# AN IDEAL COMPENSATOR MODEL OF SPEECH PERCEPTION

By

STEN KRISTIAN KNUTSEN

A thesis submitted to the

School of Graduate Studies

Rutgers, the State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Master of Science

Graduate Program in Psychology

Written under the direction of

Dave Kleinschmidt

And approved by

_____

_____

_____

New Brunswick, New Jersey

January, 2021

ABSTRACT OF THE THESIS

An ideal compensator model of speech perception

By STEN KRISTIAN KNUTSEN

Thesis Director:

Dave Kleinschmidt

One of the key issues in speech perception is how listeners are able to accurately categorize linguistic units (e.g., phonemes) from acoustic cues that contain variation due to multiple overlapping layers of information (Liberman et al., 1967). Over the years, researchers have developed various compensation procedures (e.g., vowel formant normalization) that strive to overcome this variation and increase classification accuracy. Although computationally efficient and widely used, these compensation procedures fall short conceptually as i) they are not necessarily computational models of compensation/ perception/cognition and ii) they do not allow inferences regarding classification to interact dynamically with inferences regarding compensation. In this work we outline a bayesian computational framework for speech perception and compensation, the *ideal compensator*. Because our listener model *infers how to compensate* based on a speaker's *generative model* while also *simultaneously* inferring linguistic category, we believe our approach is novel as it both increases classification accuracy and addresses the conceptual issues ignored by previous compensation models and procedures.

**Table of Contents**

# Introduction

One of the key issues in speech perception is how listeners are able to accurately infer or categorize discrete linguistic units (e.g. phonemes, syllables, words) from an acoustic signal. Categorization of linguistic units is computationally challenging because individual acoustic cues (e.g. voice onset time, segmental duration) are the product of multiple causes, both linguistic (e.g. which phoneme is intended by the talker) and non-linguistic (e.g., who is doing the speaking). As such, each cue contains multiple overlapping layers of information that have been distributed across the utterance by the talker (Liberman et al., 1967).

From a listener's perspective, much of the information encoded in a cue may be relevant to the perceptual task at hand. For instance, English vowel segment duration provides information about a vowel's tenseness or laxness and thus information useful in the task of determining vowel identity (Hillenbrand et al., 1995). However, acoustic cues are also jam-packed with information that could be useful in other perceptual tasks. For example, besides providing information about tenseness/laxness, vowel segment duration also provides information about: the number of syllables in the word that contains the vowel (Lehiste, 1972); how many times a word has been mentioned in discourse (e.g. second mention reduction; Fowler & Housum, 1987); type of speech the talker was using (e.g. plain speech or listener-directed speech; Picheny, Durlach, & Braida, 1986); the sex and regional origin of the talker (Hillenbrand et al., 1995; Jacewicz, Fox, & Salmons, 2007); talker ethnicity (Holt, Jacewicz, & Fox, 2015); the age of the talker (e.g., adult or child speech; Kim & Stoel-Gammon, 2010); whether the segment is from a

function (grammatical) word (Umeda, 1975); word frequency (Wright, 1979); the contextual predictability of the word (Bell et al., 2002); speaking rate (Kessinger & Blumstein, 1998) which varies considerably across talkers (Tsao & Weismer, 1997) and within talkers' utterances (Crystal & House, 1990); whether the word is a noun or verb (Sorensen, Cooper, & Paccia, 1978); distal prosody and lexical competition (Brown et al., 2011); whether a segment lies near a phrase boundary (Turk & Shattuck-Hufnagel, 2007); prosodic prominence (Aylett & Turk 2004; Turk, 2010); the syntactic structure of a sentence (Beach, 1991; Stromswold et al., 2002).

Unfortunately, the fact that acoustic speech cues are saturated with all sorts of information about all sorts of causes means that a considerable amount of variability is introduced into the acoustic signal. Because of this variability, information relevant to the current task is obscured or becomes "smudged" by information relevant to other tasks and contexts, making accurate speech perception difficult. The goal of this work is to describe a computational model of *how* a listener's speech perception system might *compensate*[1] for the variability inherent to information-rich cues while *simultaneously* inferring linguistic categories as intended by the talker.

