The phonetic specificity of competition: Contrastive hyperarticulation of voice onset time in conversational English

Noah Richard Nelson[a]*, Andrew Wedel[a]

*Corresponding author: NoahNelson@email.arizona.edu, Tel: +1 520 621 6897
[a] Department of Linguistics, University of Arizona, P.O. Box 210025, Tucson, AZ, 85721, USA

Abstract

Competition between words in the lexicon is associated with hyperarticulation of phonetic properties in production. This correlation has been reported for metrics of competition varying in the phonetic specificity of the relationship between target and competitor (e.g., neighborhood density, onset competition, cue-specific minimal pairs). Sampling a systematic array of competition metrics, we tested their ability to predict voice onset times in both voiced and voiceless word-initial stops of conversational English. Linear mixed effects models were compared according to their corrected Akaike's Information Criterion (AIC$_c$) values. High-performing models were evaluated using evidence ratios, with the competition metrics of top-performing models tested for significance using nested model comparisons. Words with a minimal pair defined for initial stop voicing were contrastively hyperarticulated, with shorter voice onset times for voiced stops and longer voice onset times for voiceless stops. No other competition metric reliably predicted hyperarticulation for both stop types. These results suggest that contrastive hyperarticulation is phonetically specific, increasing the perceptual distance between target and competitor.

Keywords

Competition, hyperarticulation, voice onset time, conversational speech, minimal pairs, neighborhood density

1   Introduction

A number of experimental and observational studies have reported that competition at the lexical level is correlated with hyperarticulation of phonetic properties in the target word. This correlation has been reported for English in a number of studies investigating the realization of vowel formants (e.g., Wright, 1997; 2004; Munson & Solomon, 2004; Munson, 2007; Scarborough, 2012), vowel durations (Schertz, 2013; Seyfarth, Buz, & Jaeger, 2016; but see Goldrick, Vaughn, & Murphy, 2013), degree of coarticulation (Scarborough, 2012, 2013), perseveration of voicing in coda fricatives (Seyfarth et al., 2016; Kharlamov, 2014), and initial stop voice onset time (Baese-Berk & Goldrick, 2009; Peramunage et al., 2011; Kirov & Wilson, 2012; Schertz, 2013; Fricke, 2013; Buz, Tanenhaus, & Jaeger, 2016; Fox, Reilly, & Blumstein, 2015; Fricke, Baese-Berk, & Goldrick, 2016). In each of these cases, some form of lexical competition has been found to correlate with hyperarticulation of phonetic properties of individual segments (e.g., Wright, 2004; Fricke, 2013; Buz et al., 2016; but see Goldrick et al., 2013; Gahl, 2015).

This work on competition-associated hyperarticulation is complicated, however, by the fact that 'competition' can be operationalized in a variety of ways. The two most common approaches to operationalizing competition are in terms of lexical-phonological *neighborhood density*, defined as the number of words that can be formed by adding, deleting, or substituting any single segment of the target word (often weighted for frequency; e.g., Luce & Pisoni, 1998), or in terms of *minimal pair relationships* defined for a specific phonetic cue (e.g., Baese-Berk & Goldrick, 2009). These two general approaches differ widely in terms of the specificity of the competition measure. For neighborhood density measures, competition anywhere in the word contributes to the measure, regardless of the phonetic relationship between the target and the competitor (though see Strand & Sommers, 2011, and Gahl & Strand, 2016, for a version of neighborhood density weighted according to perceptual similarity). On the other hand, the cue-specific minimal pair measure identifies the neighbor that differs from the target solely in the cue of interest. Consequently, we can think of the cue-specific minimal pair measure as being more phonetically specific than the neighborhood density measure.

Given the differences in the way these two competition metrics are defined, there is a great deal of intermediate ground between them. One way to think about this intermediate ground is in terms of a continuum of specificity in lexical competition (but see section 5.1 for discussion of alternatives). For simplicity, we conceptualize this continuum relative to a given phonetic cue in terms of two parameters that define lexical neighborhoods of varying specificity: the relative *type* of competition between target and competitor, and the relative *position* of that competition (Fig. 1). The most specific type and position on this continuum is defined by a cue-specific minimal pair competitor, which differs from the target word only in the measured cue, in the same segmental position that the cue is realized. For example, voice onset time is the primary cue distinguishing word-initial stop voicing in English (Lisker & Abramson, 1964; Lisker, 1986). Given the reference word *bill*, *pill* is a neighbor differing in the same cue (voice onset time) in the same segmental position. We can now instead define a more relaxed neighborhood for neighbors that share, for example, a manner of articulation in the same segment as the measured cue (e.g., *pill, kill, dill*…). Similarly, we can define the neighborhood to include any onset-competitor (e.g., *pill*, *will*, *mill*…). These two examples vary the phonetic *type* of competition, but hold the *position* of competition steady. We can also vary the relative position of competition. As an example, we can consider only those neighbors that differ in the second

position of the word; given the target word *bill*, neighbors differing only in the second segment include such words as *bell*, *bowl*, and *ball*.
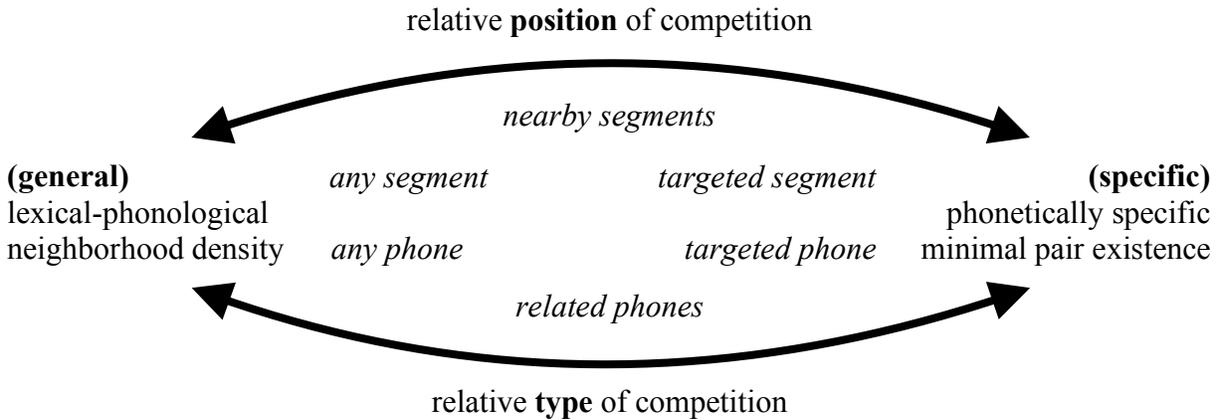
relative **position** of competition

relative **type** of competition

Fig. 1. Schematization of a continuum of specificity in competition. Single-edit competition metrics can lie at various positions along this continuum according to the relative type or position of competition with the target, in terms of neither (neighborhood density), or in terms of both (phonetically specific minimal pair competitor existence).

The present study compares a number of lexical competition metrics based on this continuum of specificity. Based on a systematic sample of alternatives from this continuum, we used mixed effects regression to predict the realization of voice onset time in a corpus of conversational English speech. We tested this sample of alternative metrics separately for both voiced and voiceless word-initial stops. While we compared a substantial set of competing hypotheses, we note that this range of metrics does not exhaust the set of plausible ways of operationalizing lexical competition. Indeed, a number of alternative metrics can be found in the literature, including neighborhood measures weighted for (i) the proportion of segmental overlap (Goldrick, Folk, & Rapp, 2010), (ii) variable Levenshtein edit distance between target and competitors (Yarkoni, Balota, & Yap, 2008), or (iii) perceptual similarity (Strand & Sommers, 2011; Gahl & Strand, 2016). Our results are considered with regard to these alternatives in the discussion.

This paper takes the following course. First, we briefly review cognitive mechanisms that have been proposed to underlie competition-associated hyperarticulation. Next we review prior research on competition-associated hyperarticulation, and the implications for each of these cognitive mechanisms. We then present results of analyses comparing a range of competition metrics for their ability to predict voice onset time in conversational speech. Finally, we discuss the implications of these results for the proposed cognitive mechanisms behind competition-associated hyperarticulation and for diachronic sound-change.

1.1   *Accounts of competition-associated hyperarticulation*

Why does competition among lexical items lead to hyperarticulation? One proposal appeals to processes of lexical selection and/or speech planning during production. In these production-internal approaches, lexical items compete for activation (e.g., Dell, 1986; see Fricke, 2013, ch. 2, for a review). This competition leads to overall higher activation levels for the target word, which in turn leads to enhancement of articulatory gestures in production (see Baese-Berk

& Goldrick, 2009, and Fricke, 2013, chs. 2 & 6, for discussion). A common implication of this hypothesis is that competition should only result in *increases* in phonetic durations such as voice onset time. Further, these accounts generally assume that competition-associated hyperarticulation should be relatively insensitive to the precise phonetic relationship between target and competitor (e.g., Baese-Berk & Goldrick, 2009; Fricke, 2013; Watson et al., 2015; see Jaeger & Buz, 2016, and Buz & Jaeger, 2016, for reviews).

Two alternative proposals predict that competition can result in targeted hyperarticulation of cues that distinguish a word from a specific competitor. According to communicative accounts (reviewed in Jaeger & Buz, 2016), speakers are implicitly aware that some words are more confusable than others, and are able to hyperarticulate the cues that maximally differentiate a given word from plausible alternatives. According to this approach, hyperarticulation is a tool that can be employed *online* by speakers in order to promote communicative efficiency. Over time, this online hyperarticulation can result in shifts in long-term lexical representations (reviewed in Hall et al., submitted). In a listener-internal approach, however, *hyper*-articulated tokens of potentially ambiguous words are more likely to be recognized, and therefore stored in memory, than *hypo*-articulated tokens. As a result, more ambiguous tokens contribute less to the exemplar cloud representing a given word or category, shifting it toward the hyperarticulated variant (see, e.g., Wedel 2006). Because this same exemplar cloud serves as the source for future productions, this process can, over time, lead to notable differences in the phonetic forms of words. Such an account does not predict that hyperarticulation happens online, but that, through filtering by listeners, words should come to be articulated in a way that enhances the specific phonetic contrasts that distinguish them. It is worth noting that all of these proposed mechanisms operate at different levels, and are therefore mutually compatible. We consider it a priori possible that all of these mechanisms contribute to the range of hyperarticulation effects found in human language.

1.2   The phonetic specificity of competition

The three cognitive mechanisms outlined above are rooted in very different assumptions about the nature of lexical competition. One critical question is how sensitive competition is to the phonetic or phonological relationship between target and competitor. According to most production-internal accounts, competition is relatively abstract, occurring among phonological representations prior to phonetic encoding (e.g., Baese-Berk & Goldrick, 2009; see also Goldrick & Rapp, 2007). For example, Fricke's (2013) Articulate As Soon As Possible Principle (AASAPP) posits that articulation unfolds on a segment-by-segment basis, and begins as soon as competition in a given segmental position is resolved. According to this model, the relative position of difference between the target and its competitors is therefore of paramount importance, but the phonetic relationships among those competitors is not relevant – a greater number of competitors defined for a particular segmental position leads to increased activation of the target segment regardless of phonetic relationship (Fricke, 2013; see also Vitevitch, Ambrüster, & Chu, 2004; Fricke et al., 2016). In contrast, both communicative and listener-internal accounts are based on perceptual confusability, such that hyperarticulation is targeted to those cues that maximize perceptual distinctiveness between competitors.

This question of how sensitive competition is to phonetic relationships is still open. In the case of word-initial stops, the focus of this paper, a number of researchers have argued that voice onset time is hyperarticulated in response to phonetically specific minimal pair competitors. This

correlation has been reported for voiceless stops (Baese-Berk & Goldrick, 2009; Peramunage et al., 2011; Kirov & Wilson, 2012; Schertz, 2013; Buz et al., 2016), but also for voiced stops, for which voice onset times have been found to *decrease* (Schertz, 2013; Wedel, Nelson, & Sharp, submitted; but see Ohala 1994, and Goldrick et al., 2013). Similar results have also been found for contrastive cues to coda voicing (vowel length and perseveration of voicing in English fricatives: Seyfarth et al., 2016; duration of fricatives and glottal pulsing in both stops and fricatives in Russian: Kharlamov, 2014). These results suggest that competition-driven hyperarticulation is *contrastive*, increasing the phonetic distance between target and competitor, and may be quite phonetically specific (e.g., Schertz, 2013; Seyfarth et al., 2016). Furthermore, Baese-Berk and Goldrick (2009) found a smaller, but significant effect of the mere lexical existence of the phonetically specific competitor, despite not being present in the experimental context (see also Peramunage et al., 2011). This additional result suggests that these effects may be partially mediated by processes that are not dependent on the presence of the competitor in the immediate context.

