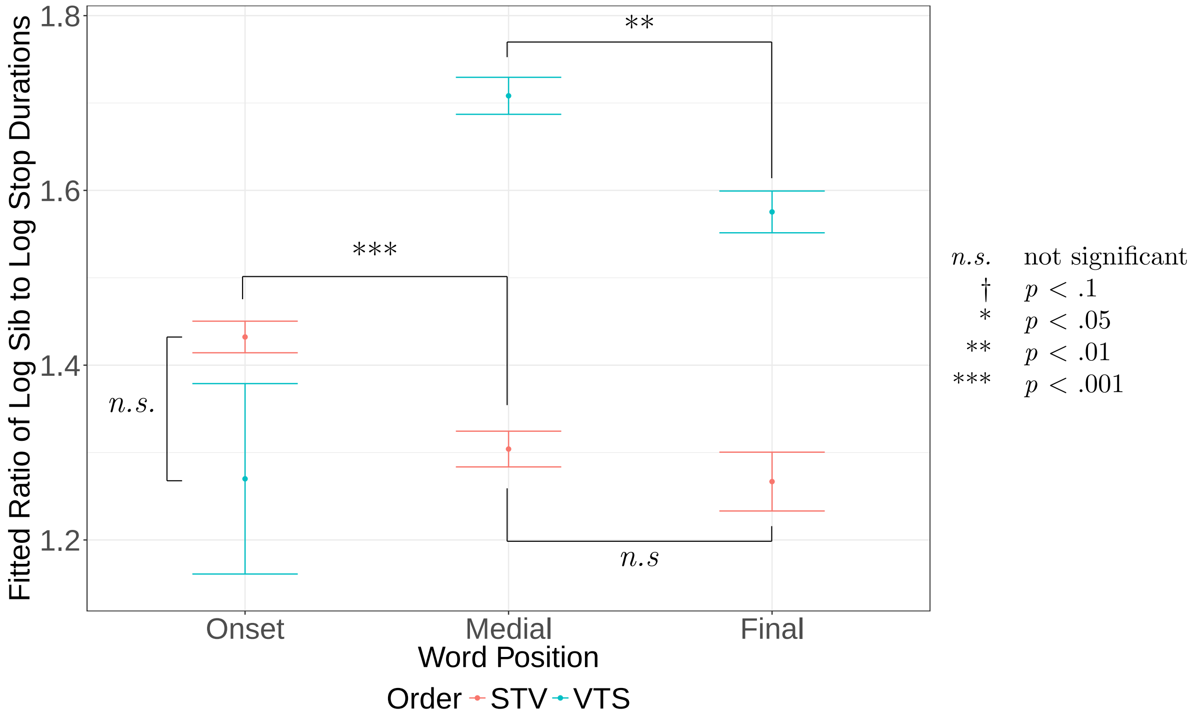

Figure 6: Model Estimates for Effect of Word Position x Cluster Order on ST Ratio

Figure 7 shows the estimated effects of the interaction between Morpheme and Voicing on sibilant duration. The difference in the ST ratio measure between clusters with voiced sibilants and clusters with voiceless sibilants is significant in tautomorphemic clusters (p < .001) and for clusters with *-s* (p < .001) and is slightly significant ($p = .068$) for clusters with *-ed*. For tautomorphemic clusters and clusters with *-s*, the ratio for voiceless sibilants is longer than voiced ones, but for the latter, the relationship is reversed. The ST ratio does not significantly differ by morpheme identity for voiceless sibilants, but the ratio for clusters with voiced sibilants preceding the *-ed* morpheme are significantly longer than that in tautomorphemic clusters or in clusters with morphemic *-s*. The results for clusters with voiced sibilants and with *-ed* must be taken with caution, however, as there are only 751 voiced tokens (6.66% of the data) and only 146 tokens with *-ed* (1.3% of the data).

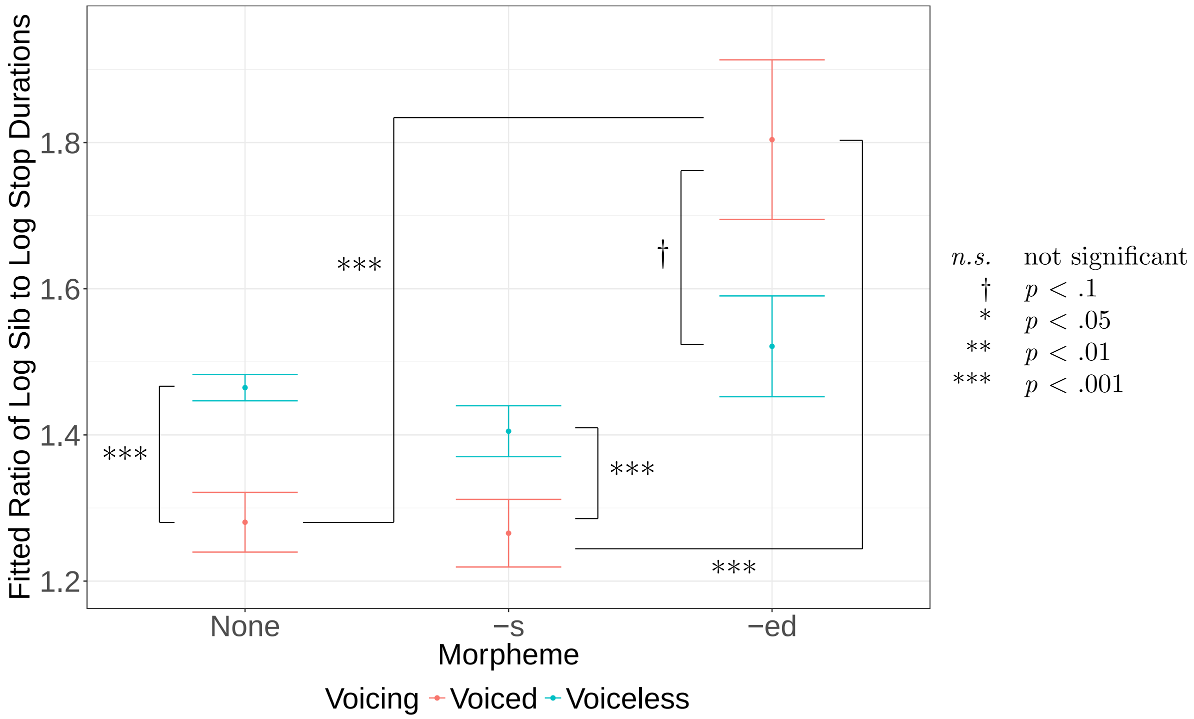

Figure 7: Model Estimates for Effect of Morpheme x Voice on ST Ratio

4.2 Variance of Sibilant-Stop Ratio

To test the hypothesis that there is greater variance in medial than onset position, STV clusters in onset position were compared to those in medial position. The SD of the ST ratio of onset STV clusters is 0.723 and that of medial ones is 0.699. The MAD of the ST ratio of onset STV clusters is 0.552 and that of medial ones is 0.515. Brown-Forsythe's Levene Test reveals that the variances are slightly significantly different: F(1,3312) = 3.327, p = .068. Thus medial clusters are *less* variable than onset position. A density graph comparing the distribution of ST ratios in onset versus medial STV clusters can be seen in Figure 8. The distributions have long right tails, with the max ST ratio being 12.404, but for visualization purposes, ratios above 5 are cut off.

