Bayesian Speech Production: Evidence from Latency and hyper-articulation

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

Both response latency and phonetic variation reflect competition among alternatives during the speech production process. A review of the literature finds an apparent contradiction in the latency results. In some tasks where latency is measured, similarity between targets and competitors results in slower reaction times. In other tasks, similar competitors appear to facilitate production times relative to non-similar competitors (though a lack of any competition at all results in the shortest response latencies). With respect to phonetic realization, experiments suggest that high levels of competition induced by sufficiently similar competitors result in hyper-articulation of target utterances. We present a Bayesian model of speech production that formalizes the selection and planning of spoken forms as noisy-channel communication among different levels of processing. The model resolves the apparent contradiction found in the latency results, and establishes a novel connection between those results and observed patterns of hyper-articulation.

Keywords: Speech production; competition; 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). A number of studies have examined how such competitive processes are reflected in the time it takes to plan speech, and in the fine-grained phonetic realization of speech sounds. The goal of this paper is to develop a unified explanation of these potentially conflicting results, which have typically been treated independently.

In various speech production tasks, response latency is affected by the relationship between the target response and any primes, distractors, or competitors in the experimental speech environment (e.g., masked priming (Ferrand et al., 1996), plan switching (Meyer & Gordon, 1985; Yaniv et al., 1990), cue distractor tasks (Gordon & Meyer, 1984; Galantucci et al., 2009; Roon, 2012). A review of these results reveals an apparent contradiction with respect to how similarity between targets and competitors affects response latency.

In some production tasks, similarity between target utterances and competitors results in *delayed* (longer) response latencies (Meyer & Gordon, 1985; Yaniv, Meyer, Gordon, Huff, & Sevald, 1990; Roelofs, 1999). One example is the plan-switching task (Meyer & Gordon, 1985), in which participants are prompted to plan to say one form (e.g., the syllable **UP**), but are sometimes cued to say an alternative (e.g., the syllable **UB**). The findings from this task are summarized in Table 1. When the alternative response is highly similar to the originally planned target response, the time to initiate the alternative is lengthened. This effect drops off rapidly with increased phonological/phonetic distance. Only alternative responses that are about one feature away from the target seem to induce a significant delay.

Table 1: Plan Switching Task: Similarity = Higher Latency

Planned	Alternative	Difference	Latency
UP	UB	**voicing**	high
UP	UT	**place**	high
UP	UD	**voicing + place**	low

In cue-distractor tasks, on the other hand, similarity seems to play the opposite role (Gordon & Meyer, 1984; Galantucci et al., 2009; Roon, 2012). In a cue-distractor task, participants are taught to associate a visual cue with a particular verbal response (e.g., the syllable **KA** or **GA**). Upon receiving the cue, the participant attempts to produce the associated response as quickly as possible. However, before the subject is able to initiate speech (e.g., at 200ms after the cue), an auditory or visual distractor is presented (e.g., the syllable **PA**).

In spite of the fact that the subject has been given instructions to ignore the distractor, it has an effect on response latency as summarized in Table 2. It seems that when the distractor is sufficiently similar to the target response, production is facilitated relative to the case when the distractor is at a greater distance. However, it is always the case that the presentation of a distractor, no matter how it is related to the target, results in some production delay relative to the no-distractor case.

Table 2: Cue-Distractor Task: Similarity = Lower Latency

Response	Distractor	Difference	Latency
KA	none	**NA**	minimal
KA	GA	**voicing**	low
KA	TA	**place**	low
KA	DA	**voicing+place**	high

Finally, high levels of competition have been shown to influence phonetic realization: salient competitors in the speech environment give rise to hyper-articulation of spoken forms.

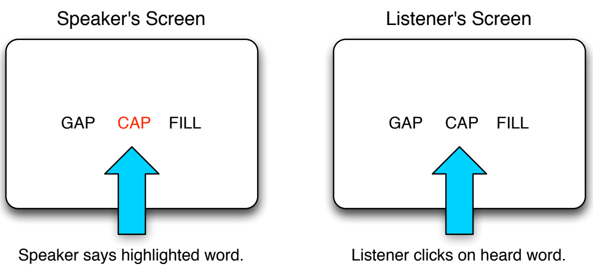
Speaker's Screen
GAP
CAP
FILL
Speaker says highlighted word.
Listener's Screen
GAP
CAP
FILL
Listener clicks on heard word.