## Of cues and compensation

Since we are trying to describe a computational model of speech perception, it is natural to approach the problem from the standpoint of a listener attempting some perceptual

[1] Throughout this paper we will use the terms *compensate* and *compensation* to broadly describe any process, procedure or framework used to overcome the lack of invariance inherent to the acoustic speech signal (cf. McMurray & Jongman, 2011)

task. From this listener perspective, a *cue* is any observed variable that might contain information relevant to the linguistic task at hand. For instance, if the linguistic task at hand is determining what vowel is being produced by a talker at this very moment, cues to the vowel's identity might include segmental duration and f1/f2 formant frequencies (Hillenbrand, Clark, & Houde, 2000; Delattre et al.,1952). However, which cues are used and *how* they are used by listeners depends entirely on the linguistic task at hand. A cue to some linguistic category in one task might also be a cue to a completely different linguistic category in another task.

For example, in the context of determining what vowel is being said by a talker, segmental duration is a cue to vowel tenseness or laxness in English: shorter durations correlate with lax vowels like /ɪ/ in *bit*, and longer durations with tense vowels like /i/ in *beat* (Hillenbrand et al., 1995). However, in the context of determining which word-final stop consonant is being said, segmental duration of the preceding vowel is a cue to whether that consonant is voiced or voiceless (Raphael, 1972; Hogan & Rozsypal,1980). On average, vowels preceding voiced stop consonants (e.g., the alveolar stop /d/ in *bid*) tend to be longer than vowels preceding voiceless ones (e.g., the alveolar stop /t/ in *bit*). In sum, given the task of inferring what vowel is being said, vowel duration is a cue to vowel tenseness/laxness, and in the context of determining the word-final stop, vowel duration is a cue to voicedness.

From a listener's perspective, this task or context-dependent nature of cues gives rise to a perceptual puzzle. For instance, suppose the listener's task at hand is to

determine whether they are hearing the English word *beat* or *bid* based on vowel segment duration alone.[2] On the one hand, the word *beat* has a tense vowel (/i/) that correlates with *longer-than-average* duration followed by a voiceless stop consonant (/t/) that correlates with *shorter-than-average* vowel duration. On the other hand, the word *bid* has a lax vowel (/ɪ/) that correlates with *shorter-than-average* duration, followed by a voiced stop (/d/) that correlates with *longer-than-average* vowel segment duration (cf. Mermelstein, 1978). As such, it's possible that these opposing influences on duration could cancel each other out, resulting in similar vowel segment durations for *beat* and *bid* (see illustration Figure 1). From a listener's perspective the key question is: *Is the duration of this vowel segment the result of a tense vowel and voiceless consonant, or a lax vowel and voiced consonant?*

To answer this question, the listener needs to somehow estimate what part of the vowel's duration provides information regarding the lax/tenseness of the vowel and what part provides information regarding the voicedness of the following word-final stop consonant. Moreover, the listener must be able to do this both in the context of identifying the vowel and in the context of determining the voicedness of the final stop. At a computational level, a listener must be able to somehow *classify* linguistic category (or, more broadly, make inferences relevant to the task at hand) while also *compensating* for the influence of hidden variables that are not relevant to the perceptual task at hand.

---

[2] This and the following examples are greatly simplified for clarity. For English listeners, F1/F2 formant frequencies are the primary cue to vowel identity in English words, while duration plays a secondary role (Ainsworth, 1972; Delattre et al., 1952; Hillenbrand, Clark, & Houde, 2000). However, non-English speakers may rely entirely on duration as a cue to vowel identity when listening to English words (Kondaurova & Francis, 2008)

However, the answer of *how* listeners accomplish this remains an open question. In this work we will describe a fully bayesian computational framework for listener compensation and classification - *the ideal compensator* - that will hopefully shed some light on this question.

Before we flesh out the details of this model framework, though, let's first briefly consider a common yet coarse-grained perspective on acoustic cues, how this perspective has given credence to a certain "style" of compensation model, and some of the conceptual drawbacks inherent to such models.

