The Specificity of Online Variation in Speech Production

Christo Kirov(kirov@cogsci.jhu.edu)
Department of Cognitive Science, 3400 N. Charles Street
Baltimore, MD 21218 USA

Colin Wilson (colin@cogsci.jhu.edu)
Department of Cognitive Science, 3400 N. Charles Street
Baltimore, MD 21218 USA

Abstract

Phonetic variation is sensitive to lexical properties of words, such as frequency and neighborhood density, as well as contextual properties, such as predictability. Previous studies of lexically-induced variation have observed that both vowels and consonants are phonetically enhanced in words from dense neighborhoods, and have suggested that this effect is modulated only by the number and frequency of the neighbors. To determine whether contextual variation is driven by cognitive processes similar to those underlying lexical enhancement, three experiments examined the effect of contextually salient neighbors on the phonetic realization of vowels and initial consonant aspiration. Enhancement was found only for consonants, and only when the neighbor differed from the target word in a single feature. Unlike lexical effects, contextually-driven phonetic enhancement reflects a highly specific competition among words, a finding that can be rationalized in terms of the utility of speaker effort within a Bayesian model of word communication.

Keywords: Speech production; lexical competition; communication; Bayesian modeling

Introduction

Competition among alternatives, and the need to resolve competition efficiently and correctly, are pervasive in speech perception and speech production (e.g., Luce & Pisoni 1998, Marslen-Wilson & Zwitserlood 1989, Dell & Gordon 2003). Listeners must determine the speaker's intended message as rapidly as possible given an inherently ambiguous signal. In speech production, words and sublexical units that are partially consistent with the intended message compete for realization at multiple levels of processing. A number of studies have examined how such competitive processes are reflected in the fine-grained phonetic realization of speech sounds.

The number and relative frequency of phonologically-similar words in the lexicon (lexical neighbors) are known to affect phonetic realization. We refer to such affects as *offline* because they appear to depend on relatively static lexical relationships among words rather than dynamic contextual factors. Researchers have found that "hard" words, those with low lexical frequency and many high frequency neighbors, tend to be phonetically enhanced relative to "easy" words with high frequency and few neighbors. Hard words beginning with aspirated consonants have longer aspiration (as measured by voice onset time, VOT) than easy words (Goldinger & Summers, 1989). They are pronounced with an expanded vowel space (Munson & Solomon, 2004; Wright, 2003), and also show increased vowel nasalization and vowel-to-vowel coarticulation (Scarborough, 2004).

Interestingly, offline phonetic enhancement effects seem to be rather general: they appear to depend only on the (frequency-weighted) density of a word's neighborhood, not on the precise phonological relationships between the word and its neighbors. For example, Scarborough (2004) found that words that were particularly confusable by their nasal consonant (i.e., had one or more lexical neighbors that differed in the position of the nasal) did not show greater vowel nasalization than words that were not similarly confusable. Words like *stem*, with minimal pair neighbor *step*, showed similar levels of nasalization as words like *plank*, with no nasal-differing neighbors in the lexicon. Similarly, Goldinger & Summers (1989) found more VOT enhancement for voiceless-initial words from dense neighborhoods than those from sparse neighborhoods, even though both sets of words had exactly one minimal pair lexical neighbor that began with a voiced sound. The generality of offline phonetic enhancement suggests that it is driven by competition among entire lexical items, not among sublexical units.

Unlike offline effects, *online* effects on phonetic realization by definition depend on the context in which a word is uttered, such as the discourse topic, transitional probabilities conditioned on preceding material, and other contextually-salient words. A classic online effect is the Lombard Reflex, a set of vocal changes that include increases in amplitude and pitch that occur when speakers attempt to talk over noise (Lau, 2008; Zhao & Jurafsky, 2009). More recently, a number of corpus-based studies have found that the contextual predictability of speech elements, including phonemes and syllables, is inversely related to their length. Less predictable elements tend to be longer (e.g., Cohen-Priva & Jurafsky 2008, Aylett & Turk 2004, van Son & Pols 2003).

This paper aims to expand our understanding of online phonetic enhancement effects, looking not just at predictability effects but at how a word's phonological neighborhood in context — the sound structure of contextually salient competitors — affects phonetic realization along several dimensions of possible hyperarticulation. This will provide further insight into how competition between similar words plays out during speech production. In previous work, Baese-Berke & Goldrick (2009) found that VOT is lengthened when a voiceless-stop initial word is pronounced in the context of a voiced-initial neighbor (in comparison to the context of a phonologically unrelated filler word). For example, *cot* shows increased initial VOT in the context of *got*, but not in

the context of *fan*.