Figure 1: Experimental paradigm.

For example, Buz & Jaeger (2012) found that word duration in a corpus of running speech is negatively correlated with distance to the nearest previously mentioned neighbor: neighbors mentioned in the recent past, against which the current word must plausibly compete, condition longer phonetic realizations. Baese-Berke & Goldrick (2009), using the same paradigm as our own experiments reviewed below, found VOT lengthening for voiceless-initial target words in the context of voiced-initial neighbors (e.g., the word CAP in the context of the word GAP).[1]

The Baese-Berke & Goldrick (2009) paradigm is designed to simulate a situation in which a speaker must accurately communicate a target word to a listener in the presence of contextually-salient competitor words that could delay recognition or cause 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 phonological neighbors of the target. After approximately 1s, the target word is 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 receive feedback indicating whether the listener has selected the intended word.

Using the same paradigm, we performed a battery of experiments (see Kirov & Wilson (2012) for details) to determine in what ways competitors could differ from the target utterance and still induce VOT hyper-articulation. The results, summarized in Table 3, point to the following generalization. Competition induces hyper-articulation only when competitors are sufficiently similar to the target word (a difference of roughly one phonological feature). The effect drops off quickly as phonological and/or phonetic distance increases.

This nonlinear relationship between feature distance and effect size mirrors the pattern found in the plan-switching task described earlier, suggesting that both effects might be linked through a common mechanism. Although, to our knowledge, no published experiment has directly attempted to correlate response latency with VOT hyper-articulation, there is some additional evidence that latency and hyper-articulation are linked. Bell et al. (2009) suggest that lexical access latency and utterance duration are correlated. Munson (personal communication) has also found that latency in a picture naming task is a good predictor of overall vowel-space expansion: longer latencies are associated with greater vowel expansion, which is a well-known type of hyper-articulation that can be conditioned by lexical competition (e.g., Wright, 2003; Munson & Solomon, 2004).

Table 3: Summary of hyper-articulation Results

Target	Competitor	Difference	Effect
CAP	DOLL	**unrelated**	X
CAP	CAD	**coda**	X
CAP	CUP	**vowel**	X
TAP	NAP	**onset voicing + nasality**	X
CAP	TAP	**onset place**	✓
CAP	GAP	**onset voicing**	✓

In this paper, we present a Bayesian model of speech production that resolves the apparent contradiction present in the latency data, and links the latency results to the hyper-articulation results, explaining why these effects share the same rapid drop-off as feature distance increases between competitors in speech production. We are not aware of previous work that has attempted to unify this body of results. Indeed, Roon (2012) has recently suggested that since the plan-switching task and cue-distractor tasks show different effects of similarity they must engage different levels of representation/processing. However, ascribing the effects to different processing levels would not make it clear why both tasks are sensitive to distance in the same phonetic/phonological space, and most importantly would not explain why some effects of similarity are inhibitory and others facilitative.

The proposed model posits that selection and planning of spoken forms involves optimal communication over noisy channels that link levels of mental processing/representation. Like well-known models of perception and recognition, our model takes Bayesian belief updating to be a fundamental component of psychological algorithms. This is in the spirit of other recent attempts to explore the mechanistic, as opposed to computational, utility of the Bayesian formalism (e.g., Sanborn et al., 2010).

[1] Voice onset time (VOT) is defined as the time between the release of a stop consonant and the start of vowel phonation.

Bayesian Word Production Introduction

Bayesian models have been productively applied to many aspects of perception (e.g., Knill & Richards, 1996; Girshick et al., 2011), including speech perception (Feldman, Morgan, & Griffiths, 2009; Norris & McQueen, 2008) and written word recognition (Norris & McQueen, 2008). In perception-oriented modeling, the mental system interprets noisy signals gained by the senses, updating internal beliefs about the external state as more and more evidence accumulates.