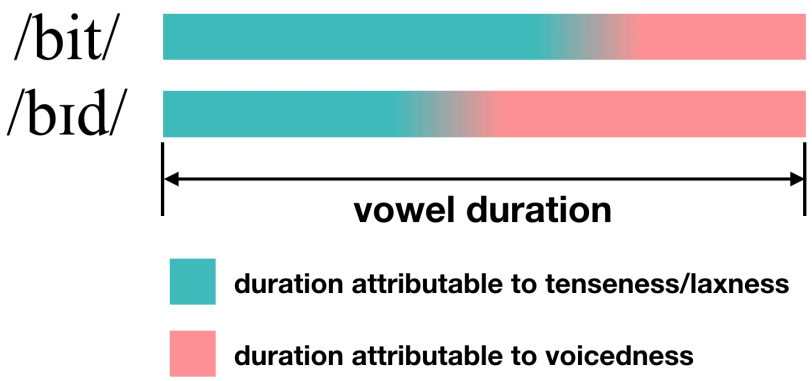


*Figure 1.* Illustration of vowel segment duration. *Is the duration of this vowel segment the result of a tense vowel and voiceless consonant, or a lax vowel and voiced consonant?*

## Breadcrumbs, red herrings and pipelines - *oh my!*

Many different approaches to compensation have been advanced in the speech perception literature. Much of the work focuses on vowel normalization (e.g., Lobanov, 1971; Syrdal & Gopal, 1986; Miller, 1989) though there has also been research on normalization of consonants such as fricatives (Strand & Johnson, 1996; Toda, 2007).

The procedures outlined in these studies have themselves been the focus of study over the past several decades (Disner, 1980; Adank, Smits, & Hout, 2004; Clopper, 2009; Fabricius, Watt, & Johnson, 2009; Flynn & Foulkes, 2011). As effective as these compensation procedures are, they are all implicitly predicated on an approach to acoustic cues that, in an era of bayesian cognitive modeling, seems to be underspecified. We have dubbed this approach the *breadcrumb and red herring* approach to cues.

In short, this perspective regards cues as containing a mixture of *breadcrumbs and red herrings*: information relevant to the perceptual task at hand is like a trail of informative *breadcrumbs*, while all irrelevant information is like a confusing jumble of *red herrings* that have been scattered about.

We can illustrate this breadcrumb/red herring approach further using the above *bit/bid/beat/bead* example. In the context of a listener trying to figure out which vowel is being said, vowel duration is like a *breadcrumb* in that it is part of a probabilistic trail of evidence left by the talker that contains information about the identity of the vowel itself (e.g., tense vowels like /i/ tend to have longer durations, on average). However, in this same context (which vowel is being spoken at present), vowel duration also behaves like a jumble of *red herrings* strewn across our path in that it contains information *irrelevant to the task at hand*. In this case, it contains information about the upcoming stop consonant's voicedness (e.g., voiceless consonants tend to be preceded by vowels with shorter durations, on average). This information, while vital in determining stop consonant voicedness, introduces variation to the tense/lax category distributions encoded

in the cue. This additional variation makes accurately inferring the correct linguistic category of the vowel more difficult, and can possibly lead the listener astray as they attempt to track the talker's linguistic intent. Thus, before a listener can infer the linguistic intent of the talker, they must "sweep away" or remove all red herrings from the path, leaving only the relevant breadcrumb trail to successful speech perception. This breadcrumb/red herring approach to cues has logically given rise to compensation procedures such as the ones mentioned above, which we will call information *pipeline* approaches to compensation.

Broadly speaking, pipeline approaches to compensation begin with raw cue values provided by the talker. Since these values contain unwanted variation due to causes that are irrelevant to the given perceptual task (red herrings), they are advanced down the pipeline to a preprocessing or compensation stage. In this compensation stage of the pipeline, the raw cue values are mathematically transformed by some compensation procedure, reducing variance due to irrelevant causes in the target cue category distributions. These compensated cue values then flow further down the information pipeline to a classifier where the linguistic intent of the speaker is finally inferred via an informative trail of "pure" breadcrumbs.

As straightforward and effective as pipeline approaches to compensation have proven themselves, there are several *conceptual* sticking points to pipeline procedures that bear our consideration.