Baese-Berke & Goldrick additionally put forward the claim that this online enhancement of VOT, and perhaps online enhancement effects in general, have the same underlying cognitive mechanism as the offline enhancement effects reviewed earlier. If this hypothesis were true, we would expect offline and online effects to be empirically parallel. In particular, we would expect to find an online analogue of every offline effect. Baese-Berke & Goldrick's VOT enhancement effect mirrors that found offline by Goldinger & Summers (1989), providing partial support of the hypothesis. However, to our knowledge researchers have yet to investigate online analogues of other offline effects, including vowel space expansion and vowel nasalization.

Furthermore, if offline and online phonetic variation are driven by the same processes of cognitive competition, we would expect the generality of offline effects to be found in online enhancement as well. Just as offline VOT enhancement does not seem to be modulated by the specific phonological structure of a word's lexical neighbors, online VOT enhancement should not be affected by the sound structure of the words that have become salient in the speech discourse. That is, any type of neighbor that is active in the speech context should induce online enhancement. Baese-Berke & Goldrick (2009) investigated only contextually salient neighbors of one kind, namely those differing in the voicing of the initial consonant, and consequently the results of that study cannot determine whether online competition is general or specific.

We examined this issue in three experiments, and found that online phonetic enhancements differ from offline effects in two significant respects. Most importantly, online effects appear to be sensitive to the phonological properties of words in the local discourse. Only competitors that have particular phonological relations with the target word — relations defined by word position and segmental makeup — induce online hyperarticulation. We show that these results are predicted if speakers expend the effort involved in phonetic enhancement only when that could contribute to listeners' recognition accuracy, assuming a Bayesian model of word recognition.

Experiments

All experiments used an experimental paradigm adapted from Baese-Berke & Goldrick (2009). The goal behind the paradigm is to simulate a situation where a speaker must accurately communicate a word to a listener even though contextually salient competitor words provide opportunities for miscommunication. The paradigm involves two participants, one playing the role of speaker and the other the role of listener. Each participant sits at a separate computer terminal, which is not visible to the other participant. In each trial of the experiment, two or more words appear on both screens: a target word along with competitor words that are sometimes neighbors of the target. After approximately 1000ms, the target word becomes highlighted on the speaker's screen, who then produces it aloud. At this point, the listener clicks the word that was heard — the same word produced by the speaker, if communication is successful. The speaker's pronunciation of the target is recorded and analyzed acoustically after the experiment. The experimental setup is illustrated in Figure 1. This paradigm has the advantage of being able to precisely control a target word's "context" (the neighbors that appear on-screen with it) and including motivation for the speakers to communicate clearly, as they are made aware if the listener fails to select the target word.

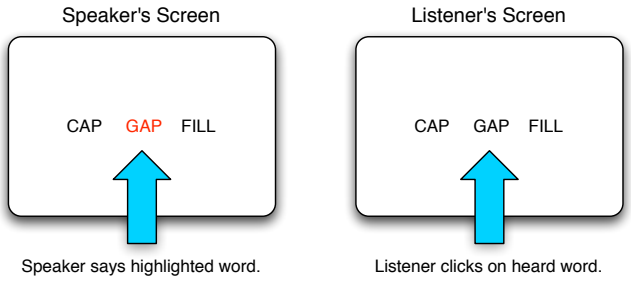

Figure 1: Experimental paradigm.

Experiment 1: Online Vowel Space Expansion

If online effects have the same underlying cause as offline effects, we expect to find an online vowel expansion effect mirroring the offline effect found (e.g., Munson & Solomon 2004). To test this hypothesis, we presented target words in the context of neighbors that differed from the target only in the vowel position. A condition where the same targets were presented with unrelated filler words was used as a baseline for comparison.

Table 1: Table of conditions for Experiment 1.

Target	Vowel	Filler
CAT	KIT	DOLL

Following previous work on vowel space dispersion, the dependent acoustic variable measured was the Euclidean distance of each target vowel from the center of each subject's vowel space (defined as the subject's mean F1 and F2 formant values). Participants (N=18) produced each of 16 target words in each condition. Target position on screen was counterbalanced across speakers. Order of conditions for a given target was also counterbalanced to avoid confounds with repetition effects. Results were analyzed using linear mixed results regression (*lme4*) in R, with condition as a fixed effect and subject and target item as random effects. Results are summarized in Table 2; p-values were obtained using Markov Chain Monte Carlo (*pMCMC*). There was no significant effect of onscreen neighbor on vowel space dispersion in comparison to onscreen filler words, suggesting that there is

no online analogue to offline vowel space dispersion effects. This provides evidence that online and offline effects do not share the same underlying cause.