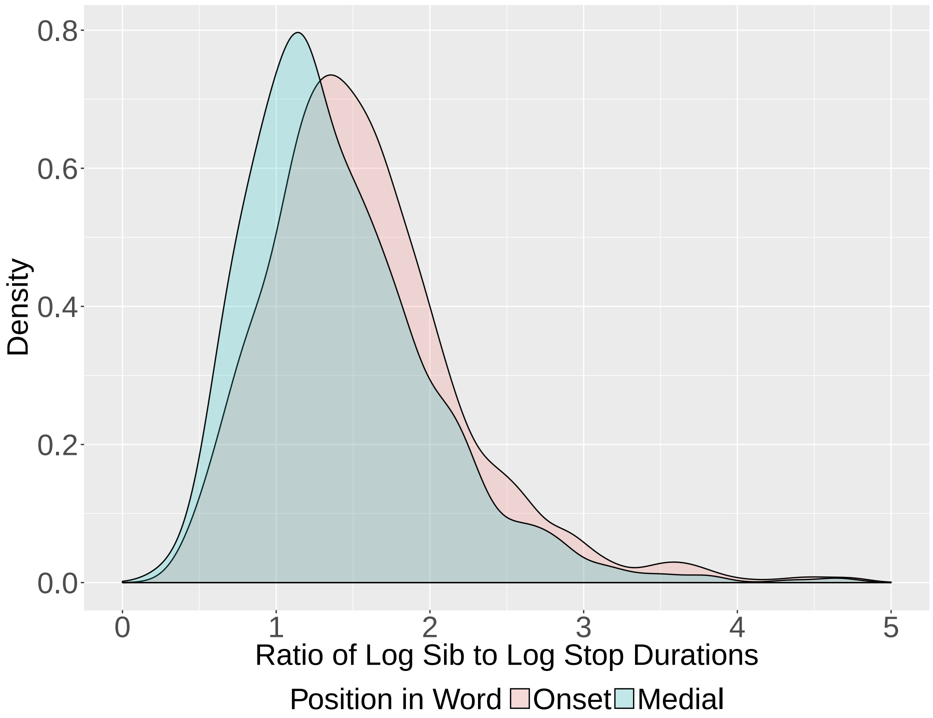

Figure 8: Density of ST Ratio in STV Clusters by Word Position

It is notable in the counts in Table 2 that there is an asymmetry in velars and labials between onset and medial position. While alveolars represent 58.29% of onset STV clusters and 57.23% of medial STV clusters, labials represent only 13.97% of onset STV clusters and velars represent only 12.25% of medial clusters. As sibilant duration differences by place of articulation were previously shown, another analysis was run with only alveolars (which had robust counts for both onset and medial position) to control for place of articulation. The SD of the ST ratio of onset STV clusters with alveolar stops is 0.81 and that of medial ones is also 0.81. The MAD of the ST ratio of onset STV clusters with alveolar stops is 0.617 and that of medial ones is 0.582. Brown-Forsythe's Levene Test reveals that the variances are not significantly different: F(1,1915) = 0.798, p = .372. A density graph comparing the distribution of ST ratios in onset versus medial STV clusters with alveolar stops can be seen in Figure 9. Once again, ratios above 5 are cut off.

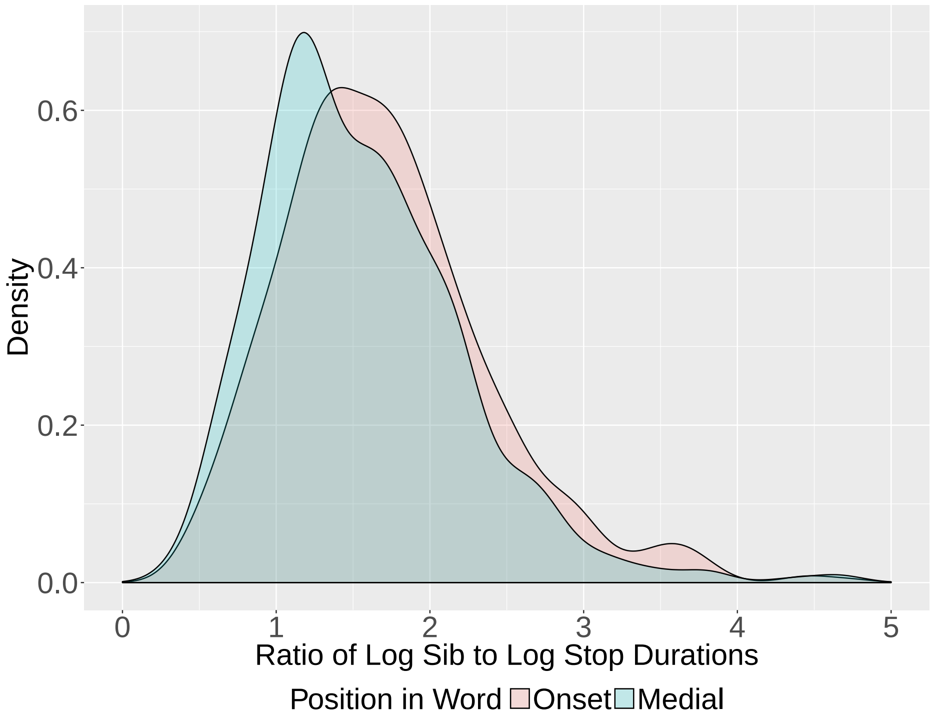

Figure 9: Density of ST Ratio in STV Alveolar Clusters by Word Position

To test the second hypothesis, STV and VTS clusters in medial position were compared. The SD of the ST ratio of medial STV clusters is 0.663 and that of medial ones is 0.872. The MAD of the ST ratio of onset STV clusters is 0.523 and that of medial ones is 0.754. Brown-Forsythe's Levene Test reveals that the variances are significantly different: F(1,1670) = 43.643, p < .001. VTS clusters are significantly more variable than STV ones in medial position. A density graph comparing the distribution of ST ratios in STV versus VTS medial clusters can be seen in Figure 10. Once again, the distributions have long right tails, with the max ST ratio being 9.947, but for visualization purposes, ratios above 5 are cut off.