The first aspect in which pipeline approaches to compensation fall short conceptually is in that they are not necessarily computational models of compensation/ perception. Regarding vowel normalization procedures, Adank, Smits, & Hout (2004) observed that although vowel-intrinsic procedures did strive to model human vowel perception, vowel-extrinsic procedures focused mostly on increasing classifier accuracy for automated speech recognition. Even when procedures *do* attempt to model human cognitive processes, the scope of such models is often limited to low-level, peripheral auditory systems which assume an 'intermediate' or pre-processing stage before category inferences are made (e.g, Syrdal & Gopal, 1986).

From the perspective of a researcher interested in cognitive models of compensation, pipeline models may very well *describe* compensation in terms of *what we would expect to see in a compensated or normalized data set* (if indeed human listeners do use compensation in speech perception). However, pipeline models do not at all address the probabilistic processes that *generated* speaker data in the first place. This means the question of *why* compensated data should look one way or another remains unanswered.

Another reason to have conceptual reservations about pipeline approaches is that by their staged nature they prohibit inferences regarding classification to interact dynamically with inferences regarding compensation. This means the listener is not allowed to sustain any uncertainty as to what portion of the cue constitutes "signal" and what portion constitutes "noise." However, as was illustrated by the *bit/bid/beat/bead*

example, phonemes that surround a target phoneme influence perceptual cues in such a way that inferences regarding the category of the target phoneme may be affected by inferences regarding the category of the adjacent phonemes (Sawusch & Pisoni 1974; Whalen, 1989). Such bidirectional, phonemic influences hint at a dynamic and probabilistic relationship between compensation and classification, something that has been little explored in compensation literature (cf. McMurray & Jongman, 2011).

Having now briefly considered the breadcrumb/red herring perspective on cues, existing pipeline-style compensation procedures, and the conceptual pitfalls inherited from both of these, let's now turn our attention to our proposed *ideal compensator* model framework. This framework is founded on principles established by *ideal listener* models of speech perception (Clayards et al., 2008). In short, such models seek to optimally infer linguistic categories given an acoustic cue. The following section introduces the ideal listener model and a perspective on acoustic cues that stands distinctly apart from the breadcrumb/red herring approach.

## Our base: the *ideal listener*

In contrast to pipeline-style compensation procedures, ideal listener models start with the assumption that the acoustic cues we use to interpret speech are *inherently ambiguous* and thus only *partially informative.* Since we are provided with only imperfect, probabilistic information regarding the linguistic categories encoded by an acoustic cue, it is important to first acknowledge that speech perception is a problem of statistically optimal *inference under uncertainty*. We can formalize this sort of inference in a bayesian

computational framework that integrates over *all* information available to a listener in a statistically optimal way.

The goal of the ideal listener model is to optimally infer the linguistic intent of the speaker or choose the message most likely intended by the speaker. This requires comparing multiple hypotheses about the speaker's intent and deciding which is most probable. Fortunately, Bayes rule provides us with means of deciding which hypothesis is most probable in a manner that is both precise and penetrable. We will now illustrate this with a highly simplified and idealized example.

Assume that our listener is using *vowel duration* as the sole cue to whether a word-final, English alveolar stop is either voiced (/d/) or voiceless (/t/) as in *bid* or *bit*. Given an observed vowel segment duration, the listener's goal is to determine how likely it is that the word-final alveolar stop is voiceless, or the probability that $C = /t/$ (where $C$ is the consonant voicedness category variable). In terms of bayesian statistics, this means the listener needs to find the posterior probability $p(C = /t/ \mid duration)$. This can be found using Bayes rule:

(1)

$$p(C = /t/ \mid duration) = \frac{p(duration \mid C = /t/)p(C = /t/)}{p(duration \mid C = /d/)p(C = /d/) + p(duration \mid C = /t/)p(C = /t/)}$$