In the Bayesian word production model developed here, shown schematically in Figure 2, the signals of interest originate and are processed wholly within the mental system. Instead of interpreting noisy signals from the external world, the levels of processing/representation studied here interpret noisy messages from other levels. Each level maintains a probability distribution over representational states, receives noisy messages from one or more other levels indicating which state it should take adopt, and in turn sends noisy messages to other levels.

As is standard in Bayesian models of perception, we take noise to be an ineluctable feature of any communication system: noise is present in a signal regardless of whether that signal originates externally (from the environment, or the senses) or internally (from another mental level). One of the simplest approaches to successful transmission over a noisy channel is to use a *repetition code*. Repeated sampling in perception can result in a more accurate representation of the external world. For the same reason, repeated transmission of the same message to a level of processing can lead it to adopt a more functionally-appropriate representational state.

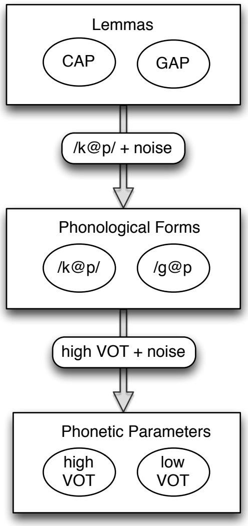

Figure 2: Bayesian Word Production Schematic

To further clarify the model, we will explain how the message passing process works, using the link between the lemma and phonology levels as an example. The construction of a message is shown schematically in Figure 3. Each possible lemma can send a characteristic message consisting of a phonological feature vector. The simulations reported here used phonologically realistic feature representations, but for reasons of space we show only part of each vector in the figure. In the construction of a message, first one lemma is sampled from the lemma distribution. The characteristic message of that lemma is then corrupted by noise and passed to the phonology level.

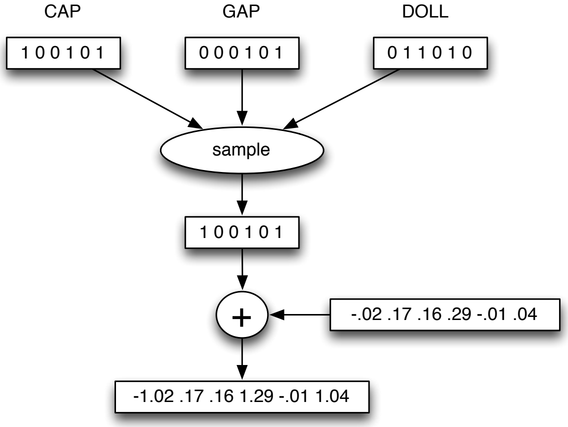

Figure 3: Message Construction

The receipt of the message by the phonology level, and the way in which the message is used to update the distribution over word forms at that level, is shown in Figure 4. Each phonological form represented by this level expects a particular message. The difference between this expected message and the message received is passed through a likelihood function to determine the probability that the message received corresponds to that particular form ($p(\text{message}|\text{form})$). The form of the likelihood function is determined by the type of noise that corrupts the message; in the simulations reported here we assume that noise in the word production system has a Gaussian distribution, but we have found that other types of noise are equally compatible with the experimental results (e.g., random flipping of binary feature values). Using the likelihood value, and the prior probability of each form, the level's probability distribution is updated according to Bayes' Rule:

$$p(\text{form}|\text{message}) \propto p(\text{message}|\text{form})p(\text{form})$$

When simulating word production using the model, a phonological form is chosen for production when it passes a high threshold probability. In the simulations reported here, the threshold is set to 0.95, which means that a form can be chosen only when it has 95% (or more) of the total probability after an instance of the Bayesian belief update. In most situations, a single message will not provide sufficient evidence for any form to reach this threshold after a single update. The necessary level of evidence is accumulated through multiple messages over time. This temporal repetition code for communication among mental levels will lead to accurate word