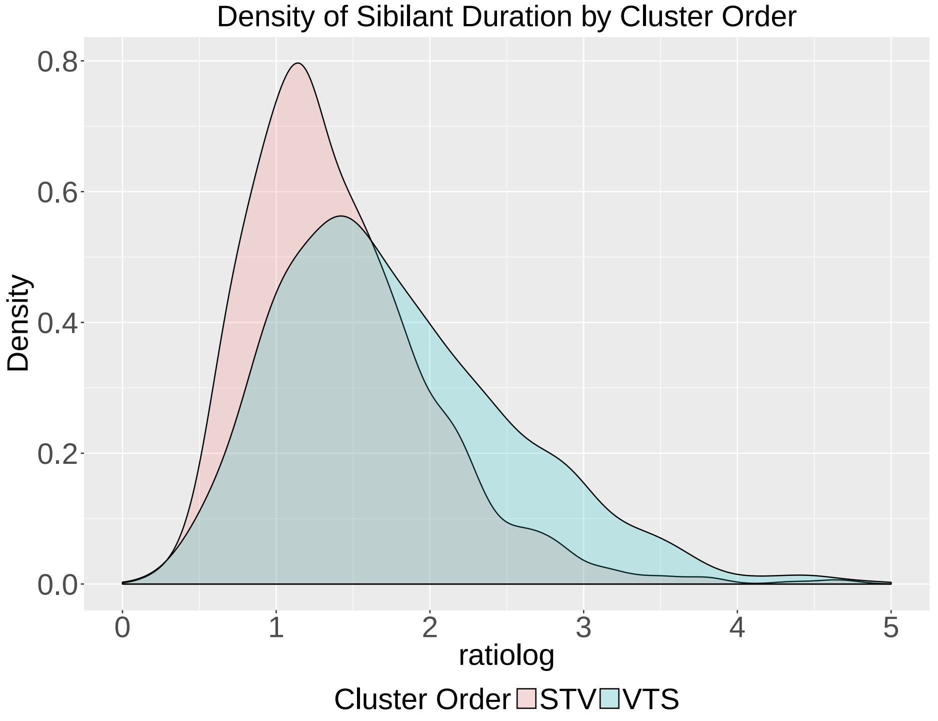

Figure 10: Density of ST Ratio in Medial Clusters by Cluster Order

Once again, the results should be treated with caution as place of articulation counts are very asymmetrical. Velars are heavily underrepresented in medial STV clusters, comprising only 12.25% of them, but are heavily overrepresented in medial VTS clusters, comprising 81.39% of them. On the other hand, alveolars and labials are heavily underpresented in medial VTS clusters, comprising only 8.52% and 10.09%, respectively. However, because of the asymmetries by place of articulation between STV and VTS clusters, the analysis cannot be limited to one place of articulation.

In the second analysis testing differences in gestural variability by word position, the linear model predicting SD by Word Position reveals that medial STV clusters are *less* variable than onset ones. Examination of residuals reveals that there are two outlier data points (over 2.5 standard deviations from the mean) from speakers 23 and 34, who had abnormally large standard deviations for their medial clusters. Prior to removal, the model did not predict medial clusters to be significantly less variable than onset ones ($p = .477$), but the difference becomes significant following removal of these two speakers ($p = .016$). On the other hand, the model predicting MAD by Word Position reveals the same directionality, but without significance ($p = .353$) and with no outliers. The estimates for the word position hypothesis for each model may be found in Tables 8 and 9.

	Estimate	SE	t	$\Pr(>\|t\|)$
Intercept	0.675	0.035	19.303	0.000
WordPosition-Medial	-0.131	0.049	-2.651	0.012

Table 8: Estimated coefficients of effect of Word Position on SD of ST Ratio

	Estimate	SE	t	$\Pr(>\|t\|)$
Intercept	0.501	0.021	23.754	0.000
WordPosition-Medial	-0.028	0.030	-0.938	0.353

Table 9: Estimated coefficients of effect of Word Position on MAD of ST Ratio

If the hypothesis that greater variance of gestural overlap is a factor in metathesis holds up, then greater variance of sibilant duration may be found in medial VTS clusters, which are more likely to metathesize into STV clusters than vice-versa. The linear model predicting SD by Cluster Order reveals that medial VTS clusters are more variable than onset ones. The two outlier data points from the word position analysis (speakers 23 and 34) are outliers here as well. Prior to removal, the greater variance of VTS clusters is only slightly significant (p = .085) but becomes much more significant after the removal of the two outliers (p < .001). The model predicting MAD by Cluster Order reveals the same directionality and also significance (p = .002) and there are no outliers. The estimates for the cluster order hypothesis for each model may be found in Tables 10 and 11.

	Estimate	SE	t	$\Pr(>\|t\|)$
Intercept	0.544	0.033	16.673	0.000
Order-VTS	0.205	0.046	4.455	0.016

Table 10: Estimated coefficients of effect of Cluster Order on SD of ST Ratio

	Estimate	SE	t	$\Pr(>\|t\|)$
Intercept	0.473	0.031	15.484	0.000
Order-VTS	0.139	0.043	3.211	0.002

Table 11: Estimated coefficients of effect of Cluster Order on MADs of ST Ratio

The analyses carried out to test variance differences reveal that medial clusters are *not* more variable than onset ones. In fact, they are *less* variable, although this difference does not seem to be very significant, if at all. The Brown-Forsythe implementation of the Levene's Test and a linear model with an SD measure for each speaker reveal significance at the p < .1 level, but looking only at alveolars or at the MAD measure (which is more robust to outliers) reveals no significance. VTS clusters, however, are significantly more variable than STV ones in all measures. Unfortunately, because of insufficient tokens, a hypothesis testing whether variance was greater across heteromorphemic clusters than across tautomorphemic ones could not be tested.

5 Discussion

This study explored the effects of word position, morphemic status, cluster order, voicing, and place of articulation on the duration of sibilant noise as normalized to its neighboring stop in a cluster in order to test whether this acoustic measure would corroborate previous findings on the degree and variability of gestural overlap of different clusters.

Word position was shown to significantly affect the magnitude of the ST ratio in a cluster and to also interact with the cluster order. Onset VTS clusters deviate greatly from others in having unusually short ST ratios, but notably, these clusters make up less than 1% of the data and are all the result of phonetic reduction through deletion of a vowel. If these aberrant clusters are ignored, two generalizations emerge. The first is that the ST ratio in medial and final STV clusters is significantly shorter than that in initial position. These findings accord with Byrd (1996), Chitoran et al. (2002), and Yanagawa (2003), who show that there is greater gestural overlap in non-initial position, which would lead to decreased sibilant noise.

Of note is the difference between medial and final position depending on cluster order. Final VTS clusters have significantly shorter ST ratios than medial ones but final STV clusters do *not* have significantly shorter ST ratios than medial ones. If metathesis were to occur in final VTS clusters, the stop would end up wedged between the sibilant and silence (unless the following word began with a vowel), rendering it perceptually weakened. Because sibilant noise can lead to auditory decoupling and, consequently, confusion in segmental order, the large decrease in the relative duration of sibilant noise in final VTS clusters as opposed to medial ones is potentially suggestive of perceptual optimization. In medial position, the longer sibilant duration may trigger segmental confusion and thus metathesis, which would lead to increased perceptibility of the stop by placing it in prevocalic position. In final position, however, metathesis would lead to *decreased* perceptibility, and so attenuating sibilant noise may be a strategy employed to maximize perception of the stop.

The analysis also revealed a significant effect of cluster order. Once again setting aside onset VTS clusters, which have shorter ST ratios than their STV counterparts (although not significantly), VTS clusters have significantly longer ST ratios than STV ones. As mentioned earlier, this does not necessarily mean that there is less gestural overlap in VTS clusters than in STV ones. Rather, it may be the case that sibilants following stops are less restricted than ones preceding stops, particularly as sibilants in final VTS clusters may theoretically be extended until one runs out of breath. Notably, out of the 274 outliers (all of which are in the right-tail), 165 (60.22%) are from final VTS clusters[7]; 68 of these are utterance-final[8] sibilants. This suggests that post-stop sibilants can theoretically be greatly lengthened, leading VTS sibilants to be naturally produced longer than STV ones.[9] While postvocalic vs. prevocalic position of a stop is one explanation for the asymmetry in intervocalic TS > ST metathesis, the data here also support a view that the greater sibilant noise in VTS

[7] The next biggest group is medial VTS clusters, of which there are 39 (14.23%)

[8] Defined as a cluster after which the following tagged segment is silence.

[9] Note, however, that despite these outliers, final VTS clusters still have a significantly lower ST ratio than medial ones.

vs. STV sequences can strengthen the phenomenon of auditory decoupling, increasing the likelihood for metathesis.