Table 2: Experiment 1: Statistical results.

Condition	Coeff	SE	t	p
Vowel Neighbor	1.707	10.696	0.160	< 0.8732

Experiment 2: Positional Specificity of Online VOT Enhancement

Baese-Berke & Goldrick (2009) found VOT lengthening in the initial segment of a target word presented with an on-screen neighbor differing in the voicing of its initial segment. This experiment tested to see if any kind of neighbor can induce this enhancement effect, as might be expected if online effects lack specificity in the way that offline effects do. Target words were presented in the context of neighbors that differed only in onset (a replication of the Baese-Berke & Goldrick study using voice-differing neighbors), vowel, or coda positions. Different neighbor types were matched for frequency (pairwise paired t-tests, all $p > 0.3$).

Table 3: Table of conditions for Experiment 2.

Target	Onset	Vowel	Coda	Filler
CAP	GAP	CUP	CAT	DOLL

Participants (N=24) produced each of 48 target words twice in one of the four conditions, so that each word appeared in all conditions every four subjects. Results are shown in Figure 2 and Table 4. Only onset-differing neighbors appear to cause a significant VOT enhancement effect over fillers. This result suggests that online enhancement effects depend at least on position-level sublexical processing and are thus more *specific* than offline effects.

Table 4: Experiment 2: Statistical results.

Condition	Coeff	SE	t	p
Onset Neighbor	2.07000	0.80555	2.570	$< 0.0102*$
Vowel Neighbor	0.03449	0.80555	0.043	< 0.9659
Coda Neighbor	0.54367	0.80555	0.675	< 0.4998

Interestingly, the effects found seem to be limited to the first production of each target word. Second productions show no VOT difference across conditions, suggesting a strong effect of repetition in this experiment. Furthermore, as shown in Figure 3, the effects found are limited to cases when the target word begins with /p/ or /t/. This may be due to a ceiling effect associated with the /k/-initial targets used in the experiment, as /k/-initial words are known to have long base VOTs that participants may find it difficult to lengthen further.

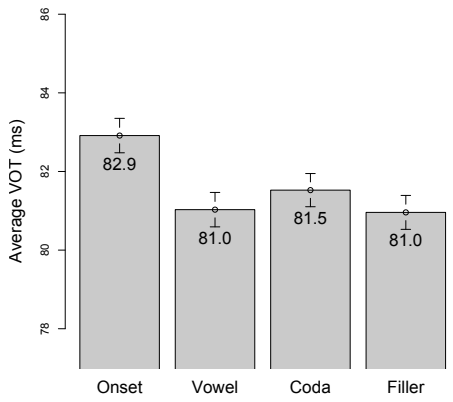

Figure 2: Experiment 2: Comparison of mean VOT across experimental conditions.

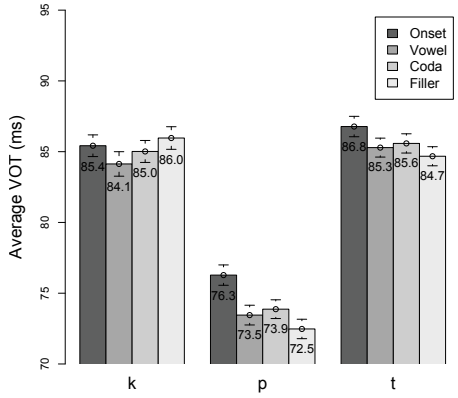

Figure 3: Experiment 2: VOT broken down by target onset phoneme and condition.

Experiment 3: Featural Specificity of Online VOT Enhancement

The goal of this follow-up experiment, consisting of two subexperiments, was to determine if online VOT enhancement involves an even lower level of sublexical processing. In particular, we tested to see if only certain kinds of onset neighbors can induce VOT enhancement. In the first subexperiment, we looked for an enhancement effect in the context of place-differing neighbors. Different neighbor types were matched for frequency (pairwise paired t-test, $p > 0.8$).

Table 5: Table of conditions for Experiment 3A.