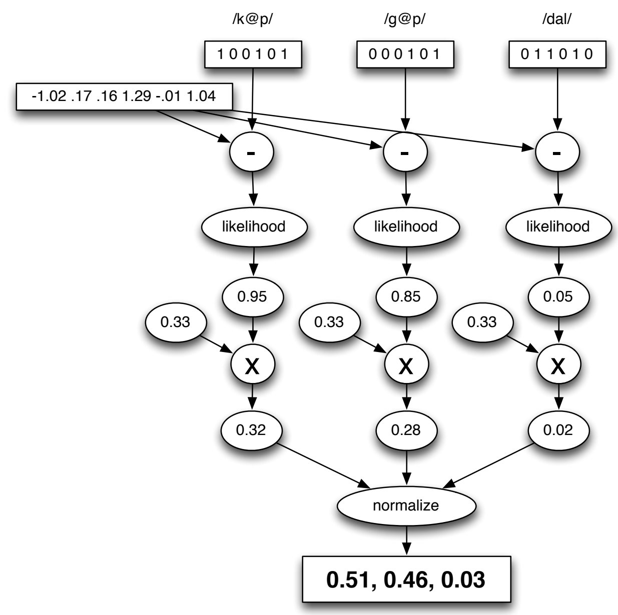

Figure 4: Bayesian Belief Updating

form selection with high probability, and lends itself well to accounting for latency and other effects observed in production.

Resolving Latency Contradictions

We will demonstrate how the model can resolve the apparent contradiction in the latency data with the help of characteristic examples. We begin with the plan switching task, where similarity between target and competitor appears to have an inhibitory effect. There are two relevant conditions. In the first case, shown in Table 4, the participant must plan to say a target utterance (the syllable **UP**), but is given a cue to say a different but similar alternative (the syllable **UB**) instead. Initially, the distribution of forms at the phonology level favors the target utterance. After the cue, this level begins to receive messages favoring the alternative. Since the target and the alternative are very similar, the likelihood function favors both of them, and the posterior distribution after each message is received is only slightly different from the prior distribution. Thus, it takes many messages (i.e., higher latency) for the alternative to reach the threshold probability required for production.

Table 4: Similar Alternative - Plan **UP** with potential alternative **UB**: Each message causes small posterior change.

	UP	**UB**
1) Initial state	0.75	0.25
2) **UB** message likelihoods	0.85	0.95
3) Updated state	0.73	.27

Table 5 shows the case when alternative response (**UD**) is substantially different from the target (**UP**). Once again, the initial distribution at the phonology level favors the target. This time, however, the likelihood function responds differently to the messages received after the response cue. Since the target and alternative are substantially different, the likelihood favors the alternative but not the target. As a result, the posterior distribution after each message is received is more significantly shifted. Since the posterior distribution experiences a larger change with each incoming message, it takes many fewer messages — hence less time — for the alternative response to reach threshold probability.

Table 5: Non-similar Alternative - Plan **UP** with potential alternative **UD**: Each message causes large posterior change.

	UP	**UD**
1) Initial state	0.75	0.25
2) **UD** message likelihoods	0.25	0.95
3) Updated state	0.44	.56

Overall then, latency is higher when the alternative response is more similar to the target, since both the alternative and the target are favored by the likelihood (i.e., there is evidence to produce both forms).

Next, we consider the cue-distractor task, where it seems a similar distractor has a facilitatory effect, relative to a different distractor. Again, there are two relevant conditions. In both cases, we will follow the setup in Roon (2012): depending on a response cue, the participant must say either **KA** or **GA**. We will assume that the **KA** cue is given, and some time has passed so that the distribution at the phonology level has shifted in favor of **KA**. In the first case, shown in Table 6, some time after the response cue the participant is presented with a distractor (**PA**) similar to the target, and a few messages corresponding to the distractor are sent to the phonology level. Since the distractor is similar to the target and different from its competitors, the likelihood function provides high evidence for the target and low evidence for any competitors, resulting in a favorable shift in posterior distribution.[2] Note that if the message received corresponded to the target exactly and not just a similar distractor, the target likelihood would be even higher, and the distribution would shift more favorably. Hence, latency is lowest when there is no distractor.