However, other researchers have reported evidence that this kind of minimal pair competition does not correlate with contrastive hyperarticulation. Goldrick, Vaughn, and Murphy (2013) used a word list reading task to elicit productions of voiced stop-initial words as well as voiced and voiceless stop-final words. For some of these words, the minimal pair defined for the voicing of their stop consonant existed (e.g., *bun ~ pun, coat ~ code*) and for others it did not (e.g., *bum ~ *pum, thud ~ *thut*). They found no significant effect of initial stop voicing minimal pair existence on the realization of voice onset time in voiced stops, nor did they find a significant effect of final stop voicing minimal pair existence on the realization of any cues to final voicing for voiceless stops (vowel duration, stop release, closure duration, or perseveration of voicing). They did find that final stop voicing minimal pair existence correlated significantly with vowel duration for final voiced stops, but the effect was opposite what would be expected under contrastive hyperarticulation. Rather than getting longer, vowels before voiced stops in words with minimal pair competitors in coda voicing were *shorter* than those in words without such a minimal pair, indicating a reduction of the contrast with the minimal pair competitor.

Finally, a number of researchers have argued that hyperarticulation results from more general competition. Fox et al. (2015) reported that phonetically specific minimal pair existence did not significantly predict voice onset time realizations in word list and sentence reading tasks, while neighborhood density did. Fricke (2013) and Fricke, Baese-Berk, and Goldrick (2016) reported that hyperarticulation of word-initial voice onset time is most robustly predicted by the number of competitors in the onset position. Fricke and colleagues argued that both cue-specific minimal pair competition and more general neighborhood density correlate with hyperarticulation of word-initial voice onset time because they *also* correlate with this position-based measure (see also Caselli, Caselli, & Cohen-Goldberg, 2015, and Vitevitch et al., 2004). Similarly, Kirov and Wilson (2012) reported that hyperarticulation of voice onset time in word-initial stops correlated with the presence of a minimal pair competitor for either stop voicing (e.g., *cap ~ gap*) or place of articulation (e.g., *cap ~ tap*), both in the initial segment. They further found that positions other than the initial segment, i.e. the vowel or coda of CVC monosyllables, did not affect initial stop voice onset time realizations (e.g., *cat ~ kit*; *cat ~ cap*). In a related study, Schertz (2013) reported contrastive voice onset time hyperarticulation in both voiced and voiceless word-initial stops associated with the presence of a voicing competitor. However, she found no effect of either place or manner of articulation competitors.

In summary, lexical competition, broadly construed, has been found to be consistently associated with contrastive hyperarticulation of voice onset time, but the nature of the associated competition has varied substantially across studies. One possible explanation for these differences is methodological. The majority of studies exploring contrastive hyperarticulation have used laboratory elicitation of defined materials in order to control for the many factors correlated with phonetic realization. Depending on experimental conditions, elicited speech may be prone to particularly clear or slow articulations that have been associated with increased phonetic durations and exaggerated articulatory gestures (e.g., Picheny et al., 1986; de Jong et al., 1993; Smiljanić & Bradlow, 2008). This "clear speech" may represent a conceptually distinct form of hyperarticulation from competition-driven contrastive hyperarticulation of the type reviewed above (for similar terminology, see Ohala, 1994; Seyfarth et al., 2016; for other discussions of different kinds of hyperarticulation, see Cho, Lee, & Kim, 2011; Schertz, 2013). We refer to this kind of generalized hyperarticulation as *clear-speech hyperarticulation*, which has been associated with increased voice onset times, particularly in voiceless stops (Smiljanić & Bradlow, 2008). As most studies examining the effects of competition on the realization of voice onset time have focused on voiceless stops (with the exceptions of Ohala, 1994, Schertz, 2013, and Goldrick et al., 2013), the predictions of both clear-speech and contrastive hyperarticulation have coincided. In these studies, clear-speech hyperarticulation may raise voice onset times toward a ceiling that makes contrastive hyperarticulation effects more difficult to detect (Kirov & Wilson, 2012; see also Wedel et al., submitted). Indeed, most of the studies that have found evidence of contrastive hyperarticulation have used task environments designed to amplify these potential effects, for example by including the minimal pair competitor in the immediate context (Baese-Berk & Goldrick, 2009, study 2; Kirov & Wilson, 2012; Seyfarth et al., 2016; Buz et al., 2016; but see Baese-Berk & Goldrick, 2009, study 1; Peramunage et al., 2011), or by explicitly indicating that the speaker's production was misperceived as the minimal pair competitor (Schertz, 2013; Buz et al., 2016).

One approach to minimize the influence of clear-speech hyperarticulation is to study speech in contexts promoting greater reduction, such as conversation (see, e.g., Gahl, Yao, & Johnson, 2012). A complementary strategy is to study phonetic cues for which the predictions of clear-speech and contrastive hyperarticulation diverge, such as in word-initial voiced stops (see Ohala, 1994, Schertz, 2013, and Goldrick et al., 2013). Contrastive hyperarticulation of voiced stops should lead to *shorter* voice onset times relative to non-hyperarticulated words (Schertz, 2013), while clear-speech conditions are not associated with shorter voice onset times (e.g., Miller, Green, & Reeves, 1986; Kessinger & Blumstein, 1997). Here, we study both voiced and voiceless stops in conversational speech, because contrastive hyperarticulation should move voice onset times in opposite directions for these two stop types (e.g., Schertz, 2013).

1.3   The present study

The present study was designed to clarify the specificity of the relationship among targets and competitors that leads to hyperarticulation. To this end, we compared a sample of competition measures from the continuum of specificity (Fig. 1) for their ability to predict the realizations of word-initial stop voice onset times in conversational English. In particular, we explored (i) the role of segmental position and (ii) the role of the phonetic relationship between segments within a given position. Though these two dimensions do not represent all possible hypotheses about how lexical competition should be operationalized, both of these dimensions

have important theoretical consequences for accounts of competition effects. For example, Fricke's (2013) AASAPP, among other production-internal accounts, predicts that hyperarticulation is targeted only to specific segmental positions, and can only lead to increased durations. Both trade-off and perception-based approaches, on the other hand, predict that hyperarticulation is targeted to specific cues and is contrastive, i.e., can be realized as either increased or decreased durations. Here, we look for contrastive hyperarticulation in the form of both longer voice onset times in voiceless stops, and shorter voice onset times in voiced stops. We investigate a sample of competing hypotheses regarding the relationship between target and competitor that gives rise to hyperarticulation, looking for these effects in conversational speech, where reduction is more likely to occur. Our study must therefore be considered alongside additional research testing alternative hypotheses, as well as investigating these phenomena in elicited speech paradigms.

2   Methods

We examined the effects of competition on the realization of voice onset time in conversational English based on a sampling of different competition metrics. These metrics took the form of modified neighborhood densities designed to sample two dimensions of relatedness in competition characterizing the differences between overall neighborhood density and cue-specific minimal pair competition. These modified neighborhood density measures targeted either the relative *position* of competition between target and competitor (i.e., *where* in the word the two words differ) or the relative *type* of competition between target and competitor (i.e., *how* the two words differ). We then used these metrics to predict voice onset time realizations of voiced and voiceless stops in the Buckeye Corpus of Conversational Speech (Pitt et al., 2005; Pitt et al., 2007). The use of conversational speech allows us to study the effects of competition in a generally more reduced context, limiting the potential for clear-speech hyperarticulation to mask the effects of contrastive hyperarticulation. The hypothesis that competition drives contrastive hyperarticulation predicts opposite effects for voiced versus voiceless stops: for voiced stops, voice onset times should decrease, while for voiceless stops, voice onset times should increase (Schertz, 2013). Including both stop types provides two distinct tests of this hypothesis.

We used linear mixed effects models to analyze the predictive relationship between competition and the realization of voice onset time. Due to the large number of hypotheses and the exploratory nature of the study, we evaluated these models using corrected Akaike's Information Criterion (AIC_c) comparisons and evidence ratios (Burnham & Anderson, 2004; Lukacs et al., 2007; Richards et al., 2011). This approach does not test for significance as in null hypothesis testing approaches to model comparison such as log-likelihood ratio tests, making it better suited to multiple hypothesis testing, in which multiple comparisons increase the risk of family-wise error (see Chamberlain, 1890, and Shadish, 1993, for discussion of the merits of testing multiple working hypotheses). In addition, evidence ratios allow for quantified statements about the relative support in favor of one model over another, ideal for comparing competing hypotheses (Burnham & Anderson, 2004; Richards et al., 2011). After comparing the models using AIC_c, we tested the top-performing model for both voiced and voiceless stops to ensure that the competition measure was contributing significantly to model fit using log-likelihood

ratio tests of nested models. If the competition measure in the top-performing model does not contribute significantly to model fit, the relative ranking of models is not clearly interpretable.

2.1   Materials

2.1.1   The data source

We used natural speech data from the Buckeye Corpus of Conversational Speech (Pitt et al., 2005; Pitt et al., 2007). The Buckeye Corpus includes 40 hours of conversation, spread over 1-hour long interviews with 40 individuals. Half of the interviewees are male, and half are female; half are over the age of 40 and half are under the age of 30. Interviews were conducted in Columbus, Ohio, and all speakers are from Columbus or surrounding regions of Ohio. The Buckeye Corpus is annotated for utterances as well as words and their syntactic category, and includes phonetic transcriptions with segment-level durations, but no sub-segmental measurements such as voice onset time.

2.1.2   Voice onset time measurements

We used measurements of the voice onset times of voiced and voiceless stops from a balanced set of 24 out of the 40 speakers in the corpus. Measurements were made by hand for stop-initial content words (labeled in the corpus as noun, verb, adjective, or adverb) of one or two syllables produced by these speakers. Our study thus differs from much previous work on competition effects in that our data includes two-syllable words. In addition, unlike in many other studies where only CV-initial words were included (e.g., Baese-Berk & Goldrick, 2009; Kirov & Wilson, 2012; Schertz, 2013; Fricke, 2013; Fox et al., 2015), we expanded our word types to include complex onsets. Because of the phonotactics of English, this meant the inclusion of words with an initial stop followed by a liquid or a glide.

Each word was annotated by hand in Praat (Boersma & Weenink, 2013, ver. 5.3.39) for the beginning of the stop closure, the beginning of the burst, and the beginning of the following sonorant. Based on those measurements, the total stop duration, the closure duration, and the voice onset time were calculated. We excluded tokens with pre-voicing, tokens with no identifiable burst, and tokens with closure durations or voice onset times that were more than 3 standard deviations from the specific speaker's mean for that stop consonant. For the remaining tokens, we calculated the proportion of the total stop duration that was taken up by the voice onset time (VOT). Henceforth, we refer to this measurement as the "VOT-length ratio." We elected to use the VOT-length ratio rather than raw voice onset times because it provides a very local control for speech rate. Speech rate is highly correlated with voice onset time in voiceless stops of English (Kessinger & Blumstein, 1998; Yao, 2007), but this correlation is largely attributable to an effect on the entire stop duration. However, an increase or decrease in the VOT-length ratio necessarily reflects a change in the voice onset time independently of any change to the total stop duration (see Smiljanić & Bradlow, 2008, for use of this and a related proportional measure of voice onset time).

We excluded high frequency discourse markers and any content words homophonous with function words (Bell et al., 2009; Gahl et al., 2012; Seyfarth, 2014). Because we based our neighborhood measures on the lemma forms of words (see section 2.1.3 below), we further excluded verbs with stem-vowel changes in their morphology (e.g., *buy ~ bought*, *come ~ came*), as the specific minimal pair competitor lemma is not consistent across the paradigm (e.g., *pie ~ buy* versus *pot ~ bought*). Of the remaining words included in our dataset, tokens were excluded

if there was no identifiable burst, if the token was immediately preceded by another stop consonant (due to unreliability of assigning the beginning of the stop closure), if the token followed an annotated pause or disfluency, or if either the aspiration or closure length of the token was more than three standard deviations from the speaker's mean.

2.1.3 Sources for competition metrics

All of the competition metrics used in both studies were calculated based on lemma forms. We used the Corpus Of Contemporary American English (COCA: Davies, 2012) to filter the Carnegie-Melon University pronouncing dictionary (CMU: Carnegie-Melon University, 2015) such that only distinct lemma forms remained. The CMU pronouncing dictionary, however, includes multiple pronunciations of many words, and so in order to prevent the same lemma from counting in any neighborhood measure more than once, we included only the first pronunciation of each distinct lemma. Only unique phonemic forms were retained, with the result that homophonic lemmas (e.g., *bear*, *bare*) correspond to the same entry. Because the CMU pronouncing dictionary is organized by orthographic form, a set of English homographs corresponding to distinct lemmas were manually retained (e.g., <tear> can be either the verb /tɛɹ/ or the noun /tɪɹ/). The full list of these homographs can be found in Appendix A. We retained the /ɑ ~ ɔ/ contrast (Durian, 2012), but collapsed /w ~ ʍ/ (Labov et al., 2006), consistent with pronunciation norms in Central Ohio.