Variance of the ST ratio was also explored with respect to word position and cluster order. Due to low token counts and the limited nature of the datasets for the variance analyses, the results must be taken with caution, but preliminarily, there is little evidence that medial STV clusters are more variable than onset ones; in fact, there is some evidence that they may be *less* variable, a finding that runs counter to Chitoran et al. (2002) and Yanagawa (2003), who show that medial clusters show greater variability in gestural timing than onset ones. Crucially, however, Chitoran et al. carry out their study on stop+stop sequences in Georgian and Yanagawa's study focuses on stop+non-fricative sequences in Hebrew. It is possible that there are conflicting findings because of the difference in the targeted phonemes, particularly for Georgian, where not even fricatives are looked at. The difference in phonotactics may also be a factor: while English allows for ST and TS clusters in medial position, only ST clusters are allowed in onset position, but Hebrew allows both ST and TS clusters in both onset and medial positions. It is possible that the restriction in the onset to ST ordering may lead onset ST clusters to have a more variable distribution than would be expected if there was competition with the opposite TS order (as for Hebrew onset ST clusters). If this is the case, even if clusters are more variable in general than onset ones, this effect may be cancelled out by the increased variability of onset ST clusters in English. Such phonotactic differences by word position may lead to language-specific behavior in the variance of clusters.

It was hypothesized that TS clusters would be more variable in the ST ratio measure than ST clusters in medial position as they are more susceptible to metathesis. This was shown to be the case, supporting Yanagawa's claim that variability can be a trigger for metathesis. A highly variable distribution of timing in gestural overlap can lead to ambiguity in the interpretation of segmental order, particularly if it overlaps with the distribution of the cluster in the opposite order. In particular, because sibilant noise masks the properties of an adjacent stop and also causes confusion of segmental order, high variability involving a sibilant-stop cluster can lead to greater uncertainty about temporal properties of the realization of the cluster.

As mentioned in the description of the methodology, clusters in which the stop was flanked by a sibilant and another consonant were removed from the analysis, due to an uncertain hypothesis about how they would pattern in comparison to STV and VTS clusters. Many of these clusters involve an adjacent sonorant, such as in ***str**ength* or *a**nts***, so it is possible that the sonorant provides an adequate transitional cue for the stop. Whether the existence of the sonorant then leads to similar behavior to STV and VTS clusters or not is a problem left to future work.

While voicing was shown not to be an independently significant factor, it does interact with morphemic status affect magnitude of the ST ratio, as predicted. Nonmorphemic clusters with voiced sibilants were shown to have a significantly shorter ST ratio than ones with voiced sibilants and the morpheme *-ed*. The relationship between voiced and voiceless sibilants is also reversed between these two groups, such tautomorphemic clusters with voiceless sibilants have higher ST ratios than ones with voiced sibilants but heteromorphemic clusters with *-ed* and voiceless sibilants have lower ST ratios than ones with voiced sibilants. Given the small token count for *-ed* clusters, however, these results must be taken with caution. The differences between morphemic and non-morphemic /s/ show both similarities and differences to Plag et al. (2017). Like Plag et al., voiceless morphemic /s/ was shown to be shorter than non-morphemic /s/, although not significantly so. Voiced morphemic /s/ was found to be slightly longer than non-morphemic /s/, but once again, this difference is not significant. Plag et al. find that voiced *plural -s* is longer than all other voiced /s/. While Plag et al. also use the Buckeye Corpus, there are various methodological differences from this study. Plag et al. use a measure of absolute duration of /s/ as well as a relative measure comparing the duration of /s/ to the entire word, whereas this study looks at the ratio of /s/ to an adjacent stop. Second, Plag et al. only compare *final* non-morphemic /s/ to morphemic /s/, whereas all nonmorphemic /s/'s, regardless of word position, are compared to morphemic ones in this study. All morphemic /s/'s were also collapsed in this study, which would particularly attenuate the effect for voiced /s/ as Plag et al. showed that morphemic /s/'s that were not the plural or plural genitive marker did not significantly differ in duration from non-morphemic /s/'s. The results in this study are not fully comparable to those of Seyfarth et al. (2018): they showed that *absolute duration* of both roots and affixes in homophones with a morpheme boundary were longer than ones without a boundary (regardless of voicing). However, if both the root and affix are lengthened, the ratio of the affix to the root should not necessarily be expected to increase, and as such, the ratio of the affix to the preceding stop may also not necessarily increase. The question of variance by morpheme status was also unable to be explored in this study, given heavily asymmetric counts between morphemic and non-morphemic clusters. In sum, the magnitude results did not reach significance (although the directions match those of Plag et al.), except for the poorly represented *-ed* tokens, and a variance analysis could not be carried out. A more controlled study observing differences between different morphemes should be carried out in future work.

Place of articulation was also observed to be a significant factor in magnitude of the ST ratio. Clusters with labial and velar stops showed significantly shorter ST ratios than those with alveolar stops. This finding was expected, since the articulation of a velar or labial stop requires little to no coarticulation with the front of the tongue, which is employed in all English sibilants. Thus, more gestural overlap may occur in the coarticulation of sibilants with an adjacent velar or labial stop, whereas in a cluster with an alveolar stop and a sibilant, the articulations may not overlap as much due to the usage of the same articulators. Clusters with labial stops also expected to exhibit greater overlap than those with velar stops since lip closure does not require any usage of the tongue while raising of the tongue body involves some displacement of the front of the tongue. While clusters with labial stops indeed showed the lowest ST ratio, the measure did not significantly differ from those with velar stops.

In showing greater variability in medial heteromorphemic clusters in Hebrew, Yanagawa (2003) makes an argument that metathesis of TS to ST in the *hitpa'el* binyan is due to the combination of word position and morpheme status. While greater variability for these factors was not found in this study, greater variability was indeed found for medial VTS clusters over STV clusters, a finding that falls in line with Yanagawa's claim as medial TS clusters flanked by vowels tend to metathesize to ST rather than vice versa. Higher ST ratios were also found for VTS clusters over STV ones, a factor that could lead to auditory metathesis as described by Blevins and Garrett (2004). The metathesis of VTSV > VSTV then may be driven not only by the bias to hear the stop in postvocalic position rather than prevocalic position, but *also* by segmental confusion via a long sibilant causing auditory decoupling in combination with ambiguity created by gestural variability. In word-final position, a speaker-driven process of perceptual optimization appears to counteract this potential confusion, which would lead to a perceptually less optimal result, through reduction of the sibilant noise.

Blevins and Garrett's 2004 typology of metathesis suggests that the four types of metathesis are distinct. However, the current study suggests that coarticulation also plays a role in what is labeled as auditory metathesis. The magnitude and variability of sibilant noise in an ST/TS cluster is affected by the degree of gestural overlap. If this leads to greater or more variable sibilant noise, it may trigger metathesis through perceptual confusion caused by auditory decoupling. In this way, coarticulatory metathesis is similar, since in stop-stop clusters, the degree of gestural overlap can lead to misperception of the initial stop as actually occurring second. Compensatory and perceptual metathesis also both involve coarticulation: the former involves extreme anticipatory coarticulation of an unstressed syllable and the latter long-distance phonetic features stretching into other segments. This study suggests that auditory metathesis also potentially has its roots in coarticulation. If auditory metathesis can be explained by coarticulation like the other subtypes, then it suggests that theories of metathesis as arising from a mismatch between speaker production and listener perception may be enhanced by exploring how, as in other sound changes (see vowel nasalization (Beddor 2009), for example), coarticulation leads to biases that can trigger the percept of metathesis.