In the second case, shown in Table 7, the distractor presented after the cue (**BA**) is substantially different from the target, but similar to the alternative response. The distractor messages now provide low evidence for the target and high evidence for its competitors, causing the posterior distribution to shift in the wrong direction. Correcting this shift requires collecting more evidence for the target, resulting in greater latency.

[2] It is typical of the cue-distractor task that the distractor is not itself a valid output, and so effectively has zero prior and posterior probability.

Table 6: Similar distractor - **PA**: Distractor message provides more evidence for target than competitors.

	KA	GA
1) Initial state	0.75	0.25
2) **PA** message likelihoods	0.85	0.25
3) Updated state	0.91	0.09

Table 7: Non-similar distractor - **BA**: Distractor message provides more evidence for competitors than target.

	KA	GA
1) Initial state	0.75	0.25
2) **BA** message likelihoods	0.25	0.85
3) Updated state	0.47	0.53

In sum, a non-similar distractor causes a larger delay than a similar distractor because it provides strong evidence for the target's competitors and creates a shift in posterior probability towards them which must be overcome. There are certain situations where this generalization will not hold. In particular, if the non-similar distractor is also very different from all competitors (e.g., if the target is very similar to all possible alternatives), then it may create a smaller posterior shift towards the competitors than a similar distractor. Such situations have not arisen in the cue-distractor experiments to date, and so remain a novel prediction of the model.

Overall, we see that if speech production is a Bayesian process as proposed in this paper, the apparent contradiction found in the latency literature is resolved. In the plan-switching task, similarity is inhibitory because messages for the correct form also support the originally planned competitor form. In the cue-distractor task, similarity facilitates responses because messages from the distractor are transient and favor the correct form more than any competitors.

Linking Latency and hyper-articulation

As shown in Figure 2, the phonology level in the model can be linked to a phonetics level that maintains a distribution over prototypical phonetic realizations. Formally, the channel between phonology and phonetics works identically to the channel between lemmas and phonology, or any other pair of connected levels. The phonology level sends messages to the phonetics level indicating which phonetic realization is preferred, and the phonetic level updates its distribution according to Bayes' rule.

The message passing between phonology and phonetics stops when a decision about which form to produce is made at the phonology level (i.e., some form achieves threshold probability). At this point, the phonetic realization of that form can be extracted as a deterministic function of the posterior distribution in the phonetic level.

Figure 5 shows the results of a series of simulations that varied the distance between the target utterance and its closest competitor in the salient-competitor paradigm of Baese-Berke & Goldrick (2009). As feature distance increases, there is a rapid drop-off in both the time it takes for the phonology level to settle on the target form and the value of the phonetic parameter associated with the form. This pattern arises with a variety of model parameterizations with respect to noise and likelihood functions.

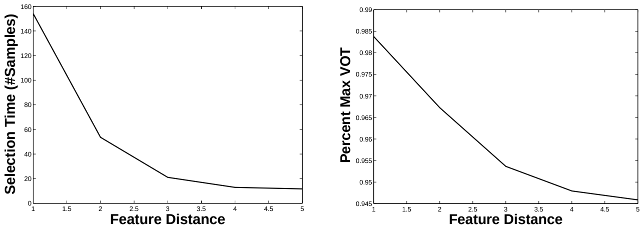

Figure 5: Simulation Results: Selection time and VOT hyper-articulation as distance between target and competitor varied from 1 to 5 features.

As previously shown for the plan-switching experiment, decisions at the phonology level take longer when competitors are very similar to the target. These longer planning times allow more messages to be sent from the phonology level to the phonetics level, so the latter will ultimately be presented with a greater amount of evidence for the max VOT prototype. This results in a more skewed posterior distribution and ultimately a longer VOT value.

The crucial result is that the modeling approach presented here predicts that hyper-articulation is a mechanical consequence of planning latency, and the two are closely correlated. This would explain why both types of effects show a similarly rapid drop as feature distance between competitors increases.