Target	Voice	Place	Filler
CAP	GAP	TAP	DOLL

Participants (N=22) produced each of 33 target words

twice in one of the three conditions. Results are shown in Figures 4 and 5 and Table 6. There is a significant VOT enhancement effect of place neighbors, and the effect is consistent across /p/, /t/, and /k/-initial targets.

Table 6: Experiment 3A: Statistical results.

Condition	Coeff	SE	t	p
Voice Neighbor	2.1361	0.9329	2.290	$< 0.0222*$
Place Neighbor	1.8506	0.9333	1.983	$< 0.0476*$

It is interesting that the VOT of /p/ lengthens in the context of /k/ and /t/, given that /k/ and /t/ tend to have longer average VOT that /p/, and thus VOT lengthening might make /p/ initial words more similar to their competitors. However, aspiration also contains spectral cues for place of articulation (e.g., labial /p/ vs. coronal /t/ or dorsal /k/; (Suchato & Punyabukkana, 2005)), and lengthening VOT may strengthen these cues .

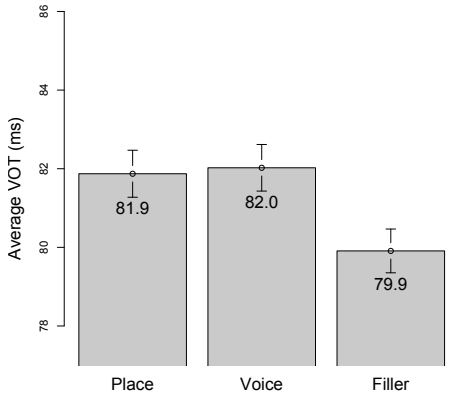

Figure 4: Experiment 3A: Comparison of mean VOT across experimental conditions.

In the second subexperiment we looked for an effect of neighbors differing in the manner of the onset. We attempted to choose neighbors that differed minimally from the targets with respect to manner, but were constrained by the phoneme inventory of English. The /p/-initial targets were paired with /f/-initial neighbors, which differ in manner and a minor place feature (labial vs. labiodental); /t/-initial targets were paired with /s/-initial neighbors, which differ in manner and stridency; and /k/-initial neighbors were paired with /h/-initial neighbors, which differ in both manner and place. Different neighbor types were matched for frequency (pairwise paired t-test, $p > 0.8$).

Participants (N=22) produced each of 36 target words twice in one of the three counterbalanced conditions. Results are shown in Figures 6 and 7 and Table 8. There appears to be no overall significant effect of manner neighbors on VOT enhancement. However, the breakdown of the results

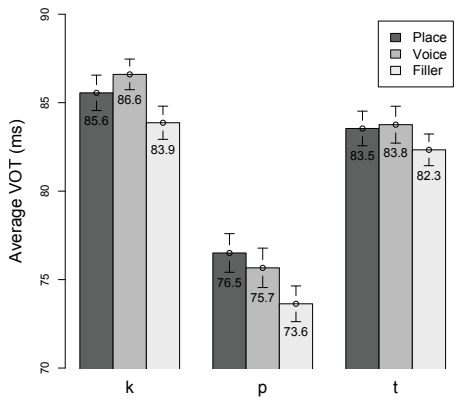

Figure 5: Experiment 3A: VOT broken down by target onset phoneme and condition.

Table 7: Table of conditions for Experiment 3B.

Target	Voice	Manner	Filler
PUN	BUN	FUN	DOLL
KILT	GUILT	HILT	DOLL
TEEM	DEEM	SEEM	DOLL

by target onset (Figure 7) indicates that there is an enhancement effect for /p/ onsets in the context of /f/ initial neighbors ($p < 0.0225$). Since /p/ is likely more similar to /f/ than /k/ is to /h/ (differing in a major place feature) or /t/ is to /s/ (differing in stridency)[1], it may be that online VOT enhancement may only occur in the context of neighbors that are sufficiently similar to the target word — about one major phonological feature away.[2]

Table 8: Experiment 3B: Statistical results.

Condition	Coeff	SE	t	p
Voice Neighbor	3.2236	0.9293	3.469	< 0.0005
Manner Neighbor	1.4489	0.9293	1.559	< 0.1192

Explaining Online Variation as Listener-Orientation: Modeling Speech Perception

It has been hypothesized that language is designed to facilitate effective communication between speakers and listeners

[1] Although /p/ and /f/ tend to pattern together as a natural class more often than /k/ and /h/ or /t/ and /s/, the effects found might not be due to their apparent similarity; instead, they may be a property of /p/-initial targets. It is difficult to disentangle this question using English stimuli, since another stop/fricative pair as similar as /p/ and /f/ does not exist in the phoneme inventory.