Applying Bayes rule reveals several other crucial pieces of statistical information needed to calculate the posterior. First, the listener model must know the *likelihood function*, i.e., $p(duration \mid C = /t/)$, or the probability of a durational token given that it was the

talker's intent to produce a /t/. Likelihood functions for both /d/ and /t/ categories are represented by the illustrated probability density functions in Figure 2, Model A, top. Second, the listener model must know the *prior probability* of the hypothesis that the category of the consonant is voiced, i.e., $p(C = /t/)$. By multiplying the prior by likelihood in the numerator in (1), we update the listener's prior beliefs regarding consonant voicedness. Lastly, the posterior probability that the talker is saying /t/ is captured by dividing the combined posterior and likelihood for /t/ by the sum of *all* of the possible hypotheses (the denominator in (1)). This effectively normalizes our updated prior by the marginal likelihood, ensuring that *all possible posterior probabilities*, i.e., $p(C = /d/ \mid duration)$ and $p(C = /t/ \mid duration)$, add up to one. These normalized posterior values are reflected in the classification function for the /t/-/d/ voicing distinction (Figure 2, Model A, bottom).

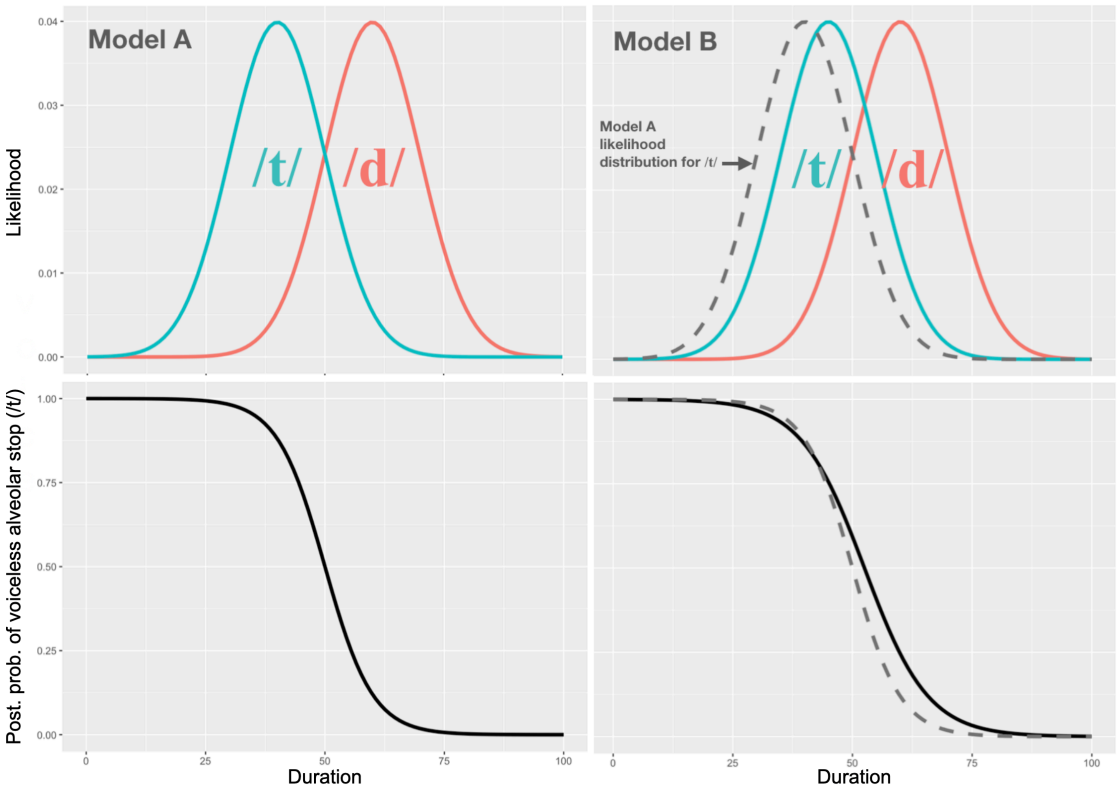


*Figure 2:* Illustrations of idealized /t/-/d/ category distributions for vowel segment durations. Model A top illustrates listener likelihood distributions for /t/-/d/ categories; bottom shows the listener's posterior classification function for Model A. Model B top shows "real-world" likelihood distributions in order to illustrate the "mismatch" between listener assumptions about likelihood functions and the "actual" functions in the real world; bottom shows classification function for Model B (solid line) in comparison to Model A (dashed line). Durational units are for illustrative purposes only.