6 Directions

This exploratory study exploited data from a large acoustic corpus in an attempt to reproduce findings from previous smaller-scale articulatory experiments. As naturalistic data leads to various confounds because of its unbalanced nature, several factors were controlled for in an effort to isolate contributions to magnitude and variability of sibilant noise. Greater magnitude and variability was predicted to occur in clusters that would be more likely to metathesize, a prediction that was borne out most robustly for medial clusters with postvocalic stops (VTS), an environment that is known to metathesize diachronically. These findings support Blevins and Garrett's claim that longer sibilant noise can lead to metathesis and Yanagawa's claim that variability of gestural timing can lead to metathesis. One direction for future study is to assess differences in variances with more sophisticated methods from multilevel modeling that can account for all the data without the need to subset the data into severely limited datasets. If greater magnitude and variability of sibilant noise increase the likelihood of metathesis, then metathesis should be expected to be more common in conditions that favor long and/or variable sibilants. Following from the findings of this study, asymmetries should also be expected in sibilant-stop metathesis by place of articulation. In particular, the finding that sibilants are longer when adjacent to alveolar stops than when adjacent to velar or labial stops suggests that *ts* > *st* could be a more likely metathesis than *ks* > *sk* or *ps* > *sp*. Such a hypothesis could be tested via observation of diachronic data.

The findings in this study also stand to be corroborated by data on articulatory timing. To answer this question, I am currently in the process of analyzing data from an articulatory X-ray microbeam database (Westbury 1994). The findings from this database will bear on the question of whether the simplifying assumptions of gestural overlap and sibilant noise in this study are valid. This study also provides a set of factors to be tested in experimental paradigms. Production experiments may be carried out to observe whether the findings on sibilant noise magnitude and variability are reproduceable when factors are controlled for carefully. A perception experiment culd also be carried out to test the theory that greater sibilant noise will increase the likelihood of segmental order confusion. Further work should also explore acoustic (and articulatory) corpora available in other languages in order to tease apart effects due to differences in language phonotactics.

This study contributes to the growing literature on phonetic corpus work and provides support for the use of acoustic data as a proxy for articulatory data given the convergence of most previous findings. It also provides a lens into production factors that can provide a catalyst for metathesis to occur. While perceptual explanations have been well-accepted for metathesis, the understanding of the contribution of production factors is still in its infancy.

References

Douglas Bates and Martin Mächler and Ben Bolker and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67:1–48.

Beddor, Patrice Speeter. 2009. A coarticulatory path to sound change. *Language* 85:785–821.

Blevins, Juliette and Garrett, Andrew. 1998. The Origins of Consonant-Vowel Metathesis. *Language* 74:508–556.

Blevins, Juliette and Garrett, Andrew. 2004. The evolution of metathesis. In *Phonetically based phonology*, ed. Bruce Hayes, Robert Kirchner, and Donca Steriade, 117–56. Cambridge University Press.

Bloomfield, Leonard. 1962. *The Menomini Language*. New Haven: Yale University Press.

Bregman, Albert S. 1990. *Auditory Scene Analysis*. Massachusetts: MIT Press.

Bregman, Albert S and Campbell, Jeffrey. 1971. Primary auditory stream segregation and perception of order in rapid sequences of tones. *Journal of experimental psychology* 89:244.

Browman, Catherine P. and Goldstein, Louis. 1988. Some notes on syllable structure in articulatory phonology. *Phonetica* 45:140–155.

Browman, Catherine P. and Goldstein, Louis. 1989. Articulatory gestures as phonological units. *Phonology* 6:201–251.

Morton B. Brown and Alan B. Forsythe. 1974. Robust tests for the equality of variances. *Journal of the American Statistical Association* 69:364–367. URL `http://www.jstor.org/stable/2285659`.

Byrd, Dani. 1995. C-centers revisited. *Phonetica* 52:285–306.

Byrd, Dani. 1996. Influences on articulatory timing in consonant sequences. *Journal of phonetics* 24:209–244.

Chitoran, Ioana and Goldstein, Louis and Byrd, Dani. 2002. Gestural overlap and recoverability: Articulatory evidence from Georgian. *Laboratory Phonology* 7:419–447.

Cho, Taehong. 2001. Effects of morpheme boundaries on intergestural timing: Evidence from Korean. *Phonetica* 58:129–62.

Connell, Bruce. 1994. The structure of labial-velar stops. *Journal of Phonetics* 22:441–476.

Crystal, David. 1997. *A Dictionary of Linguistics and Phonetics*. Oxford: Blackwell.

Flipsen, Peter and Shriberg, Lawrence and Weismer, Gary and Karlsson, Heather and McSweeny, Jane. 1999. Acoustic characteristics of/s/in adolescents. *Journal of Speech, Language, and Hearing Research* 42:663–677.

John Fox and Jangman Hong. 2009. Effect displays in R for multinomial and proportional-odds logit models: Extensions to the effects package. *Journal of Statistical Software* 32:1–24. URL `http://www.jstatsoft.org/v32/i01/`.

Fromkin, Victoria A. 1971. The Non-Anomalous Nature of Anomalous Utterances. *Language* 47:27–52.

Fujimura, Osamu and Macchi, M. J. and Street, L. A. 1978. Perception of Stop Consonants with Conflicting Transitional Cues: A Cross-Linguistic Study. *Language and Speech* 21:337–346.

Gallois, Cindy and Giles, Howard. 2015. Communication accommodation theory. *The international encyclopedia of language and social interaction* 1–18.

Giles, Howard. 1973. Accent mobility: A model and some data. *Anthropological linguistics* 87–105.

Giles, Howard and Taylor, Donald M and Bourhis, Richard. 1973. Towards a theory of interpersonal accommodation through language: Some canadian data. *Language in society* 2:177–192.

Graff, Peter and Scontras, Gregory. 2012. Metathesis as Asymmetric Perceptual Realignment. In *WCCFL XXVIII*.

Grammont, Maurice. 1923. L'interversion. In *ANTIΔΩPON: Festschrift Jacob Wackernagel zur Vollendung des 70. Lebensjahres am 11. Dezember 1923*, 72–77. Göttingen: Vandenhoeck & Ruprecht.

Grammont, Maurice. 1933. *Traité de phonétique*. Paris: Delagrave.

Harkema, Henk. 1999. Dutch schwa is not a long vowel. Master's thesis, UCLA.

Heid, Sebastian and Hawkins, Sarah. 2000. An Acoustical Study of Long Domain /r/ and /l/ Coarticulation. *Proceedings of the 5th Seminar on Speech Production: Models and Data* 77–80.

Hock, Hans Heinrich. 1985. Regular metathesis. *Linguistics* 23:529–546.

Hockett, Charles. 1958. *A Course in Modern Linguistics*. Macmillan.

Hoole, Philip and Pouplier, Marianne and Beňuš, Štefan and Bombien, Lasse. 2013. Articulatory coordination in obstruent-sonorant clusters and syllabic consonants: Data and modelling. *Proceedings of ratics3* 79–94.

Hume, Elizabeth. 1998. The role of perceptibility in Consonant/Consonant metathesis. In *WCCFL XVII Proceedings*, 293–307. Stanford: CSLI.

Hume, Elizabeth and Seo, Misun. 2004. Metathesis in Faroese and Lithuanian: from speech perception to Optimality Theory. *Nordic Journal of Linguistics* 27:35–60.

Hume, Elizabeth V. 2004. The indeterminacy/attestation model of metathesis. *Language* 80:203–237.

Jones, Kyle. 2016. The perception of stop/sibilant clusters in Modern Hebrew. In *Poster presented at LabPhon 15: Speech Dynamics and Phonological Representations*. Ithaca, NY: Cornell University.