Finally, we filtered this lemmatized CMU pronouncing dictionary based on contextual diversity, excluding any forms appearing in less than 0.5% of films in the SUBTLEX-US database. This was done to reduce the contribution of jargon or uncommon words to the resulting lexicon. Contextual diversity has been shown to be a better predictor of lexical decision accuracy and reaction time than frequency (Brysbaert & New 2009). The 0.5% cutoff results in a lexicon of 11 692 lemmas; for comparison, a frequency cut-off of 1 per million produces a lexicon of similar size (12 811 lemmas). This lexicon served as the source file when calculating all of our competition metrics.

2.2 Competition metrics

We created a systematic sample of competition metrics from the continuum of specificity as conceptualized in Figure 1. This continuum is defined by the two metrics that have dominated the literature on competition-induced hyperarticulation: lexical-phonological neighborhood density and cue-specific minimal pair competitor existence. For comparability with prior work, we included both of these competition measures in our study. Lexical-phonological neighborhood density was defined as the tally of all words that can be derived from the target by adding, deleting, or replacing any single phoneme of the target word (Luce & Pisoni 1998). Henceforth, we will refer to this measure as the 'overall neighborhood density.' Phonetically specific minimal pair competitor existence was coded as a binary neighborhood density. If switching the voicing value of the initial stop consonant of the target lemma resulted in a unique lemma in our lexicon (e.g., *bat ~ pat*), then this measure was coded as 1; otherwise, it was coded as 0. We refer to this metric as 'minimal pair competitor existence.'

To explore the space between these measures on the continuum, we used a sample of intermediate neighborhood density measures targeting either the relative position or type of competition. For the relative position of competition, we calculated three modified neighborhood densities, each targeting a different segmental position, or set of segmental positions, within the

word (Fig. 2). These three neighborhood densities targeted the first segmental position, the second segmental position, and the rest of the word (from the third segment to the last segment). We chose to include a metric for competition in the onset position because competition in this position has been found to be predictive in a number of studies (Vitevitch et al., 2004; Goldrick et al., 2010; Fricke, 2013; Caselli et al., 2015; Fricke et al., 2016). We included a metric for competition in the second segment because, as stop bursts often contain cues to the formants of following sonorants (Suchato & Punyabukkana 2005), hyperarticulation of bursts might arise when there are many competitors in the following segment. Finally, we included a metric for competition elsewhere in the word to evaluate whether phonetic variation is insensitive to segmental position. Note that all of these neighborhoods include only neighbors formed by substituting a single segment in the relevant position or set of positions; none of the modified neighborhood densities used in this study included additions or deletions of segments.
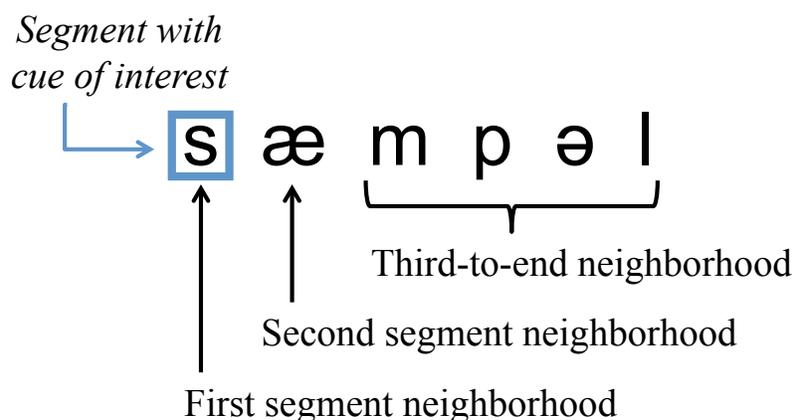
Segment with
cue of interest

s æ m p ə l

Third-to-end neighborhood

Second segment neighborhood

First segment neighborhood

Fig. 2. Neighborhoods for the relative position of competition in the word *sample*. Each neighborhood density references a segmental position or set of segmental positions. The tally of the density of each of these neighborhoods is increased by one if replacing the segment in one of that neighborhood's segmental positions results in a unique lemma.

For the relative type of competition, we calculated modified neighborhood densities targeting three types of competition, each within the first segmental position: the place of articulation, manner of articulation, and voicing value of the competitor relative to the target. For each of these, we calculated two neighborhood densities, one for competitors that share the feature with the target and one for competitors that have a different value for that feature relative to the target. This resulted in a total of six modified neighborhood densities targeting the relative type of competition (Fig. 3). This set of position- and type-based competition measures was not intended to provide an exhaustive survey of all possible competition relationships, but rather was chosen to provide a reasonable sample of possible phonological relationships between target and competitor (for similar approaches, see Kirov & Wilson, 2012; Schertz, 2013).
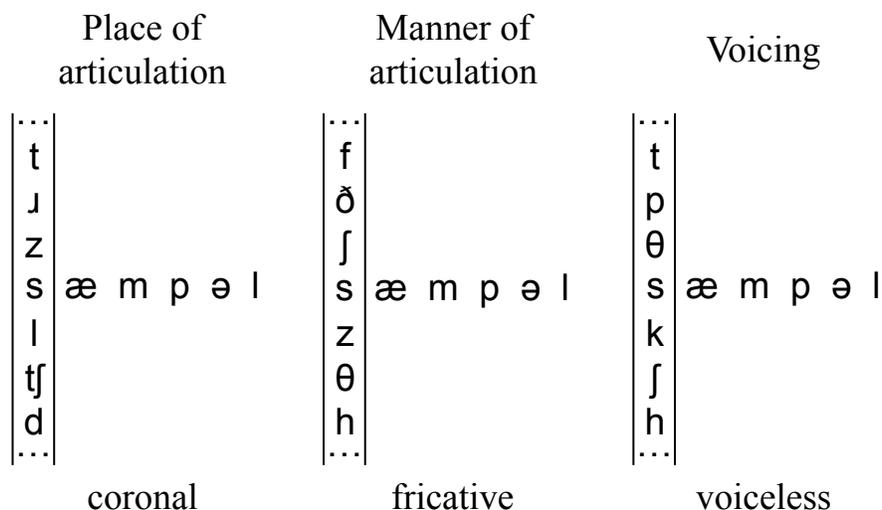
Fig. 3. The three types of same-feature competition included in our study, for the word *sample*. Each neighborhood represented here targets competitors with the same featural value as the target. Three other neighborhoods targeting competitors with different values for each of these three features were also calculated. The density of each of these neighborhoods is increased by one if replacing the first segment with another segment fitting that neighborhood's targeted featural type results in a unique lemma.

In order to calculate these competition metrics, we sorted the various phonemes of English into a simplified set of possible feature values. As our dependent measure was voice onset time, we sorted all consonants into one of three places of articulation: labial, coronal, or dorsal[1]. If a consonant could not be fully described by one of these feature values, that consonant was coded as the closest approximate place of articulation of these three (e.g., /f/ was coded as labial, as it is at least partly labial in articulation; /h/ was coded as dorsal, as dorsal is the closest place of articulation in our list to glottal). All consonants were additionally coded as either plosive or non-plosive, where "plosive" was defined as all oral stops and affricates (/p|b|t|d|k|g|t͡ʃ|d͡ʒ/), and "non-plosive" contained every other consonant, including nasal stops, approximates, and fricatives. Finally, all consonants were coded as either voiced or voiceless according to their standard phonemic transcription.

For a given target word, all neighborhood density measures were calculated by extracting the lemma form of the word and identifying the phonemic form of that lemma in our lexicon (sec. 2.1.3). We then identified each distinct lemma in that lexicon that fit the criteria of each of our competition metrics using regular expressions, and added it to the appropriate metric's neighborhood. Thus, each neighborhood density measure received a count of +1 for each unique lemma entry in our lexicon that could be formed by replacing the appropriate phoneme of the target word with another single phoneme, provided that the substitution fit the criteria of the neighborhood in question. In an effort to further test the range of possibilities in lexical competition, we also created versions of each of these neighborhood density measures that were weighted for frequency. We will turn to the description of these frequency-weighted measures in section 4.

---

[1] The single exception to this is /w/, which was coded as both labial and dorsal.

Competition metric	Segment(s) and feature(s) targeted	Examples for *pat* (/pæt/)	
Overall ND	Any segment Any feature	*spat* *at* *pit*	(/spæt/) (/æt/) (/pɪt/)
First segment ND	First segment only Any feature	*cat* *fat*	(/kæt/) (/fæt/)
Second segment ND	Second segment only Any feature	*pit* *pot*	(/pɪt/) (/pɔt/)
Third-to-end ND	Any segment after the second segment Any feature	*pan* *patch* *pack*	(/pæn/) (/pætʃ/) (/pæk/)
Same place of articulation ND	First segment only Same place of articulation	*fat* *mat*	(/fæt/) (/mæt/)
Different place of articulation ND	First segment only Different place of articulation	*cat* *sat*	(/kæt/) (/sæt/)
Same manner of articulation ND	First segment only Same manner of articulation	*cat* *bat*	(/kæt/) (/bæt/)
Different manner of articulation ND	First segment only Different manner of articulation	*fat* *that*	(/fæt/) (/ðæt/)
Same voicing ND	First segment only Same voicing value	*cat* *sat*	(/kæt/) (/sæt/)
Different voicing ND	First segment only Different voicing value	*that* *rat*	(/ðæt/) (/ræt/)
Minimal pair competitor existence	First segment only Same place of articulation Same manner of articulation Different voicing value	*bat*	(/bæt/)

Table 1. The competition metrics with descriptions and examples. Examples are non-exhaustive, provided for the target word *pat*.

2.3   *Mixed effects models*

The competition metrics were included in linear mixed effects models conducted in R (R Core Team 2016) using the lme4 package (Bates et al. 2015). Models were evaluated separately for voiced and voiceless stops due to different expectations for the effects of competition on the realization of voice onset times for each stop type (e.g., Schertz, 2013). For each of our stop types, each competition metric was included in a separate linear mixed effects regression model along with a consistent set of control predictors that have been shown to influence the realization of either voice onset time specifically, or word or segment durations more generally. For each subset of the data (voiced and voiceless stops), an additional model with only the control predictors and no competition metric was included as a baseline to which the other models could be compared (the "base model"). Thus, for both voiced and voiceless stops, the models in each set differed only in terms of the particular competition metric included (when one was included at all).

2.3.1   Factors of interest

<u>VOT-length ratio</u>: The dependent measure used in this study was the proportion of total stop duration taken up by the voice onset time (VOT-length ratio). This was calculated as the total voice onset time (from release of the closure to the end of aspiration) divided by the total stop duration (from initialization of the closure to the end of aspiration; Smiljanić & Bradlow, 2008; see also Wedel et al., submitted). As a proportion, the VOT-length ratios all lay on a scale between 0 and 1.

<u>Competition metrics</u>: Each model either included one of the competition metrics described above, or did not include any competition metric (base models). If the model included a competition metric, that value was centered, and then linearly transformed to a scale between -1 and 1 (note that this does not apply to the logical minimal pair existence factor).

2.3.2   Control predictors (fixed effects)

We included the following control predictors in all models. All continuous variables were centered, and then linearly transformed to a scale between -1 and 1 to facilitate model convergence.

<u>Stop phoneme</u>: The identity of the initial stop phoneme (/p|t|k/, /b|d|g/).

<u>Speech rate</u>: The number of phonemic vowels (i.e. syllables) per second (Bell et al., 2003). Speech rate is highly correlated with durational cues, including voiceless stop voice onset time (e.g., Yao, 2007).

<u>Contextual diversity</u>: The percent of films in which the word appears in the SUBTLEX-US database. Word familiarity measures correlate with reduction (e.g., word frequency: Bell et al., 2009), and contextual diversity has been found to be a better predictor of behavioral data than other familiarity measures such as frequency (Brysbaert & New, 2009).

<u>Following liquid</u>: Logical factor for whether or not the initial stop was followed by a liquid (/l|ɹ|ɚ/). Visual inspection of the data indicated that words followed by a liquid had considerably longer voice onset times.

<u>Number of syllables</u>: Either 1 or 2, coded as a factor. This factor was included under the assumption that bisyllabic words would have systematically shorter voice onset times on average.

<u>Conditional (forward/backward) bigram probability</u>: The log-transformed conditional bigram probability of each word given the preceding word (backward conditional bigram probability) and the following word (forward conditional bigram probability), based on a combination of the Fisher English Training Part 2 corpus (Cieri et al., 2005) and the Buckeye Corpus of Conversational Speech (Pitt et al., 2005; Pitt et al., 2007). Conditional probability is predictive of durations in the Buckeye corpus (Seyfarth, 2014).