[2] All of the place neighbors in experiment 3A differ from the target in just one place feature.

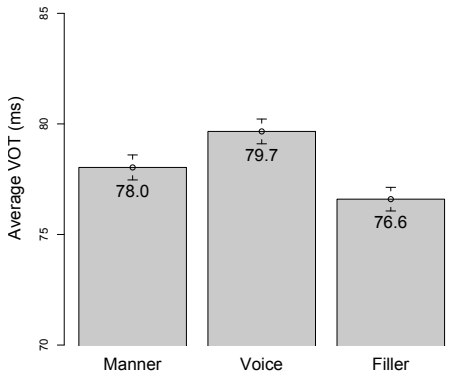

Figure 6: Experiment 3B: Comparison of mean VOT across experimental conditions.

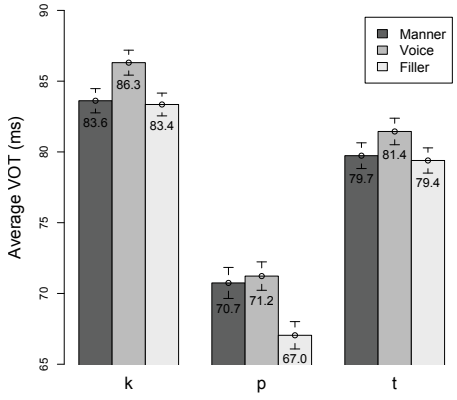

Figure 7: Experiment 3B: VOT broken down by target onset phoneme and condition.

(e.g., Lindblom 1990, Genzel & Charniak 2002, Levy 2006, Jaeger 2010, Frank 2008). This suggests that speakers may alter their speech in an effort to aid the recognition accuracy and speed of listeners.

This section presents a preliminary computational model of speech perception intended to help explain the specificity of online hyperarticulatory effects by allowing a comparison of the efficacy of different speech modifications in aiding listeners. The model is an extension of Norris's Shortlist B (Norris & McQueen, 2008). It assumes that word recognition is a Bayesian process (Norris & McQueen, 2008; Feldman, Morgan, & Griffiths, 2009) that modifies a posterior distribution over possible input words as an acoustic signal unfolds. The word with maximum posterior probability after a certain amount of input is recognized. The ratio of the recognized word's probability to that of its competitors determines how robust the match is — how likely it is to remain error-free given noisier input. It is this ratio of posterior probabilities that expresses the concept of competition between alternatives in the model.

The posterior probability of each possible input word (W_t) is equivalent to the likelihood that the word generated the signal seen so far (S), multiplied by the prior probability of the word, divided by the total probability of the signal being generated by any word:

$$P(W_t|S) = \frac{P(S|W_t)P(W_t)}{\sum_i P(S|W_i)P(W_i)}$$

The likelihood function $P(S|W_t)$ is equal to the likelihood that some prefix of the sequence of phonemes that make up the word (PP_t) generated the signal seen so far:

$$P(S|W_t) = \sum_i P(S|PP_{ti})$$

Unlike Shortlist B, the present model does not assume a fixed amount of time per phoneme (see also Scharenborg, 2009). $P(S|PP_t)$ can be broken down into a sum over possible segmentations (SS) of the signal into phonemes.

$$P(S|PP_t) = \sum_i P(SS_i)$$

The likelihood that a particular phoneme generated a portion of the signal in a segmentation is a function specified for the model, and is intended to be empirically realistic (e.g., the likelihood that a voiced phoneme like /g/ generated a large amount of of aspiration is low).

As an example, the model can be used to simulate the results found in Experiment 2. It receives input incrementally in 5ms frames, each containing one feature: C for closure, A for aspiration, and V for vowel (e.g. *gap* would be represented as [CAAAV...C]). In addition, the model only needs to distinguish between a target word and its on-screen competitor, resulting in a collapsed prior distribution over words (e.g. $P(cap) = P(gap) = 0.5$). Simulation using this model indicate that when the target word is *cap* and it is pronounced in the context of *gap*, the ratio of posterior probabilities monotonically improves in its favor as VOT increases.