Jongman, Allard and Wayland, Ratree and Wong, Serena. 2000. Acoustic characteristics of english fricatives. *The Journal of the Acoustical Society of America* 108:1252–1263.

Kelly, John and Local, John K. 1986. Long-domain resonance patterns in English. *International Conference on Speech Input/Output; Techniques and Applications* 304–309.

Kiparsky, Paul. 1982. Lexical morphology and phonology. In *Linguistics in the morning calm*, 3–91. Seoul: Hanshin Publishing Company.

Kökeritz, Helge. 1945. The Reduction of Initial kn and gn in English. *Language* 21:77–86.

Kuppens, Toon and Yzerbyt, Vincent Y. 2014. Predicting variability: Using multilevel modelling to assess differences in variance. *European Journal of Social Psychology* 44:691–700.

Alexandra Kuznetsova and Per B. Brockhoff and Rune H. B. Christensen. 2017. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software* 82:1–26.

Ladefoged, Peter. 2001. *Vowels and consonants: An introduction ot the sounds of languages*. Malden: Blackwell Publishing.

Lehmann, Winfred P. 1962. *Historical Linguistics: An Introduction*. New York: Holt, Rinehart, & Winston.

Lenth, Russell. 2018. *emmeans: Estimated marginal means, aka least-squares means*. URL `https://CRAN.R-project.org/package=emmeans`, r package version 1.1.3.

Levelt, Willem J. M. and Roelofs, Ardi and Meyer, Antje S. 1999. A theory of lexical access in speech production. *Behavioral and Brain Sciences* 22:1–75.

Linville, Sue Ellen. 2001. *Vocal aging*. Singular Thomson Learning.

Losiewicz, Beth L. 1995. Word frequency effects on the acoustic duration of morphemes. *The Journal of the Acoustical Society of America* 97:3243–3243.

Lunden, Anya and Renoll, Kelsey. 2015. Position and stress as factors in long-distance consonant metathesis. Working paper.

Makashay, Matthew. 2001. Lexical effects in the Perception of Obstruent Ordering. *Studies on the Interplay of Speech Perception and Phonology, OSUWPL* 55:88–116.

Marslen-Wilson, William D. 1987. Functional Parallelism in Spoken Word-Recognition. *Cognition* 25.

Mielke, Jeff. 2001. Turkish /h/ deletion: evidence for the interplay of speech perception and phonology. In *Proceedings of NELS 32*.

Mielke, Jeff and Hume, Elizabeth. 2001. Consequences of word recognition for metathesis. In *Surface Syllable Structure and Segment Sequencing*, ed. Elizabeth Hume, Norval Smith, and Jeroen van de Weijer, 135–158. Leiden: Holland Institute of Generative Linguistics.

Montreuil, Jean-Pierre. 1981. The Romansch 'Brat'. *Papers in Romance* 3:67–76.

Ohala, John J. 1981. The listener as a source of sound change. In *Papers from the Parasession on Language and Behavior*, ed. Carrie S. Masek, Robert A. Hendrick, and Mary Frances Miller, 178–203. Chicago: Chicago Linguistics Society.

Ohala, John J. 1993. The phonetics of sound change. In *Historical linguistics: Problems and perspectives*, ed. Charles Jones, 237–278. London: Longman.

Osthoff, Hermann and Brugmann, Karl. 1878. *Morphologische Untersuchungen auf dem Gebiete der indogermanischen Sprachen*. Leipzig: Hirzel.

Pardo, Jennifer S. 2006. On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America* 119:2382–2393.

Penny, Ralph J. 2002. *A history of the Spanish language*. Cambridge: Cambridge University Press, 2 edition.

Pitt, Mark and Dilley, Laura and Johnson, Keith and Kiesling, Scott and Raymond, William and Hume, Elizabeth and Fosler-Lussier, Eric. 2007. *Buckeye Corpus of Conversational Speech (2nd release) [www.buckeyecorpus.osu.edu]*. Columbus, OH: Department of Psychology, Ohio State University.

Pitt, Mark and McQueen, James. 1998. Is compensation for coarticulation mediated by the lexicon? *Journal of Memory and Language* 39:347–370.

Plag, Ingo and Homann, Julia and Kuntner, Gero. 2017. Homophony and morphology: The acoustics of word-final S in English. *Journal of Linguistics* 53:181–216.

Powell, J. V. 1985. An Occurrence of Metathesis in Chimakuan. *Oceanic Linguistics Special Publications, For Gordon H. Fairbanks* 20:105–110.

R Core Team. 2017. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

Recasens, Daniel. 1989. Long range coarticulation effect for tongue dorsum contact in VCVCV sequences. *Speech Communication* 8:293–307.

Recasens, Daniel and Fontdevila, Jordi and Pallarè, Maria Dolors. 1992. Alveolar-palatal correlations in coarticulatory activity for a selected group of Catalan consonants. *Bulletin de la Communication Parlée* 2.

Seyfarth, Scott and Garellek, Marc and Gillingham, Gwendolyn and Ackerman, Farrell and Malouf, Robert. 2018. Acoustic differences in morphologically-distinct homophones. *Language, Cognition and Neuroscience* 33:32–49. doi: 10.1080/23273798.2017.1359634.

Spencer, Andrew. 1996. *Phonology*. Oxford: Blackwell.

Steriade, Donca. 2001. Directional asymmetries in place assimilation: a perceptual account. In *The role of speech perception in phonology*, ed. Elizabeth Hume and Keith Johnson, 219–50. Brill Academic Pub.

Stoel-Gammon, Carol. 1985. Phonetic inventories, 15-24 months: A longitudinal study. *Journal of Speech and Hearing Research* 28:505–512.

Tapio, Sophia. 2008. The effects of frequency and composition on production duration on morphological processing. Master's thesis, MIT.

Trudgill, Peter. 1981. Linguistic accommodation: Sociolinguistic observations on a sociopsychological theory. *Papers from the Parasession on Language and Behavior* 218–237.

Walsh, Thomas and Parker, Frank. 1983. The duration of morphemic and non-morphemic [s] in English. *Journal of Phonetics* 11:201–206.

Westbury, John R. 1994. *X-ray microbeam speech production database user's handbook, version 1.0*. http://www.medsch.wisc.edu/ milenkvc/pdf/ubdbman.pdf, Madison, WI.

Wheeler, Max W. 2005. Cluster reduction: deletion or coalescence? *Catalan Journal of Linguistics* 4:57–82.

Wright, Richard. 2001. Perceptual cues in contrast maintenance. In *The Role of Perception in Phonology*, ed. Elizabeth Hume and Keith Johnson, 251–277. New York: Academic Press.

Yanagawa, Mariko. 2003. Metathesis in Modern Hebrew: An Analysis in Articulatory Phonology. In *Proceedings of the 15th International Congress of the Phonetic Sciences*, 1671–1674. Rundle Mall: Causal Productions.