<u>Previous mention</u>: Logical factor for whether or not the lemma (not necessarily the specific word) occurred earlier in the relevant transcript. Repetitions of words are more reduced than initial productions (Bell et al., 2009).

<u>Phonotactic probability</u>: The log-transformed average two-phoneme sequence probability, based on relative position within the word, calculated using the IPhOD2 database's unstressed, unweighted biphoneme probability calculator (Vaden et al., 2009).

2.3.3   Random effects

All models included random intercepts for both Speaker and Lemma. In addition, for each model with a competition metric, we included a correlated random slope for that metric on the Speaker intercept (Barr et al., 2013). We did not include random slopes for our control predictors as this would lead to problems with model convergence, and we had no principled reason to include any particular slopes over others.

2.4   Model evaluation procedure

Models were first evaluated according to their corrected Akaike's Information Criterion (AIC$_c$), an information-theoretic approach to model comparison based on entropy (see Burnham & Anderson, 2004), using the AICcmodavg package (Mazerolle, 2016). The AIC$_c$ value can be thought of in terms of information loss, where a lower AIC$_c$ value corresponds to less information loss and therefore a more accurate model. However, AIC$_c$ values by themselves are essentially meaningless, and must be interpreted relative to alternative AIC$_c$ values. To do this, models are included in a candidate set and ranked according to their AIC$_c$ value. These models are given weights based on their normalized log-likelihood, and compared based on these weights (Burnham & Anderson, 2004). This relative evaluation is operationalized in terms of the change in AIC$_c$ value between the current model and the top-performing model (Δ AIC$_c$). In general, models with a Δ AIC$_c \leq 2$ are considered to have substantial support relative to the top-performing model (i.e., they are not deemed considerably inferior to that model), and models with a Δ AIC$_c \geq 10$ are taken to have very little support (Burnham & Anderson, 2004). In addition, AIC$_c$ comparison allows evidence ratios to be calculated. Evidence ratios compare two models directly, and are calculated as the ratio of the AIC$_c$ weights of the two models (Richards et al., 2011). These evidence ratios allow us to describe the models in terms of the amount of evidence in favor of the better model with respect to the other.

In addition to the AIC$_c$ model comparison, we further tested the top-performing model of each set (voiced and voiceless stop models) for the significance of the competition metric. To test for statistical significance, we subjected these models to nested model comparison using the log-likelihood ratio test. In this case, the models in question were compared to the corresponding model with the fixed effect of the competition metric removed. Thus, for both voiced and voiceless stops separately, the model with the lowest AIC$_c$ value was compared to a restricted version of itself without the fixed effect of the relevant competition metric.

3   Results

Results are presented for voiced (section 3.1) and voiceless (section 3.2) stops separately. Summaries of the top-performing models can be found in Appendices B (voiced) and C (voiceless).

3.1   Voiced stops

The voiced stop data included 2 267 observations meeting our criteria from 24 speakers, distributed over 293 lemmas. Variance inflation factors (VIFs) for all of the predictors of interest

were less than 1.9, indicating low multi-collinearity between the factors of interest and control predictors. The overall neighborhood density metric had the highest VIF of all factors of interest (1.80), while minimal pair existence had the smallest VIF (1.21). VIFs less than 2 are generally not cause for concern (see O'Brien, 2007 and Belsley, Kuh, & Welsch, 1980/2005, ch. 3, for discussion).

3.1.1 AIC_c comparison

The AIC_c comparison table for all voiced stop models is presented in Table 2. Models are ranked according to their AIC_c value, with a lower AIC_c value corresponding to a better model fit. The model including a factor for minimal pair existence had the lowest AIC_c value ($\Delta AIC_c = 0$), and therefore the best overall fit to the data. The model including a factor for neighbors with the same place of articulation as the target was the second-best model (same place ND: $\Delta AIC_c = 4.49$), but evidence ratios indicate that there was about 9 times more evidence in favor of the minimal pair model. The only other model with a $\Delta AIC_c < 10$ was the neighborhood density for second segment competitors (second segment ND: $\Delta AIC_c = 8.13$), with about 58 times more evidence in favor of the minimal pair model. AIC_c comparisons suggest virtually no evidence in favor of any of the other competition metrics in the voiced stop dataset (AIC_c weight < 0.01; $\Delta AIC_c > 10$). These results suggest that the existence of the phonetically specific minimal pair competitor in the lexicon was the best predictor of voice onset time values for voiced stops in this dataset.

Rank	Model	K	AIC_c	ΔAIC_c	AIC_cWt	Cum.Wt	LL
1	Minimal pair exist	20	-4131.63	0.00	0.88	0.88	2086.00
2	Same place ND	20	-4127.14	4.49	0.09	0.97	2083.76
3	Second segment ND	20	-4123.49	8.13	0.02	0.99	2081.93
4	Overall ND	20	-4121.26	10.36	0.00	0.99	2080.82
5	First segment ND	20	-4118.50	13.13	0.00	1.00	2079.43
6	Base model	17	-4118.22	13.41	0.00	1.00	2076.24
7	Different manner ND	20	-4117.89	13.74	0.00	1.00	2079.13
8	Third-to-end ND	20	-4117.28	14.35	0.00	1.00	2078.83
9	Same voicing ND	20	-4116.93	14.70	0.00	1.00	2078.65
10	Different voicing ND	20	-4116.49	15.14	0.00	1.00	2078.43
11	Same manner ND	20	-4115.87	15.76	0.00	1.00	2078.12
12	Different place ND	20	-4115.83	15.80	0.00	1.00	2078.10

Table 2. AIC_c comparison table for all models predicting word-initial voiced stop VOT-length ratios. Models are ranked in order of AIC_c value. K = the number of estimable parameters in the model. AIC_c = the corrected AIC value. ΔAIC_c = the change in corrected AIC value between each model and the top performing model. AIC_cWt = the relative likelihood that the present model is the best model, presented as a proportional weight. $Cum.Wt$ = the cumulative AIC_c weight of the present model and all higher-ranked models. LL = the log-likelihood of the model.

3.1.2 *The top model: minimal pair existence*

The minimal pair existence model is summarized in Appendix B. As predicted, minimal pair existence correlates with a *decrease* in the VOT-length ratio, suggesting that these words are contrastively hyperarticulated away from their voice onset time competitor. As the top-performing model, we evaluated whether the minimal pair existence factor significantly contributed to model fit using nested model comparison. Removal of the fixed effect of minimal pair existence significantly affected model fit ($\chi^2(1) = 16.319, p < 0.001$), indicating that minimal pair existence is a significant predictor of voiced stop VOT-length ratios.

3.2 Voiceless stops

The voiceless stop data included 3 690 observations meeting our criteria from 24 speakers, distributed over 417 lemmas. VIFs for all of the predictors of interest were less than 1.8, indicating low multi-collinearity between the factors of interest and control predictors. The overall neighborhood density metric had the highest VIF of all factors of interest (1.70), while minimal pair existence had the smallest VIF (1.10).

3.2.1 AIC_c comparison

The AIC_c comparison table for all voiceless stop models is presented in Table 3. The model including a factor for minimal pair existence had the lowest AIC_c value ($\Delta AIC_c = 0$), and therefore the best overall fit to the data. The model including a factor for the second segment neighborhood density was the second-best model ($\Delta AIC_c = 2.56$), and evidence ratios indicate that there was only about 4 times more evidence in favor of the minimal pair model. Relative to all other models, the minimal pair model had limited support ($\Delta AIC_c > 2$). However, the minimal pair model had more substantial support relative to the base model ($\Delta AIC_c = 5.82$), with about 18 times more evidence in favor of the minimal pair existence model over the base model. These results suggest that the existence of the phonetically specific minimal pair competitor in the lexicon was the best predictor of voiceless stop voice onset times, but that the support for this model over many of the other models was limited.

Rank	Model	K	AIC_c	ΔAIC_c	AIC_cWt	Cum.Wt	LL
1	Minimal pair exist	20	-6124.87	0.00	0.55	0.55	3082.55
2	Second segment ND	20	-6122.31	2.56	0.15	0.70	3081.27
3	Same place ND	20	-6121.32	3.55	0.09	0.79	3080.77
4	Same manner ND	20	-6120.24	4.63	0.05	0.85	3080.24
5	Different voice ND	20	-6120.24	4.63	0.05	0.90	3080.24
6	Overall ND	20	-6119.39	5.48	0.04	0.94	3079.81
7	Base model	17	-6119.05	5.82	0.03	0.97	3076.61
8	First segment	20	-6117.66	7.21	0.01	0.98	3078.95
9	Different manner ND	20	-6116.30	8.57	0.01	0.99	3078.27
10	Different place ND	20	-6115.79	9.08	0.01	0.99	3078.01
11	Third-to-end ND	20	-6114.44	10.43	0.00	1.00	3077.34
12	Same voice ND	20	-6113.72	11.15	0.00	1.00	3076.98

Table 3. AIC_c comparison table for all models predicting word-initial voiceless stop VOT-length ratios. Models are ranked in order of AIC_c value. K = the number of estimable parameters in the model. AIC_c = the corrected AIC value. ΔAIC_c = the change in corrected AIC value between each model and the top performing model. AIC_cWt = the relative likelihood that the present model is the best model, presented as a proportional weight. $Cum.Wt$ = the cumulative AIC_c weight of the present model and all higher-ranked models. LL = the log-likelihood of the model.

3.2.2 The top model: minimal pair existence

The minimal pair existence model is summarized in Appendix C. As predicted, minimal pair existence correlates with an *increase* in the VOT-length ratio, suggesting that these words are contrastively hyperarticulated away from their voice onset time competitor. As the top-performing model, we evaluated whether the minimal pair existence factor significantly contributed to model fit using nested model comparison. Removal of the fixed effect of minimal pair existence significantly affected model fit ($\chi^2(1) = 5.25$, $p < 0.05$), indicating that minimal pair existence is a significant predictor of voiceless stop VOT-length ratios.

3.3   Summary and discussion

These results support the claim that minimal pairs are hyperarticulated away from each other in a manner that enhances the phonetic contrast between them. We found evidence of this contrastive hyperarticulation in the form of both longer voice onset times in voiceless stops, and shorter voice onset times in voiced stops. For both voiced and voiceless stops, the best model in terms of AIC_c included a factor for cue-specific minimal pair existence that contributed significantly to model fit. This suggests that competition defined for the cue of interest is the best predictor of contrastive hyperarticulation in word-initial stops of conversational English. We also found limited evidence that competition among neighbors that share a place of articulation in their initial segment (same place ND), as well as competition in the following segment (second segment ND), can influence voice onset time in word-initial stops (these were the second- and third-best models in terms of AIC_c for both voiced and voiceless stops); however, the evidence in support of these models was relatively weak (all $\Delta AIC_c > 2$).

In this dataset, we find evidence of contrastive hyperarticulation in both voiced and voiceless stops. To our knowledge, only three prior investigations of contrastive hyperarticulation have examined voiced stops (Ohala, 1994; Schertz, 2013; Goldrick et al., 2013). Schertz (2013) found that voiced stops were realized with shorter voice onset times in clarifications following misperception as their voiceless counterpart, paralleling our results here. Ohala (1994) found the same trend, but it was not significant. Goldrick, Vaughn, and Murphy (2013), however, found that word-initial voiced stops were not articulated differently depending on whether they had an initial stop voicing minimal pair in a word list reading task. These differences in results may be due, in part, to the use of different tasks. Studies employing word or sentence reading tasks to look for contrastive hyperarticulation of minimal pairs have sometimes found small effects (Baese-Berk & Goldrick, 2009, study 1; Peramunage et al., 2011; but see Fricke et al., 2016), but have sometimes failed to find any effect (e.g., Goldrick et al., 2013; Fox et al., 2015). On the other hand, studies in which cue-specific minimal pairs compete directly in communicative tasks have reported contrastive hyperarticulation effects more consistently or more robustly (Baese-Berk & Goldrick, 2009, study 2; Kirov & Wilson, 2012; Schertz, 2013; Seyfarth et al., 2016; Buz et al., 2016; but see Ohala, 1994). This suggests that contrastive hyperarticulation effects of the type reported here may be more easily detected under conditions promoting communication (Buz et al., 2016). That we find evidence of contrastive hyperarticulation of voiced stops may partly reflect the communicative nature of the speech in this dataset.