Overall, the model has certain formal properties that make it suitable for explaining the experimental results presented. First, the posterior odds in favor of a target word can't be improved by hyperarticulating those parts of the word that are identical to its competitors. Doing this would equally increase the likelihood that the signal was generated by the target and it's competitor. Second, the improvement in posterior odds gained by hyperarticulating the differing parts of the target is minimal if the target and its competitor are not sufficiently similar to each other, since in that case the likelihood that the competitor generated the signal, $P(\text{signal}|\text{competitor})$, remains near zero throughout the recognition process, and consequently so does the competitor's posterior probability. In other words, if the target and competitor are different enough it would take a signal that is

very unlikely to have been generated by the target for it to be even slightly likely to have been generated by the competitor. Together, these two properties predict the specificity of online effects found experimentally. Speakers only seem to hyperarticulate when there is sufficient utility gained from the extra effort (Lindblom, 1990).

Conclusions and Future Research

In summary, the experiments presented here indicate that offline and online effects of phonetic enhancement may *not* share the same underlying mechanisms. Not all offline effects appear to have online analogues, as evidenced by the apparent lack of a significant online vowel space expansion effect. In addition, online effects appear to be more *specific* than offline effects. Online enhancement can be caused only by neighbors in the speech context that are minimally different from the target word (differing by approximately one phonological feature). These findings are compatible with a system in which speakers attempt to aid listener comprehension, with a Bayesian model of word recognition indicating which speech changes are helpful and which aren't.

However, the latter finding opens the possibility that vowels may indeed be subject to online hyperarticulation in principle, but that Experiment 1 did not include competitors that were similar enough to induce this effect. In particular, the vowel neighbors used in the experiment were not controlled to be minimally different from the vowels in the target words (in terms of backness and height features). Experiments 2 and 3 suggest that minimal difference is essential for inducing online effects, and future experiments will explore whether online vowel enhancement can be induced by minimally-different vowel neighbors.

References

Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relashionships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, *47*(1), 31-56.

Cohen-Priva, U., & Jurafsky, D. (2008, April). Phone information content influences phone duration. In *Conference on prosody and language processing.*

Dell, G. S., & Gordon, J. K. (2003). Neighbors in the lexicon: Friends or foes? In N. O. Schiller & A. S. Meyer (Eds.), *Phonetics and phonology in language comprehension and production: Differences and similarities* (p. 9-39). Mouton, New York.

Feldman, N. H., Morgan, J. L., & Griffiths, T. L. (2009). The influence of categories on perception: Explainng the perceptual magnet effect as optimal statistical inference. *Psychological Review*, *116*(4), 752-782.

Frank, A. F., & Jaeger, T. F. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. In *CogSci* (Vol. 30, p. 939-944).

Genzel, D., & Charniak, E. (2002). Entropy rate constancy in text. In *Proceedings of the 40th annual meeting of the association for computational linguistics (ACL)* (p. 199-206). Philadelphia.

Goldinger, S., & Summers, W. V. (1989). Lexical neighborhoods in speech production: A first report. In *Research on Speech Percetion Progress Report* (p. 331-342). Bloomington.

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, *61*, 23-62.

Lau, P. (2008). *The Lombard Effect as a communicative phenomenon* (Tech. Rep.). UC Berkeley.

Levy, R., & Jaeger, T. F. (2006). Speakers optimize information density through syntactic reduction. In *NIPS* (Vol. 19, p. 849-856).

Lindblom, B. (1990). Explaining phonetic variation: a sketch of the H&H theory. In W. Hardcastle & A. Maschal (Eds.), *Speech Production and Speech Modeling* (p. 403-439). Kluwer Academic Publishers.

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

Marslen-Wilson, W., & Zwitserlood, P. (1989). Accessing spoken words: The importance of word onsets. *Journal of Experimental Psychology: Human Perception and Performance*, *15*, 576-585.

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

Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review*, *115*(2), 357-395.

Scarborough, R. A. (2004). *Coarticulation structure and the lexicon*. Unpublished doctoral dissertation, UCLA.

Scharenborg, O. (2009). Using durational cues in a computational model of spoken-word recognition. In *INTERSPEECH* (p. 1675-1678).

Son, R. van, & Pols, L. C. (2003). How efficient is speech? In *ICPhS* (Vol. 25, p. 171-184).

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

Wright, R. (2003). Factors of lexical competition in vowel articulation. In J. Local, R. Ogden, & R. Temple (Eds.), *Phonetic Interpretation: Papers in Laboratory Phonology VI* (p. 75-87). Cambridge University Press.

Zhao, Y., & Jurafsky, D. (2009). The effect of lexical frequency and Lombard Reflex on tone hyperarticulation. *Journal of Phonetics*, *37*(2), 231-247.