Appendix A: Empirical Data

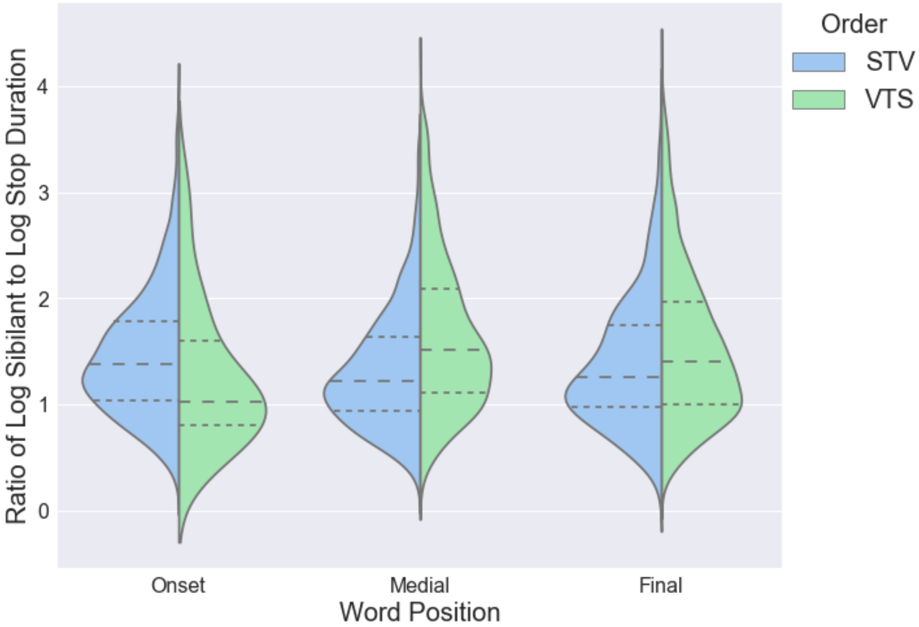

Figure 11: Interaction of Word Position x Cluster Order and ST Ratio

Interestingly, VTS sibilants in word-initial position are significantly shorter than

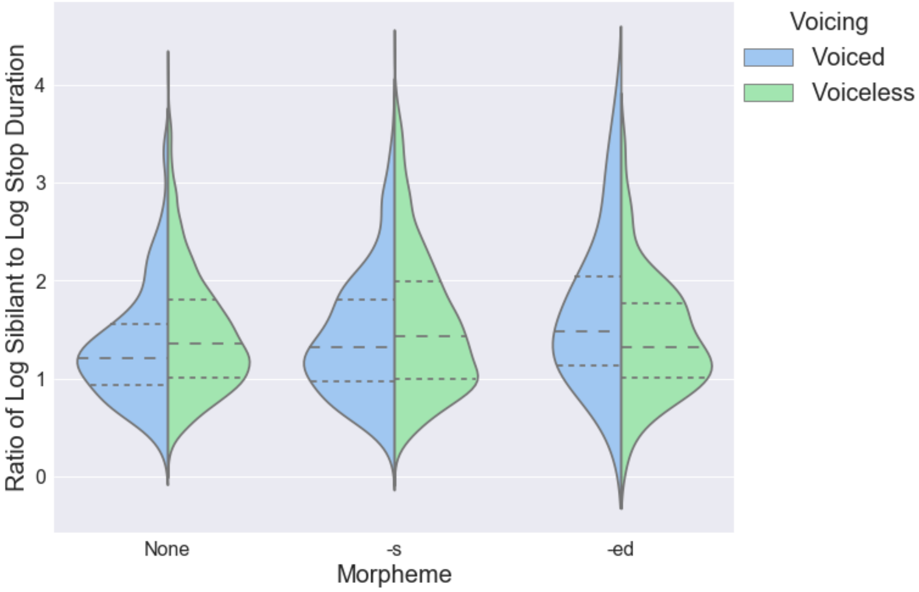

Figure 12: Interaction of Morpheme x Voice and ST Ratio

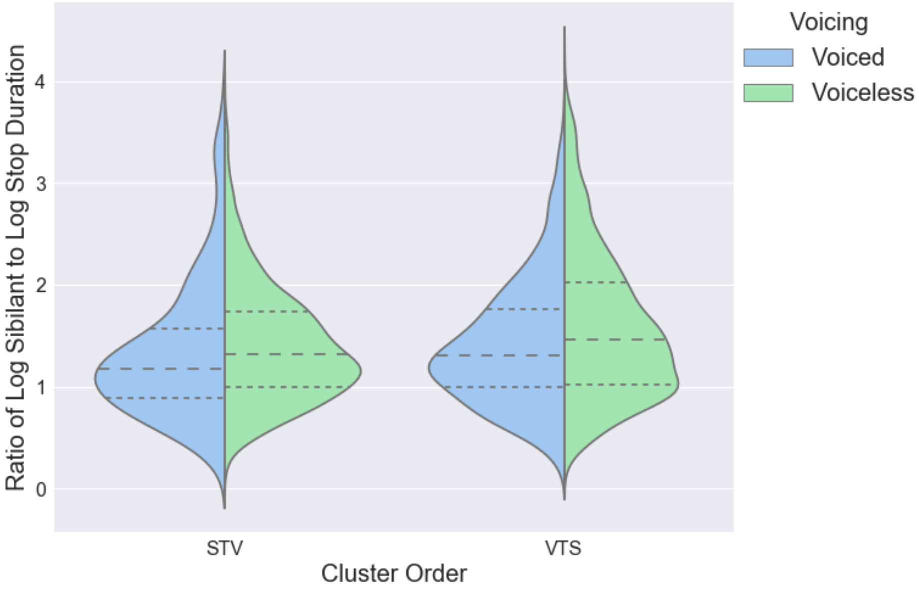

Figure 13: Interaction of Voice x Cluster Order and ST Ratio

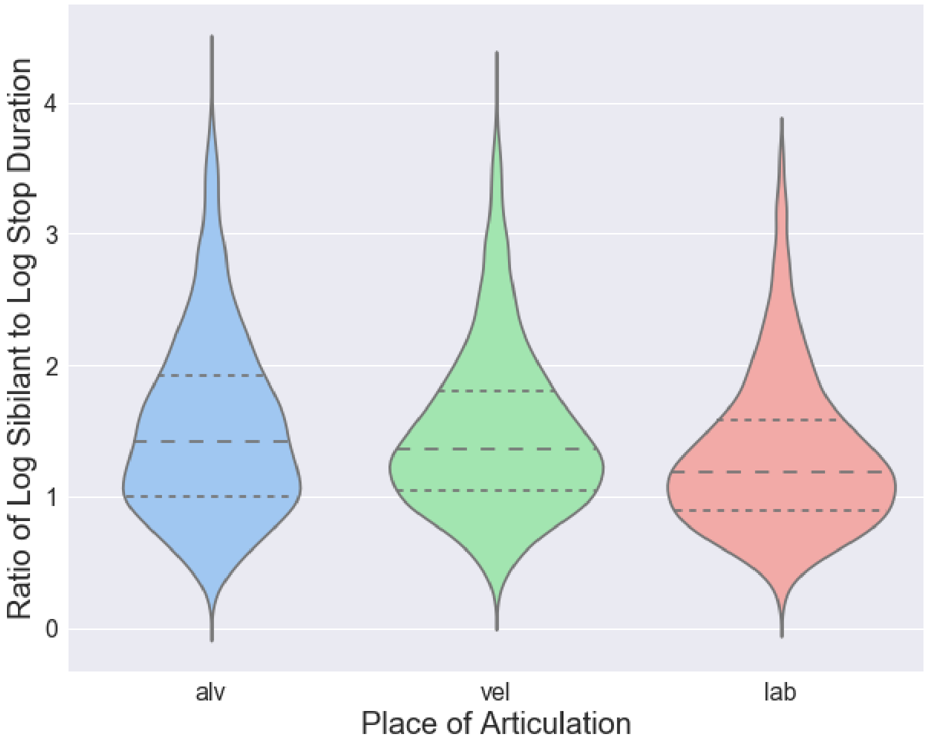

Figure 14: Place of Articulation and ST Ratio

Appendix B: Counts for Speakers by Condition

For the variance analyses, token counts for speakers by each combination of factors were determined. Due to low counts, only tokens with voiceless sibilants were considered and counts were collapsed across place of articulation. The gender of the speaker (f = female, m = male), age of the speaker (o = old, y = young), and gender of the interviewer (f = female, m = male), are provided in order in parentheses following the speaker number. The conditions are in the following order: 1) Word Position: O = onset, M = medial, F = final. 2) Morpheme Boundary Existence: N = no, Y = yes. 3) Cluster Order: STV, VTS. Only Onset STV, Medial STV, and Medial VTS conditions had enough speakers with enough tokens for analysis. While Final VTS heteromorphemic clusters also had enough counts, there was no direct comparison available for any hypotheses: Final STV heteromorphemic and Final VTS tautomorphemic counts are too few and there are no morphemic onset or medial clusters. The 23 speakers with at least 30 tokens for all three conditions were considered for analysis. These speakers and their counts are bolded in the table. Appendix C shows the variance measures for each speaker.