4   Alternative analyses based on frequency-weighted neighborhood measures

We repeated our analyses using frequency-weighted versions of the competition metrics described above. Following Luce and Pisoni (1998), these frequency-weighted measures were constructed as the ratio of the log-transformed target word frequency over the total neighborhood frequency, defined as the sum of the log-transformed frequencies of every member of the neighborhood (including the target). This method of frequency-weighting neighborhood density is commonly used in studies of competition-induced hyperarticulation (e.g., Munson, 2007; Scarborough, 2013; Buz & Jaeger, 2016). This measure is typically used in the choice of experimental materials to bin test items into 'hard' and 'easy' categories. Hard words have low frequency relative to their (relatively numerous) neighbors, while easy words have high frequency relative to their (relatively less numerous) neighbors (e.g., Wright, 2004). Log-

transformed frequencies were extracted directly from SUBTLEX-US (Brysbaert & New, 2009). Henceforth, we will refer to these frequency-weighted ratios as 'neighborhood frequency' measures.

As an illustration of the simplest case, the minimal pair neighborhood frequency consists of the log-transformed word frequency of the target divided by the sum of the log-transformed frequencies of the target and its voicing minimal pair competitor. Thus, if the target does not have a voicing minimal pair competitor (e.g., *bright*), the minimal pair neighborhood frequency is 1 (log-transformed target frequency divided by log-transformed target frequency + 0). If the target does have a minimal pair competitor, however, the denominator will include the frequency of that competitor, will therefore be larger than the numerator, and the resulting value of the ratio will be less than 1. Consequently, these neighborhood frequency measures range in value from 0 to 1, where 1 indicates that the target has no neighbors as defined for that neighborhood[2].

We repeated our mixed effects analyses as reported above for these neighborhood frequency measures. As before, the competition metrics were centered, and each appeared in its own model as both a fixed effect and as a random slope on speaker. Results are presented for voiced (section 4.1) and voiceless (section 4.2) stops separately.

4.1 *Voiced stops*

As before, the voiced stop data included 2 267 observations meeting our criteria from 24 speakers, distributed over 293 lemmas. VIFs for all of the predictors of interest were less than 1.8, indicating low multi-collinearity between the factors of interest and control predictors. The frequency-weighted metric corresponding to all neighbors with the same voicing value as the target had the highest VIF of all factors of interest (1.73), while the frequency-weighted minimal pair factor had the smallest VIF (1.17).

4.1.1 *AIC$_c$ comparison*

The AIC$_c$ comparison table for all voiced stop models is presented in Table 4. The model including the frequency-weighted minimal pair factor had the lowest AIC$_c$ value (Δ AIC$_c$ = 0), and therefore the best overall fit to the data. The model including a factor for the second segment neighborhood frequency was the second-best model (second segment ND: Δ AIC$_c$ = 4.44), and evidence ratios indicate that there was about 9 times more evidence in favor of the minimal pair model. The only other model with limited support was the model including a factor for neighbors that share their place of articulation with the target (same place ND: Δ AIC$_c$ = 4.52). All other models had little or no support (all Δ AIC$_c$ > 12). This result suggests that a frequency-weighted measure of cue-specific minimal pair competition is a better predictor of voiced stop voice onset times than other frequency-weighted neighborhood measures, mirroring the results for unweighted neighborhood densities.

---

[2] Mathematically, a value of 0 is impossible, but the value can come arbitrarily close to 0 in theory.

Rank	Model	K	AIC$_c$	Δ AIC$_c$	AIC$_c$Wt	Cum.Wt	LL
1	Minimal pair NF	20	-4132.31	0.00	0.82	0.82	2086.34
2	Second segment NF	20	-4127.86	4.44	0.09	0.91	2084.12
3	Same place NF	20	-4127.78	4.52	0.09	0.99	2084.08
4	Same voicing NF	20	-4120.05	12.26	0.00	1.00	2080.21
5	Different manner NF	20	-4119.53	12.78	0.00	1.00	2079.95
6	Base model	17	-4118.22	14.09	0.00	1.00	2076.24
7	First segment NF	20	-4118.02	14.28	0.00	1.00	2079.20
8	Third-to-end NF	20	-4117.78	14.52	0.00	1.00	2079.08
9	Overall NF	20	-4116.44	15.86	0.00	1.00	2078.41
10	Same manner NF	20	-4115.58	16.73	0.00	1.00	2077.98
11	Different voicing NF	20	-4115.37	16.94	0.00	1.00	2077.87
12	Different place NF	20	-4114.57	17.74	0.00	1.00	2077.47

Table 4. AIC$_c$ comparison table for all models predicting word-initial voiced stop VOT-length ratios. Models are ranked in order of AIC$_c$ value. K = the number of estimable parameters in the model. AIC_c = the corrected AIC value. $Δ AIC_c$ = the change in corrected AIC value between each model and the top performing model. AIC_cWt = the relative likelihood that the present model is the best model, presented as a proportional weight. $Cum.Wt$ = the cumulative AIC$_c$ weight of the present model and all higher-ranked models. LL = the log-likelihood of the model.

4.1.2   *The top model: frequency-weighted minimal pair competition*

The frequency-weighted minimal pair model is summarized in Appendix D. As predicted, the correlation between this factor and the VOT-length ratio is positive, indicating that voice onset time decreases as competition increases (i.e., as the frequency ratio decreases). This suggests that these words are contrastively hyperarticulated away from their voice onset time competitor. We evaluated whether the minimal pair existence factor significantly contributed to model fit using nested model comparison, and found that removal of the fixed effect of minimal pair existence significantly affected model fit ($χ^2(1) = 17.13$, $p < 0.001$). As with the unweighted neighborhood densities, this indicates that cue-specific minimal pair competition is a significant predictor of initial voiced stop voice onset times.

4.2   *Voiceless stops*

As before, the voiceless stop data included 3 690 observations meeting our criteria from 24 speakers, distributed over 417 lemmas. VIFs for all of the predictors of interest were less than 1.6, indicating low multi-collinearity between the factors of interest and control predictors. The second segment neighborhood frequency metric had the highest VIF of all factors of interest (1.53), while the frequency-weighted minimal pair factor had the smallest VIF (1.10).

4.2.1   *AIC$_c$ comparison*

The AIC$_c$ comparison table for all voiceless stop models is presented in Table 5. The model including the overall neighborhood frequency had the lowest AIC$_c$ value ($Δ AIC_c = 0$), and therefore the best overall fit to the data. No other model had substantial support (all $Δ AIC_c > 12$), and evidence ratios indicate that there was about 552 times more evidence in favor of the overall neighborhood frequency model over the next-best model (first segment NF: $Δ AIC_c = 12.63$). This result stands in stark contrast to the results of our other three analyses, in which cue-specific minimal pair competitor existence or neighborhood frequency provided the most predictive models of voice onset times.

Rank	Model	K	AIC$_c$	Δ AIC$_c$	AIC$_c$Wt	Cum.Wt	LL
1	Overall NF	20	-6141.40	0.00	1	1	3090.82
2	First segment NF	20	-6128.78	12.63	0	1	3084.50
3	Different voicing NF	20	-6126.32	15.08	0	1	3083.27
4	Second segment NF	20	-6123.87	17.53	0	1	3082.05
5	Different manner NF	20	-6122.92	18.48	0	1	3081.58
6	Same place NF	20	-6122.75	18.65	0	1	3081.49
7	Minimal pair NF	20	-6122.59	18.82	0	1	3081.41
8	Same manner NF	20	-6122.18	19.22	0	1	3081.21
9	Different place NF	20	-6121.11	20.29	0	1	3080.67
10	Base model	17	-6119.05	22.35	0	1	3076.61
11	Same voicing NF	20	-6117.03	24.38	0	1	3078.63
12	Third-to-end NF	20	-6116.15	25.25	0	1	3078.19

Table 5. AIC$_c$ comparison table for all models predicting word-initial voiceless stop VOT-length ratios. Models are ranked in order of AIC$_c$ value. K = the number of estimable parameters in the model. AIC_c = the corrected AIC value. ΔAIC_c = the change in corrected AIC value between each model and the top performing model. AIC_cWt = the relative likelihood that the present model is the best model, presented as a proportional weight. $Cum.Wt$ = the cumulative AIC$_c$ weight of the present model and all higher-ranked models. LL = the log-likelihood of the model.

4.2.2 The top model: overall neighborhood frequency

The overall neighborhood frequency model is summarized in Appendix E. As predicted, the correlation between this factor and the VOT-length ratio is negative, indicating that voice onset time increases as competition increases (i.e., as the frequency ratio decreases). We evaluated whether the overall neighborhood frequency factor significantly contributed to model fit using nested model comparison, and found that removal of the fixed effect of overall neighborhood frequency significantly affected model fit ($\chi^2(1) = 9.74$, $p < 0.01$).

4.2.3 A closer look at the neighborhood frequency metric

For the analyses of both frequency-weighted and unweighted competition metrics in the voiced stop data, as well as the analysis of unweighted measures in the voiceless stop data, cue-specific minimal pair competition provided the best predictor of voice onset time realizations. However, for frequency-weighted competition metrics in the voiceless stop data, the opposite measure as defined by our continuum of specificity was most predictive: overall neighborhood density.

Why might this result be so different from the others? One possibility lies in the distribution of data along the neighborhood frequency variable. A feature of the neighborhood frequency metric is that items with no neighbors have a neighborhood frequency of 1, because the ratio for words with no neighbors is the frequency of the word divided by itself. When items are included that have no neighbors, this can result in a distribution in which the majority of values are concentrated at the lower range, with an isolated peak at 1 corresponding to words with no neighbors. In the voiceless dataset, the bulk of the data is distributed between overall neighborhood frequency values of 0 and 0.2 (Fig. 4). There is a substantial gap in the observed values in the upper range, followed by a single peak corresponding to a neighborhood frequency ratio of 1. Extreme values can exert undue leverage on regression models, and notably, studies using neighborhood frequency to contrast 'hard' vs. 'easy' words have excluded words with no neighbors, i.e. those with a neighborhood frequency value of 1 (e.g., Wright, 2004; Munson & Solomon, 2004; Munson, 2007; Scarborough, 2013). To ask whether words with no neighbors

contributed to this anomalous result, we removed from the dataset the 337 observations of the 69 lemmas with no neighbors and repeated the analysis. 67 of these lemmas were bisyllabic, and the remaining 2 monosyllabic lemmas were the phonotactically unusual *prompt* and *puke*. The resulting dataset retained 3 353 observations from 24 speakers, distributed over 348 lemmas.
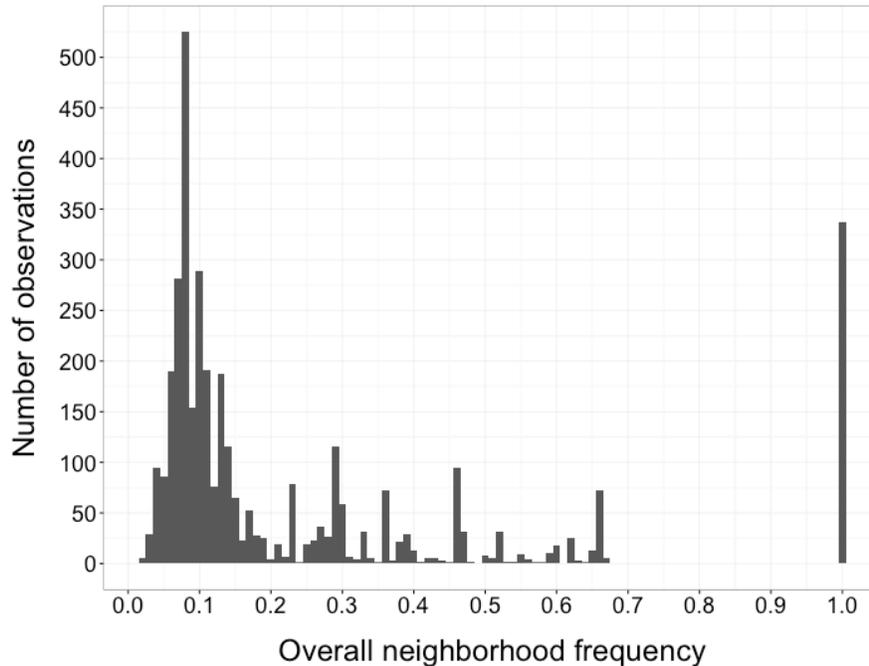

Fig. 4. The distribution of observations in the voiceless stop dataset by overall neighborhood frequency.