General Discussion and Future Research

We have presented a Bayesian model of word production that resolves an apparent contradiction found in latency-centered word production studies, and links latency results with results describing hyper-articulation. The fundamental claim of the model is that the selection and preparation of spoken forms should be formalized as Bayesian communication among levels of the speech production system. The model occupies a unique place in the overall space of production models, having distinct advantages and avenues for further development.

Most modeling based on interactive activation (e.g., Dell & Gordon, 2003) has not attempted to explain latency data, focusing instead on what errors the model makes after running for a predetermined amount of time. While the Bayesian model presented in this paper can be pushed to make errors by increasing the level of noise, it is left to future research to determine if the error distribution predicted conflicts with the available empirical data.

Some models, including Roon's dynamic field theory (DFT)-based production model (2012) and Roelofs'

WEAVER++ (1997), have addressed latency results, but have not been simultaneously used to explain hyper-articulation results. In addition, the extant models do not appear to address the full range of latency effects, focusing only on those cases, such as the cue-distractor task, in which similarity between targets and competitors appears to facilitate production. It remains to be seen whether cases in which similarity has an inhibitory effect, including the results from plan-switching tasks, can be accounted for by the models in their present form.

Up to this point, we have focused on modeling empirical data where the speaker's potential utterances were limited to a small closed set. Many studies deal with situations where any word in the lexicon is a potential output utterance. These studies typically examine the effects of global lexical factors such as frequency and neighborhood density on word production. Words with low lexical frequency and high neighborhood density tend to be produced with an expanded vowel space and longer VOT (Munson & Solomon, 2004; Wright, 2003; Goldinger & Summers, 1989).

An important question to pursue with respect to our model (or any model) is whether or not it can scale up to explain these results. One convenient feature of the Bayesian model is that the likelihood calculation performed when updating the distribution in a level quickly rules out competitors that differ significantly from the target utterance (i.e., their likelihood is close to 0). This means that selection among a large open set of potential outputs quickly begins to resemble selection between a small closed set of the type used in our experiments and simulations.

References

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

Bell, A., Brenier, J. M., 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.

Buz, E., & Jaeger, T. F. (2012). Effects of phonological confusability on speech duration (poster). In *The 23th CUNY sentence processing conference.*

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.

Ferrand, L., Segui, J., & Grainger, J. (1996). Masked priming of word and picture naming: The role of syllabic units. *Journal of Memory and Language*, *35*(708-723).

Galantucci, B., Fowler, C. A., & Goldstein, L. M. (2009). Perceptuomotor compatibility effects in speech. *Attention, Perception, & Psychophysics*, *71*(1), 1138-1149.

Girshick, A. R., Landy, M. S., & Simoncelli, E. P. (2011). Cardinal rules: Visual orientation perception reflects knowledge of environmental statistics. *Nature Neuroscience*, *14*(7), 926-934.

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.

Gordon, P. C., & Meyer, D. E. (1984). Perceptual-motor processing of phonetic features in speech. *Journal of Experimental Psychology: Human Perception and Performance*, *10*(2), 153-178.

Kirov, C., & Wilson, C. (2012). The specificity of online variation in speech production. In *Proceedings of the 34th annual meeting of the cognitive science society.* Sapporo, Japan.

Knill, D., & Richards, W. (1996). *Perception as bayesian inference*. Cambridge, UK: Cambridge University Press.

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.

Meyer, D. E., & Gordon, P. C. (1985). Speech production: Motor programming of phonetic features. *Journal of Memory and Language*, *24*, 3-26.

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.

Roelofs, A. (1997). The WEAVER model of word-form encoding in speech production. *Cognition*, *64*, 249-284.

Roelofs, A. (1999). Phonological segments and features as planning units in speech production. *Language and Cognitive Processes*, *14*, 173-200.

Roon, K. (2012). *The dynamics of phonological planning*. Unpublished doctoral dissertation, New York University.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, *117*(4), 1144-1167.

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.

Yaniv, I., Meyer, D. E., Gordon, P. C., Huff, C. A., & Sevald, C. A. (1990). Vowel similarity, connectionist models, and syllable structure in motor programming of speech. *Journal of Memory and Language*, *29*(1), 1-26.