Speaker	O.N.STV	O.N.VTS	M.N.STV	M.N.VTS	F.Y.STV	F.Y.VTS	F.N.STV	F.N.VTS
1 (f,y,f)	**48**	1	**45**	**40**	4	57	9	12
2 (f, o, m)	60	0	27	41	2	68	10	5
3 (m,o,m)	**47**	0	**37**	**31**	2	37	10	5
4 (f,y,f)	89	0	13	15	3	92	3	9
5 (f,o,f)	64	1	26	34	5	99	13	5
6 (m,y,f)	52	0	17	26	1	42	6	5
7 (f,o,f)	**77**	1	**73**	**37**	5	142	18	18
8 (f,y,f)	59	2	28	38	1	97	17	7
9 (f,y,f)	**50**	1	**67**	**65**	2	45	13	4
10 (m,o,f)	**72**	1	**108**	**46**	1	114	10	10
11 (m,y,m)	**76**	1	**68**	**44**	5	66	12	8
12 (f,y,m)	**88**	0	**42**	**37**	4	192	14	12
13 (m,y,f)	**74**	1	**94**	**50**	2	123	14	10
14 (f,o,f)	**45**	0	**83**	**52**	3	48	13	11
15 (m,y,m)	**59**	2	**84**	**64**	0	62	8	8
16 (f,o,m)	**164**	3	**66**	**44**	7	171	15	35
17 (f,o,m)	57	0	20	29	2	55	11	4
18 (f,o,f)	**96**	1	**41**	**51**	2	120	17	7
19 (m,o,f)	**121**	1	**80**	**49**	0	111	11	11
20 (f,o,f)	10	0	21	18	0	22	7	4
21 (f,y,m)	38	1	43	19	3	55	6	2
22 (m,o,f)	41	0	55	24	3	27	5	6
23 (m,o,m)	**65**	0	**51**	**36**	1	112	14	7
24 (m,o,m)	48	1	60	24	3	22	12	5
25 (f,o,m)	**86**	2	**39**	**36**	4	139	20	4
26 (f,y,f)	36	0	22	22	2	67	10	3
27 (f,o,m)	**78**	0	**44**	**30**	2	41	6	4
28 (m,y,m)	73	1	34	25	3	158	8	3
29 (m,o,f)	**100**	1	**50**	**30**	1	106	7	8
30 (m,y,m)	**72**	1	**55**	**36**	3	115	12	12
31 (f,y,m)	44	1	26	34	1	97	9	8
32 (m,y,f)	**86**	1	**42**	**39**	4	68	14	14
33 (m,y,f)	**117**	0	**58**	**54**	6	128	16	16
34 (m,y,m)	**101**	3	**55**	**46**	3	136	21	7
35 (m,o,m)	**108**	1	**67**	**50**	11	137	11	4
36 (m,o,f)	50	0	24	19	3	34	5	6
37 (f,y,m)	57	2	14	30	0	52	5	2
38 (m,o,m)	101	3	47	27	0	127	15	19
39 (f,y,m)	**88**	0	**47**	**54**	4	105	10	13
40 (m,y,f)	65	0	27	36	3	101	12	7

Table 12: Token counts for each condition by speaker

Appendix C: Variance Measures of Sibilant Noise

Speaker	Gender	Age	Interviewer Gender	Onset SD	Medial SD	Onset MAD	Medial MAD
1	f	y	f	0.462	0.478	0.413	0.302
3	m	o	m	0.824	0.538	0.408	0.562
7	f	o	f	0.434	0.500	0.340	0.384
9	f	y	f	0.640	0.394	0.447	0.298
10	m	o	f	0.533	0.499	0.485	0.475
11	m	y	m	0.778	0.826	0.619	0.651
12	f	y	m	0.625	0.541	0.685	0.408
13	m	y	f	0.568	0.400	0.416	0.343
14	f	o	f	1.215	0.553	0.397	0.503
15	m	y	m	0.685	0.541	0.596	0.521
16	f	o	m	0.879	0.595	0.460	0.507
18	f	o	f	0.672	0.690	0.482	0.581
19	m	o	f	0.938	0.604	0.550	0.353
23	m	o	m	0.563	1.641	0.530	0.481
25	f	o	m	0.592	0.377	0.582	0.373
27	f	o	m	0.775	0.475	0.591	0.495
29	m	o	f	0.423	0.408	0.436	0.441
30	m	y	m	0.644	0.731	0.634	0.665
32	m	y	f	0.429	0.397	0.417	0.340
33	m	y	f	0.639	0.620	0.567	0.599
34	m	y	m	0.770	1.256	0.561	0.596
35	m	o	m	0.808	0.612	0.449	0.428
39	f	y	m	0.603	0.634	0.455	0.569
Mean				0.674	0.622	0.501	0.473
Median				0.640	0.541	0.482	0.481

Table 13: Variances of ST Ratio by Speaker in Onset vs. Medial STV Clusters

Speaker	Gender	Age	Interviewer	STV	VTS	STV	VTS
			Gender	SD	SD	MAD	MAD
1	f	y	f	0.478	0.531	0.302	0.604
3	m	o	m	0.538	0.517	0.562	0.417
7	f	o	f	0.500	0.877	0.384	0.755
9	f	y	f	0.394	0.821	0.298	0.684
10	m	o	f	0.499	0.633	0.475	0.550
11	m	y	m	0.826	0.747	0.651	0.774
12	f	y	m	0.541	0.693	0.408	0.752
13	m	y	f	0.400	0.967	0.343	0.758
14	f	o	f	0.553	0.782	0.503	0.361
15	m	y	m	0.541	0.894	0.521	0.651
16	f	o	m	0.595	0.749	0.507	0.432
18	f	o	f	0.690	0.664	0.581	0.441
19	m	o	f	0.604	0.820	0.353	0.720
23	m	o	m	1.641	0.478	0.481	0.396
25	f	o	m	0.377	0.715	0.373	0.292
27	f	o	m	0.475	0.798	0.495	0.701
29	m	o	f	0.408	0.787	0.441	0.540
30	m	y	m	0.731	0.716	0.665	0.793
32	m	y	f	0.397	0.366	0.340	0.362
33	m	y	f	0.620	0.769	0.599	0.845
34	m	y	m	1.256	1.015	0.597	0.878
35	m	o	m	0.612	1.223	0.428	0.760
39	f	y	m	0.634	0.658	0.569	0.601
Mean				0.622	0.749	0.473	0.612
Median				0.541	0.749	0.481	0.651

Table 14: Variances of ST Ratio by Speaker STV vs. VTS Medial Clusters