The AIC_c comparison table for all voiceless stop models is presented in Table 6. The model including the overall neighborhood frequency still had the lowest AIC_c value ($\Delta AIC_c = 0$), and therefore the best overall fit to the data. However, the overall neighborhood frequency factor no longer contributed significantly to model fit ($\chi^2(1) = 2.74$, $p > 0.05$; see Appendix E for details). Furthermore, two other models now varied minimally from the overall neighborhood frequency model: the first segment neighborhood frequency model ($\Delta AIC_c = 2.07$) and the minimal pair neighborhood frequency model ($\Delta AIC_c = 2.27$). Evidence ratios indicate that there was about 3 times more evidence in favor of the overall neighborhood frequency model relative to both the first segment neighborhood frequency model and the minimal pair neighborhood frequency model. These results indicate that overall neighborhood frequency is still the most predictive of these frequency-weighted competition metrics in this voiceless stop dataset, but the effect of this factor was not significant and the evidence in support of this model over the two subsequent models in the AIC_c ranking is limited. We confirmed that unweighted minimal pair existence was still significantly predictive of voice onset times in this subset of the voiceless stop data ($\chi^2(1) = 5.38$, $p < 0.05$), suggesting that the lack of significance for the overall neighborhood frequency factor is not purely due to a reduction of power.

Rank	Model	K	AICc	Δ AICc	AICcWt	Cum.Wt	LL
1	Overall NF	20	-5555.97	0.00	0.38	0.38	2798.11
2	First segment NF	20	-5553.90	2.07	0.13	0.51	2797.08
3	Minimal pair NF	20	-5553.71	2.27	0.12	0.63	2796.98
4	Different voicing NF	20	-5553.42	2.55	0.11	0.74	2796.84
5	Same place NF	20	-5552.42	3.55	0.06	0.80	2796.34
6	Second segment NF	20	-5552.10	3.87	0.05	0.85	2796.18
7	Base model	17	-5551.72	4.26	0.04	0.90	2792.95
8	Different manner NF	20	-5551.33	4.46	0.04	0.94	2795.79
9	Same manner NF	20	-5551.05	4.92	0.03	0.97	2795.65
10	Different place NF	20	-5549.75	6.23	0.02	0.99	2795.00
11	Same voicing NF	20	-5548.95	7.02	0.01	1.00	2794.60
12	Third-to-end NF	20	-5546.59	9.38	0.00	1.00	2793.42

Table 6. AICc comparison table for all models predicting word-initial voiceless stop VOT-length ratios in the subset of data for which overall neighborhood frequency is less than 1. Models are ranked in order of AICc value. K = the number of estimable parameters in the model. AIC_c = the corrected AIC value. $Δ AIC_c$ = the change in corrected AIC value between each model and the top performing model. AIC_cWt = the relative likelihood that the present model is the best model, presented as a proportional weight. $Cum.Wt$ = the cumulative AICc weight of the present model and all higher-ranked models. LL = the log-likelihood of the model.

4.3   Summary and discussion

The results for voiced stops were nearly identical for both frequency-weighted and unweighted measures. The competition metric corresponding to cue-specific minimal pair competition produced the best models in both cases, where minimal pair competition correlated with shorter voice onset times. The results for voiceless stops, however, varied substantially. While the analysis for unweighted competition metrics indicated that cue-specific minimal pair existence was the best predictor of voiceless stop voice onset times, the analysis for frequency-weighted competition metrics provided more evidence in favor of the overall neighborhood frequency measure. However, we noted that the neighborhood frequency metric results in a value of 1 for words without any neighbors, regardless of their frequency. This cluster of data at one extreme of the distribution creates the possibility for leverage, where a relatively small number of observations can have undue influence on the regression model. What is more, studies using this neighborhood frequency measure to categorize words as 'easy' or 'hard' have excluded words with no neighbors (e.g., Wright, 2004; Munson & Solomon, 2004; Munson, 2007; Scarborough, 2013). We repeated our analysis of the voiceless data with these words excluded and found that the overall neighborhood frequency model had the greatest support, but the effect of overall neighborhood frequency was no longer significant. These results suggest that, at least within this dataset, neighborhood frequency measures are not robustly predictive of voice onset time in voiceless stops.

We also found limited evidence in favor of both the first segment and minimal pair neighborhood frequency measures predicting voice onset times in voiceless stops (first segment: $Δ AIC_c = 2.07$; minimal pair: $Δ AIC_c = 2.27$). These findings parallel those of Fricke (2013), who reported that overall neighborhood density, minimal pair competitor existence, and the number of competitors differing in their onset ("rhyme neighbors" in her terminology) all correlated with hyperarticulation of voiceless stop voice onset times in the Buckeye Corpus (see also Fricke et al., 2016). However, Fricke reported that onset competition was a better predictor of voice onset time than either minimal pair competitor existence or overall neighborhood density, concluding that each of these other measures correlate with voice onset time because

they also correlate with the position-dependent measure of onset competition. Instead, we found that overall neighborhood density was a better predictor, but was not significant when we excluded words with no neighbors. Some methodological considerations may help to explain this different result. For example, Fricke only included monosyllabic words beginning with simplex onsets, while we included bisyllabic words and words with complex onsets. This difference affects our various neighborhood measures, particularly the neighborhoods corresponding to onset/first segment competition.

5   General Discussion

We identified a set of lexical competition metrics that sampled a conceptual space between the phonetically specific measure of cue-defined minimal pair existence, and the phonetically more general measure of neighborhood density. These measures were tested for their ability to predict voice onset time in voiced and voiceless word-initial stops of conversational English. Previous studies have suggested that competition from a cue-defined minimal pair competitor induces contrastive hyperarticulation of voice onset time (e.g., Baese-Berk & Goldrick, 2009; Schertz, 2013; Buz et al., 2016), and we found support for this hypothesis in the form of shorter voice onset times in voiced stops and longer voice onset times in voiceless stops (see also Wedel et al., submitted). However, previous studies using a variety of speech elicitation paradigms have suggested that less phonetically specific measures of competition also correlate with hyperarticulation of voice onset time in voiceless stops (e.g., Kirov & Wilson, 2012), in some cases more strongly than the cue-specific minimal pair competitor measure (Fricke, 2013; Fricke et al., 2016; Fox et al., 2015). In this natural speech dataset, we found that these other metrics of competition were less predictive of the realization of voice onset time in both voiced and voiceless stops.

We also conducted analyses in which we weighted our competition metrics for the relative frequencies of neighbors. For the voiced stops, this neighborhood frequency approach did not alter the results. The model including a frequency-weighted factor for the existence of an initial stop voicing minimal pair remained the most predictive of our models. For voiceless stops, however, overall neighborhood frequency provided the most predictive model. Upon further inspection, we noted the possibility that the extreme neighborhood frequency values contributed by words with no neighbors could unduly leverage model outcomes. With these data removed, the model including overall neighborhood frequency remained the most predictive in terms of AIC_c, but the contribution of overall neighborhood frequency to the model was no longer significant. Further, the models including factors for first segment and minimal pair neighborhood frequencies were minimally different from the overall neighborhood frequency model in terms of AIC_c.

5.1   Alternative metrics of lexical competition

The competition measures tested in this study represent only a subset of the large and multi-dimensional hypothesis space surrounding lexical competition effects. One way to expand this hypothesis space is to extend the continuum beyond a single-phoneme edit distance between target and competitor, so that similar words with a multi-phoneme edit distance can still contribute to competition. An example of such a measure in the literature is to weight lexical items for their Levenshtein edit distance, so that edit distances greater than 1 contribute to the

neighborhood, but less so than more closely related words (Yarkoni et al., 2008). A related, but different approach is to weight the neighbors for proportion of position-dependent segmental overlap with the target (Goldrick et al., 2010). Using either of these approaches, *trap* would not only have such neighbors as *trip*, *tap*, and *track*, but also neighbors such as *trick* and *tarp*, for which more than one segment are different. Proportion of segmental overlap has been found to be predictive of spoken and written errors in subjects with acquired language impairment, especially when additionally weighted for frequency, syntactic category, and onset overlap (i.e., the Lex-Form Composite: Goldrick et al., 2010). However, broader competition metrics such as these are less phonetically specific than the range of neighborhood measures tested in our study. We find that hyperarticulation of word-initial voice onset time in conversational English is primarily predicted by a competitor differing solely in word-initial voicing, suggesting that these broader alternatives should be less predictive of contrastive hyperarticulation in voice onset time.

Another competition metric from the literature compares the target to all words in the lexicon, weighting them for their perceptual similarity (Phi-square density: Strand & Sommers, 2011). Under this approach, *fist* and *fish* would be considered stronger competitors than *fist* and *fit*, because /s/ and /ʃ/ are perceptually similar, while the absence of /s/ is perceptually salient. This measure has been found to be predictive of word recognition accuracy, but whether it contributes predictive power above and beyond overall neighborhood density is unclear, and it was not found to be significantly predictive of word durations in the Buckeye corpus (Gahl & Strand, 2016). Although the phi-square density measure is gradiently sensitive to similarity, it considers perceptual relationships without respect to a particular cue. As such, neighbors that are perceptually similar in the segment or cue of interest, such as the cue-specific minimal pair competitor, are given no priority over neighbors differing in other positions or cues. It is unclear to what extent this measure should correlate with contrastive hyperarticulation of individual cues, but it may provide a fruitful model for studying the role of perceptual similarity in contrastive hyperarticulation (cf. Buz et al., 2016).

5.2   *Other cue-types*

If these results are representative, we expect to find similar contrastive hyperarticulation of other types of cues, where this hyperarticulation specifically increases phonetic distance between competitors. In another study of conversational speech, Wedel, Nelson, and Sharp (submitted) found evidence of hyperarticulation in vowels in the form of F1-F2 Euclidean distance. Vowels in words with minimal pairs defined by a nearby competitor vowel (e.g., *pit ~ pet*) were farther apart in F1-F2 space than the same vowels in words without such minimal pairs (e.g., *ship ~ *shep*). This difference was not explained by more general competition in the form of overall neighborhood density. In conjunction with the results reported here, this suggests that contrastive hyperarticulation of phonetic cues arises in response to competition with nearby competitors within that cue, and that this holds across two very different cue types (voice onset time and F1-F2 Euclidean distance).

5.3   *Cognitive mechanisms*

What do our results have to say about the cognitive mechanisms proposed to underlie competition-based hyperarticulation? In production-internal accounts that appeal to interactive cascading activation, the set of competitors has often been defined broadly as the set of lexical

neighbors differing by one edit-distance in any position. More recent accounts, however, have been modified to include differential neighborhood effects at different positions within words (e.g., Vitevitch et al., 2004; Goldrick et al., 2010; Fricke, 2013). For example, Fricke (2013) proposed that interactive cascading activation during motor planning is processed on a segment-by-segment basis. Each progressive segment is articulated as soon as competition is resolved, with greater competition leading to higher activation levels and, consequently, longer phonetic durations. According to this Articulate As Soon As Possible Principle, competition is sensitive to segmental position within the word, but is insensitive to the phonetic relationships among competitors within that segmental position. Consequently, in the case of word-initial voice onset time, this model predicts that the overall competition within the word onset should be the best predictor of competition-induced hyperarticulation.

Our findings did not support this hypothesis. In general, the existence of the minimal pair competitor defined by the initial stop voicing contrast was most predictive of voice onset times for both voiced and voiceless stops. This result is difficult to reconcile with any model of competition effects that does not consider the specific phonetic contrast between target and competitor. On the other hand, this result is consistent with communicative and listener-internal approaches, according to which the contrastive cues to phonemic and/or lexical categories can be specifically targeted for hyperarticulation (e.g., Pierrehumbert, 2002; Wedel, 2006; Buz et al., 2016; Hall et al., submitted). Notably, there is evidence that these cues may be especially targeted for hyperarticulation when they are most confusable. Buz, Tanenhaus, & Jaeger (2016) elicited productions of voiceless stop-initial monosyllabic words of English in a cooperative task. Speakers produced longer voice onset times on average when the minimal pair competitor was present in the experimental context, replicating past findings (e.g., Baese-Berk & Goldrick, 2009). Crucially, in a post-hoc analysis they found evidence that speakers were achieving this shift in average voice onset time not by increasing the target voice onset time of their productions, but by reducing the distribution of productions near the category boundary, where small changes to voice onset time have detectable effects on perception (McMurray, Aslin, & Tanenhaus, 2002). This suggests that speakers specifically avoid production variants that would result in greater perceptual similarity with a competitor.

If speakers contrastively hyperarticulate by avoiding more confusable productions, we expect that contrastive hyperarticulation may be less evident in conditions favoring clear-speech hyperarticulation, in which category contrasts may be inherently clearer. In elicited speech paradigms such as those reported above, voice onset times might be distinct enough a priori that the voiced and voiceless categories are completely non-overlapping[3]. In such a case, speakers may not attempt to hyperarticulate further, as the added effort would offer little in communicative benefit. On the other hand, if less formal registers induce greater overall reduction, we would expect phonetically contrastive cues to be less distinctive overall. In such a scenario, increasing articulatory effort to enhance this contrast would be worthwhile in words with minimal pair competitors defined for the relevant cue, but not for words without such a competitor.

What is more, the communicative aspect of this story is likely playing a further role. The studies that have found contrastive effects most consistently have involved (i) communication with a (sometimes simulated) human partner or (ii) clarifications of previous utterances (e.g.,

---

[3] Unfortunately, most of the relevant studies have investigated only voiced or voiceless stops, but not both. The one exception is Schertz (2013), who reports changes to voice onset time for each condition, but not raw voice onset time values. Consequently, it is difficult to test this hypothesis based on the existing data.

Baese-Berk & Goldrick, 2009, study 2; Kirov & Wilson, 2012; Schertz, 2013; Seyfarth et al., 2016; Buz et al., 2016). These conditions highlight the possibility that hyperarticulation is part of a communicative goal (Buz et al., 2016; Hall et al., submitted). Our conversational data similarly involve communication with an interlocutor (see also Wedel et al., submitted). In contrast, studies that have not found evidence of contrastive hyperarticulation of the kind reported here have largely used word or sentence reading tasks to elicit productions (e.g., Goldrick et al., 2013; Fox et al., 2015; but see Ohala, 1994; for studies finding minimal pair effects using list reading paradigms, see Baese-Berk & Goldrick, 2009, study 1; Peramunage et al., 2011). It is possible that, in such tasks, cue-defined minimal pair competitor existence fails to predict contrastive hyperarticulation as reliably as in more communicative tasks because there would be little communicative benefit to this hyperarticulation[4].

5.4   Implications for sound change

In what follows, we assume that consistent utterance-level bias in phonetic output can shift long-term mental representations, which can in turn shape the trajectory of sound change in a speech community (discussed in Wedel, 2012; Seyfarth, 2014). In this dataset, we observed that only competitors specifically distinguished from target words by the voice onset time cue robustly predicted contrastive hyperarticulation of that cue. This suggests in turn that contrastive hyperarticulation – and associated longer-term sound change – will be best predicted by the finer grained, cue-specific minimal pair relationships in the lexicon, rather than more abstract neighborhood relationships. Conversely, it suggests that more general measures of the functional load of phonetic cues such as phoneme-level entropy (Hockett, 1967; Surendran & Niyogi, 2006) will be predictive of sound change only through their correlation with the probability that a cue defines a minimal pair contrast. This is consistent with crosslinguistic evidence that phoneme contrasts distinguishing more minimal pairs are significantly less likely to merge over time (Wedel et al., 2013).

Further, the finding that competition results in significant shortening of voice onset time for voiced stops supports the notion that contrastive hyperarticulation creates greater perceptual distance to a competitor (Seyfarth et al., 2016; Buz et al., 2016; Wedel et al., submitted; see Hall et al., submitted, for discussion), rather than resulting in a general exaggeration of duration or extent of articulatory gestures. Slow or clear speech conditions, which are associated with increased phonetic durations, have not been found to correlate with shortening of voice onset time for initial voiced stops (Miller et al., 1986; Kessinger & Blumstein, 1997). This in turn suggests that phonetic distinctions that are on average more confusable should be more likely to trigger contrastive hyperarticulation. As reviewed above, this is supported by evidence that, for voiceless stops, contrastive hyperarticulation primarily suppresses shorter voice onset time productions that would be close to the voiced-voiceless category boundary (Buz et al., 2016). Taken together, these two observations predict that contrastive hyperarticulation will exert a greater influence on the trajectory of change for cue-distinctions (i) which are more perceptually confusable, and (ii) which distinguish a greater number of cue-specific minimal pairs. A

---

[4] In the case of Fox et al., (2015), greater phonological neighborhood density was nonetheless predictive of longer voice onset time realizations (see also Fricke et al., 2016). Alongside previous findings correlating neighborhood density with longer phonetic durations (see Fricke, 2013, chs. 2 & 6 for discussion), this raises the question of whether cue-defined minimal pair competition and overall neighborhood density index different kinds of competition and, consequently, different kinds of hyperarticulation.

corollary of this prediction is that contrastive-hyperarticulation is more likely to exert an effect through modulating production of reduced/rapid speech, where phonetic distinctions tend to be reduced, rather than through effects on careful/slow speech, where phonetic distinctions tend to be more robust.

5.5   Conclusion

We have argued that voice onset time is contrastively hyperarticulated in conversational speech in a way that increases the perceptual distance between lexical minimal pairs defined specifically for the voice onset time cue. The effect of minimal pair existence was more robust for voiced stops, for which voice onset times were found to decrease under competition. These results are most consistent with models of hyperarticulation effects that consider the fine-grained phonetic relationships among competitors, such as listener-oriented models that consider the communicative goals of speakers (e.g., to be understood). Furthermore, this is consistent with prior work arguing that fine-grained difference in individual-level productions of lexical minimal pairs can help to explain patterns of sound change such as phoneme merger (Wedel et al., 2013). In addition, we noted that prior work has emphasized (i) voiceless stop hyperarticulation, for which increased voice onset times are predicted under both contrastive and clear-speech hyperarticulation, and (ii) elicited speech paradigms, some of which may not provide motivation for contrastive hyperarticulation in speakers (e.g., word list or sentence reading tasks), and which may instead promote clear-speech hyperarticulation more generally. We argued that such studies are not ideally suited to inducing or detecting contrastive hyperarticulation, potentially explaining why some of these studies find small effects, but others do not. Furthermore, we noted that the pattern of results in the literature and in the data presented here suggest the importance of considering language as a tool for communication, where speakers hyperarticulate cues that facilitate effective communication, but are less likely to hyperarticulate when there is little benefit to doing so.

Acknowledgements

Appendix A

Homographs retained in final source lexicon

abuse	*advocate*	*allied*	*alternate*	*appropriate*	*articulate*	*bass*
bow	*buffet*	*close*	*combine*	*compact*	*complex*	*compound*
concert	*conduct*	*conflict*	*console*	*content*	*contract*	*convict*
decrease	*desert*	*dove*	*intimate*	*invalid*	*laminate*	*lead*
learned	*live*	*minute*	*moped*	*object*	*polish*	*present*
produce	*progress*	*read*	*rebel*	*record*	*refuse*	*resign*
resume	*separate*	*subject*	*tear*	*use*	*wind*	*wound*

List of English homographs manually retained in our source lexicon. Additional homographs not present on this list either involve inflectional variants (e.g., *confines* (n.) ~ *confines* (v., inflected)), have pronunciations varying only in

stress, which in our phonemic representations do not differ (e.g., *exploit* (n.) ~ *exploit* (v.)), or involve a form not represented in CMU at all (e.g., *blessed* (v.) ~ *blessed* (adj., not represented)).

Appendix B

Factor		Estimate	Std. Error	t-value
(Intercept)		0.241	0.013	18.654
MPexist = TRUE		-0.036	0.009	-4.146
Stop Phoneme				
	b	-0.107	0.008	-13.511
	g	0.034	0.010	3.417
SpeechRate		0.026	0.017	1.499
FollLiquid = TRUE		0.086	0.008	10.671
ContextDiversity		-0.010	0.015	-0.658
logBiphoneProb		-0.007	0.039	-0.174
logForBigramProb		0.039	0.010	3.864
logBackBigramProb		-0.012	0.010	-1.232
NumSyllables = 2		-0.007	0.007	-0.986
PrevMen = TRUE		0.006	0.004	1.335
Syntactic Category				
	Noun	-0.003	0.009	-0.354
	Adverb	0.017	0.013	1.261
	Verb	0.024	0.011	2.114

B1a. Summary of fixed effects for the model including minimal pair existence for voiced stops.

Groups	Name	Variance	Std. Dev.	Corr.
Lemma	(Intercept)	0.0009	0.030	
Speaker	(Intercept)	0.0012	0.035	
	MPexist = TRUE	0.0001	0.011	-0.71
Residual		0.0086	0.093	

B1b. Summary of random effects for the model including minimal pair existence for voiced stops.

Appendix C

Factor		Estimate	Std. Error	*t*-value
(Intercept)		0.557	0.015	37.98
MPexist = TRUE		0.023	0.010	2.34
Stop Phoneme				
	k	-0.018	0.008	-2.26
	p	-0.137	0.008	-17.23
SpeechRate		-0.033	0.015	-2.21
FollLiquid = TRUE		0.039	0.007	5.88
ContextDiversity		-0.019	0.013	-1.46
logBiphoneProb		-0.001	0.036	-0.03
logForBigramProb		0.007	0.009	0.79
logBackBigramProb		-0.012	0.009	-1.41
NumSyllables = 2		-0.0004	0.006	-0.08
PrevMen = TRUE		-0.0004	0.004	-0.12
Syntactic Category				
	Noun	0.004	0.010	0.42
	Adverb	0.015	0.023	0.65
	Verb	-0.006	0.011	-0.50

C1a. Summary of fixed effects for the model including minimal pair existence for voiceless stops.

Groups	Name	Variance	Std. Dev.	Corr.
Lemma	(Intercept)	0.0010	0.032	
Speaker	(Intercept)	0.0015	0.039	
	MPexist = TRUE	0.0004	0.020	0.08
Residual		0.0103	0.101	

C1b. Summary of random effects for the model including minimal pair existence for voiceless stops.

Appendix D

Factor		Estimate	Std. Error	*t*-value
(Intercept)		0.228	0.012	18.883
MP neighb. freq.		0.084	0.020	4.296
Stop Phoneme				
	b	-0.108	0.008	-13.771
	g	0.033	0.010	3.383
SpeechRate		0.026	0.017	1.493
FollLiquid = TRUE		0.085	0.008	10.603
ContextDiversity		-0.013	0.015	-0.906
logBiphoneProb		-0.008	0.039	-0.202
logForBigramProb		0.039	0.010	3.863
logBackBigramProb		-0.012	0.010	-1.252
NumSyllables = 2		-0.007	0.007	-1.007
PrevMen = TRUE		0.006	0.004	1.342
Syntactic Category				
	Noun	-0.003	0.009	-0.362
	Adverb	0.018	0.013	1.318
	Verb	0.025	0.011	2.218

D1a. Summary of fixed effects for the model including minimal pair neighborhood frequency for voiced stops.

Groups	Name	Variance	Std. Dev.	Corr.
Lemma	(Intercept)	0.0009	0.030	
Speaker	(Intercept)	0.0010	0.032	
	MP neighb. freq.	0.0005	0.023	0.64
Residual		0.0086	0.093	

D1b. Summary of random effects for the model including minimal pair neighborhood frequency for voiced stops.

Appendix E

Factor	Estimate	Std. Error	t-value
(Intercept)	0.554	0.015	37.90
Overall neighb. freq.	-0.048	0.015	-3.26
Stop Phoneme			
k	-0.015	0.008	-1.91
p	-0.132	0.008	-16.52
SpeechRate	-0.034	0.015	-2.29
FollLiquid = TRUE	0.043	0.007	6.47
ContextDiversity	-0.021	0.013	-1.65
logBiphoneProb	-0.003	0.035	-0.08
logForBigramProb	0.008	0.009	0.92
logBackBigramProb	-0.014	0.008	-1.61
NumSyllables = 2	0.007	0.006	1.20
PrevMen = TRUE	-0.0008	0.004	-0.22
Syntactic Category			
Noun	0.004	0.010	0.37
Adverb	0.012	0.023	0.55
Verb	-0.008	0.011	-0.67

E1a. Summary of fixed effects for the model including overall neighborhood frequency for voiceless stops.

Groups	Name	Variance	Std. Dev.	Corr.
Lemma	(Intercept)	0.0010	0.031	
Speaker	(Intercept)	0.0016	0.040	
	Overall neighb. freq.	0.0013	0.036	-0.16
Residual		0.0102	0.101	

E1b. Summary of random effects for the model including overall neighborhood frequency for voiceless stops.

Factor	Estimate	Std. Error	*t*-value
(Intercept)	0.558	0.015	37.13
Overall neighb. freq.	-0.043	0.026	-1.68
Stop Phoneme			
k	-0.013	0.008	-1.64
p	-0.132	0.008	-15.71
SpeechRate	-0.030	0.016	-1.92
FollLiquid = TRUE	0.049	0.007	6.66
ContextDiversity	-0.018	0.014	-1.33
logBiphoneProb	-0.015	0.037	-0.40
logForBigramProb	0.006	0.009	0.68
logBackBigramProb	-0.015	0.009	-1.65
NumSyllables = 2	0.009	0.006	1.41
PrevMen = TRUE	0.001	0.004	0.20
Syntactic Category			
Noun	-0.003	0.011	-0.25
Adverb	-0.001	0.025	-0.04
Verb	-0.014	0.012	-1.12

E2a. Summary of fixed effects for the model including overall neighborhood frequency for voiceless stops when words with no single-phoneme edit distance neighbors are removed.

Groups	Name	Variance	Std. Dev.	Corr.
Lemma	(Intercept)	0.0010	0.032	
Speaker	(Intercept)	0.0013	0.037	
	Overall neighb. freq.	0.0014	0.037	-0.95
Residual		0.0103	0.102	

E2b. Summary of random effects for the model including overall neighborhood frequency for voiceless stops when words with no single-phoneme edit distance neighbors are removed.

References

Baese-Berk, M. M. & Goldrick, M. (2009). Mechanisms of interaction in speech production. *Language and Cognitive Processes,* 24, 527-554.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68, 255-278.

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software,* 67(1), 1-48. doi:10.18637/jss.v067.i01.

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., & Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *The Journal of the Acoustical Society of America*, *113*(2), 1001-1024.

Bell, A., Brenier, J., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, 60, 92–111.

Belsley, D. A., Kuh, E. & Welsch, R. E. (1980/2005). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: Wiley.

Boersma, P., & Weenink, D. (1992–2013). Praat: Doing phonetics by computer (Version 5.3.39). Amsterdam.

Brysbaert, M. & New, B. (2009). Moving beyond Kucera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods,* 41, 997-990.

Burnham, K. P. & Anderson, D. R. (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods and Research,* 33(2), 261-304. doi: 10.1177/0049124104268644.

Buz, E., Tanenhaus, M. K., & Jaeger, T. F. (2016). Dynamically adapted context-specific hyper-articulation: Feedback from interlocutors affects speakers' subsequent pronunciations. *Journal of Memory and Language*, *89*, 68-86.

Buz, E., & Jaeger, T. F. (2016). The (in)dependence of articulation and lexical planning during isolated word production. *Language, Cognition and Neuroscience*.

Carnegie Mellon University. (2015). CMUdict: The CMU Pronouncing Dictionary 0.7. http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

Caselli, N. K., Caselli, M. K., & Cohen-Goldberg, A. M. (2015). Inflected words in production: Evidence for a morphologically rich lexicon. *The Quarterly Journal of Experimental Psychology,* 1-23.

Chamberlin, T. C. (1890). The method of multiple working hypotheses. *Science*, *15*(366), 92-96.

Cho, T., Lee, Y., & Kim, S. (2011). Communicatively driven versus prosodically driven hyper-articulation in Korean. *Journal of Phonetics, 39*(3), 344-361.

Cieri, C., Graff, D., Kimball, O., Miller, D., & Walker, K. (2005). Fisher English Training Part 2. Linguistic Data Consortium, Philadelphia.

Davies, M. (2012). Corpus of Contemporary American English (1990-2012). Available at http://corpus.byu.edu/coca/.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological review*, *93*(3), 283.

Durian, D. (2012). A New Perspective on Vowel Variation across the 19th and 20th Centuries in Columbus, OH. Doctoral Dissertation: The Ohio State University.

Fox, N. P., Reilly, M., & Blumstein, S. E. (2015). Phonological neighborhood competition affects spoken word production irrespective of sentential context. *Journal of Memory and Language*, 83, 97–117.

Fricke, M. D. (2013). *Phonological encoding and phonetic duration* (Doctoral dissertation). University of California, Berkeley.

Fricke, M. D., Baese-Berk, M. M., & Goldrick, M. (2016). Dimensions of similarity in the mental lexicon. *Language, Cognition and Neuroscience*, *31*(5), 639-645.

Gahl. S. (2015). Lexical competition in vowel articulation revisited: Vowel dispersion in the Easy/Hard database. *Journal of Phonetics,* 49, 96–116.

Gahl, S., & Strand, J. F. (2016). Many neighborhoods: Phonological and perceptual neighborhood density in lexical production and perception. *Journal of Memory and Language*, *89*, 162-178.

Gahl, S., Yao, Y., & Johnson, K. (2012). Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of memory and language*, *66*(4), 789-806.

Goldrick, M., Folk, J. R., & Rapp, B. (2010). Mrs. Malaprop's neighborhood: Using word errors to reveal neighborhood structure. *Journal of Memory and Language, 62*(2), 113-134.

Goldrick, M., & Rapp, B. (2007). Lexical and post-lexical phonological representations in spoken production. *Cognition*, *102*(2), 219-260.

Goldrick, M., Vaughn, C., & Murphy, A. (2013). The effects of lexical neighbors on stop consonant articulation. *Journal of the Acoustical Society of America, 134*, EL172-EL177.

Hall, K. C., Hume, E., Jaeger, T. F., & Wedel, A. (submitted). The Message Shapes Phonology.

Jaeger, T. F. (2013). Production preferences cannot be understood without reference to communication. *Frontiers in psychology*, *4*, 230.

Jaeger, T. F. & Ferreira, V. (2013). Seeking predictions from a predictive framework. *Behavioral and Brain Sciences, 36*(4), 359-360.

Jaeger, T. F. & Buz, E. (2016). Signal reduction and linguistic encoding. In E. M. Fernández & H. S. Cairns (Eds.), *Handbook of psycholinguistics*. Wiley-Blackwell.

de Jong, K., Beckman, M. E., & Edwards, J. (1993). The interplay between prosodic structure and coarticulation. *Language and Speech,* 36, 197-212.

Kessinger, R. H. and Blumstein, S. E. (1997). Effects of speaking rate on voice-onset time in Thai, French, and English. *Journal of Phonetics, 25*(2), 143–168.

Kessinger, R. H., & Blumstein, S. E. (1998). Effects of speaking rate on voice-onset time and vowel production: Some implications for perception studies. *Journal of Phonetics*, *26*(2), 117-128.

Kharlamov, V. (2014). Incomplete neutralization of the voicing contrast in word-final obstruents in Russian: Phonological, lexical, and methodological influences. *Journal of Phonetics, 43*, 47-56.

Kirov, C., & Wilson, C. (2012). The specificity of online variation in speech production. *34th Annual meeting of the Cognitive Science Society,* Sapporo, Japan.

Kondaurova, M. V., & Francis, A. L. (2008). The relationship between native allophonic experience with vowel duration and perception of the English tense/lax vowel contrast by Spanish and Russian listeners. *The Journal of the Acoustical Society of America, 124,* 3959–3971. http://dx.doi.org/10.1121/1.2999341.

Labov, W., Ash, S., & Boberg, C. (2006). The atlas of North American English: Phonetics, phonology, and sound change. Berlin: Mouton de Gruyter.

Lisker, L. (1986). "Voicing" in English: a catalogue of acoustic features signaling/b/versus/p/in trochees. *Language and speech*, 29(1), pp??

Lisker, L., & Abramson, A. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word,* 20, 384–422.

Liu, R. & Holt, L. L. (2015). Dimension-based statistical learning of vowels. Journal of Experimental Psychology: Human Perception & Performance. ePub ahead of print.

Luce, P. A. & Pisoni, D.B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing,* 19, 1-36.

Lukacs, P. M., Thompson, W. L., Kendall, W. L., Gould, W. R., Doherty, P. F., Burnham, K. P., & Anderson, D. R. (2007). Concerns regarding a call for pluralism of information theory and hypothesis testing. *Journal of Applied Ecology,* 44, 456-460.

Mazerolle, M. J. (2016). AICcmodavg: Model selection and multimodel inference base on (Q)AIC(c). R package version 2.0-4. http://CRAN.R-project.org/package=AICcmodavg.

McMurray, B., Tanenhaus, M., and Aslin, R. (2002). Gradient effects of within-category phonetic variation on lexical access, *Cognition,* 86(2), B33-B42. PMCID: PMC2630474.

Miller, J. L., Green, K. P., & Reeves, A. (1986). Speaking rate and segments: A look at the relation between speech production and speech perception for the voicing contrast. *Phonetica, 43*(1-3), 106–115.

Munson, B. (2007). Lexical access, lexical representation, and vowel production. In J. Cole & J. I. Hualde (Eds.), *Laboratory phonology 9: Phonology and phonetics,* 201–227. Berlin: Mouton.

Munson, B., & Solomon, N. P. (2004). The effect of phonological neighborhood density on vowel articulation. *Speech, Language, and Hearing Research,* 47, 1048–1058.

O'brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity*, *41*(5), 673-690.

Ohala, J. J. (1994). Acoustic study of clear speech: a test of the contrastive hypothesis. In *Proceedings of the International Symposium on Prosody,* 75–89, Yokohama, Japan.

Peramunage, D., Blumstein, S. E., Myers, E. B., Goldrick, M., & Baese-Berk, M. M. (2011). Phonological neighborhood effects in spoken word production: An fMRI study. *Journal of Cognitive Neuroscience, 23*(3), 593–603.

Peterson, R. R., & Savoy, P. (1998). Lexical selection and phonological encoding during language production: Evidence for cascaded processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24, 539-557.

Picheney, M., Durlach, N. & Braida, L. (1986). Speaking clearly for the hard of hearing II: acoustic characteristics of clear and conversational speech. *Journal of Speech and Hearing Research,* 29, 434-446.

Pierrehumbert, J. (2002). *Word-specific phonetics* (Vol. 7, pp. 101-139). Berlin: Mouton de Gruyter.

Pitt, M., Johnson, K., Hume, E., Kiesling, S., Raymond, W. (2005). The Buckeye Corpus of conversational speech: labeling conventions and a test of transcriber reliability. *Speech Communication,* 45, 90-95.

Pitt, M., Dilley, L., Jonson, K., Kiesling, S., Raymond, W., Hume, E., & Foler-Lussier, E. (2007). *Buckeye Corpus of Conversational Speech (2nd release)*. Department of Psychology, Ohio State University, Columbus, OH.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Richards, S. A., Whittingham, M. J., & Stephens, P. A. (2011). Model selection and model averaging in behavioural ecology: the utility of the IT-AIC framework. *Behavioral Ecology and Sociobiology,* 65(1), 77-89.

Scarborough, R. (2012). Lexical similarity and speech production: Neighborhoods for nonwords. *Lingua*, *122*(2), 164-176.

Scarborough, R. (2013). Neighborhood-conditioned patterns in phonetic detail: Relating coarticulation and hyperarticulation. *Journal of Phonetics*, *41*(6), 491-508.

Schertz, J. (2013). Exaggeration of featural contrasts in clarifications of misheard speech in English. *Journal of Phonetics,* 41, 249–263.

Seyfarth, S. (2014). Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition*, *133*(1), 140-155.

Seyfarth, S., Buz, E., & Jaeger, T. F. (2016). Dynamic hyperarticulation of coda voicing contrasts. *Journal of the Acoustical Society of America,* 139(2), EL31 - EL37.

Shadish, W. R. (1993). Critical multiplism: A research strategy and its attendant tactics. *New directions for program evaluation*, *1993*(60), 13-57.

Smiljanić, R. & Bradlow, A. R. (2008). Stability of temporal contrasts across speaking styles in English and Croatian. *Journal of Phonetics*, 36(1). 91–113.

Strand, J. F., & Sommers, M. S. (2011). Sizing up the competition: Quantifying the influence of the mental lexicon on auditory and visual spoken word recognition. *The Journal of the Acoustical Society of America*, *130*(3), 1663-1672.

Suchato, A., & Punyabukkana, P. (2005, September). Factors in classification of stop consonant place of articulation. In *INTERSPEECH* (pp. 2969-2972).

Vaden, K. I., Halpin, H. R., & Hickok, G. S. (2009). Irvine Phonotactic Online Dictionary, Version 2.0. [Data file]. Available from http://www.iphod.com.

Vitevitch, M. S., Armbrüster, J., & Chu, S. (2004). Sublexical and lexical representations in speech production: effects of phonotactic probability and onset density. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(2), 514.

Watson, D. G., Buxó-Lugo, A., & Simmons, D. C. (2015). The effect of phonological encoding on word duration: Selection takes time. In E. Gibson & L. Frazier (Eds.), *Explicit and implicit prosody in sentence processing,* 85–98. Switzerland: Springer International Publishing. doi:10.1007/978-3-319-12961-7{_}5.

Wedel, A. B. (2006). Exemplar models, evolution and language change. *The linguistic review*, *23*(3), 247-274.

Wedel, A., Kaplan, A., & Jackson, S. (2013). High functional load inhibits phonological contrast loss: A corpus study. *Cognition, 128*, 179-186.

Wedel, A., Nelson, N. R., & Sharp, R. (submitted). The phonetic specificity of contrastive hyperarticulation in natural speech.

Wright, R. (1997). Lexical competition and reduction in speech: A preliminary report. Indiana University research on spoken language processing progress report no. 21 (pp. 471–485).

Wright, R. (2004). Factors of lexical competition in vowel articulation. In Local, J., Ogden, R. & Temple, R. (Eds.), Papers in laboratory phonology VI, 26–50. Cambridge: Cambridge University Press.

Yao, Y. (2007). Closure duration and VOT of word-initial voiceless plosives in English in spontaneous connected speech. *UC Berkeley Phonology Lab Annual Report* (183-225).

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review, 15*(5), 971-979.