As discussed above, the likelihood function is critical to obtaining optimal classification of linguistic categories. From a listener's perspective, the likelihood function makes predictions about what cue values (in this case, vowel segment duration) are likely to occur given the linguistic category intended by the talker (in this case, stop consonant voicing). However, these likelihood functions are subjective and may not perfectly represent the actual cue distributions in the real world. For example, a listener

may assume the likelihood distribution for /t/ (Figure 2, Model A) is the ideal likelihood function for that category. But suppose the *actual* likelihood distribution for /t/ in the "real world" is shifted to the right (see Figure 2, Model B, top). The mismatched likelihood distribution subsequently shifts the final classification function to the right, shifting the /t/-/d/ category boundary. The effect of this shift is that cue values that were previously ambiguous are now more a bit more likely to predict /t/ as the category. Conversely, if our listener selects the likelihood distributions in Model B in a situation where the actual likelihood distributions are those in Model A, cue values that often predicted the /t/ category would predict voicedness at chance levels (an ambiguous alveolar stop). Such mismatches in likelihood distributions lead to less accurate classification and poor overall comprehension.

Mismatches can also occur when sources of information or causal variables are not included in the listener model. For instance, perhaps the "shifted" /t/ distribution in the "real world" Model B (illustrated in Figure 2) is due to some other cause (linguistic or not) that has simply not been factored into the listener model. How might a listener model incorporate other talker and linguistic information in such a way that classification mismatches are less likely to occur?

## Compensation as model selection based on "fixed" context

In the above ideal listener example, we demonstrated how an ideal listener model can interpret acoustic cues despite the fact that these cues are inherently ambiguous and only partially informative. We also demonstrated that despite being charged with the

simplified task of determining the voicedness of a single, word-final consonant from vowel duration, mismatches between a listener's likelihood function and the actual real-world likelihood distributions can negatively impact perception.

In this next example, our listener's goal is still to determine whether a talker intended on producing a voiced or voiceless alveolar stop (/d/ or /t/, respectively) based on vowel segment duration. However, following from the *bit/beat/bid/bead* example, we want to infer the final stop's voicedness conditioned on *context*, i.e., whether the preceding vowel is tense (/i/) or lax (/ɪ/).

In terms of bayesian statistics, the listener needs to find the posterior probability of the target consonant's voicedness, not only conditioned on the observed duration, but also on *V*, or the vowel tenseness category variable. We can spell this out formally as $p(C \mid duration, V)$. Applying Bayes rule we arrive at:

(2)

$$p(C \mid duration, V) \propto p(duration \mid C, V)p(C)$$

where $p(duration \mid C, V)$ is the likelihood. Since our likelihood is now conditioned on two categorical variables having two levels each, there are now a total of four possible hypotheses:

(3)

$i)$ $p(duration \mid C = /t/, V = /I/)$ as in *bit*

$ii)$ $p(duration \mid C = /d/, V = /I/)$ as in *bid*

$iii)$ $p(duration \mid C = /t/, V = /i/)$ as in *beat*

$iv)\ \ p(duration \mid C = /d/, V = /i/)$ as in *bead*

The key question we now want to ask is: How might the listener model use contextual knowledge (about vowel tenseness/laxness) in determining the voicedness of the consonant?

For a moment, let's assume that the vowel preceding the consonant is tense (i.e., $V = /i/$). In this context, the model sheds all likelihoods where $V = /I/$ leaving only $p(duration \mid C, V = /i/)$ (Figure 3, top). This is reflected in the posterior $p(C \mid duration, V = /i/)$ (Figure 3, bottom) where the optimal /t/-/d/ classification function falls to the right of the 50 unit durational midpoint. As such, the range of cue values that correlate with a high posterior probability of /t/ is *larger* than the range of cue values that correlate with a low posterior probability of /t/ (or a high probability of /d/). This means that when operating in a "fixed" /i/ context, the listener will more often favor a /t/ (voiceless) interpretation of the alveolar stop.

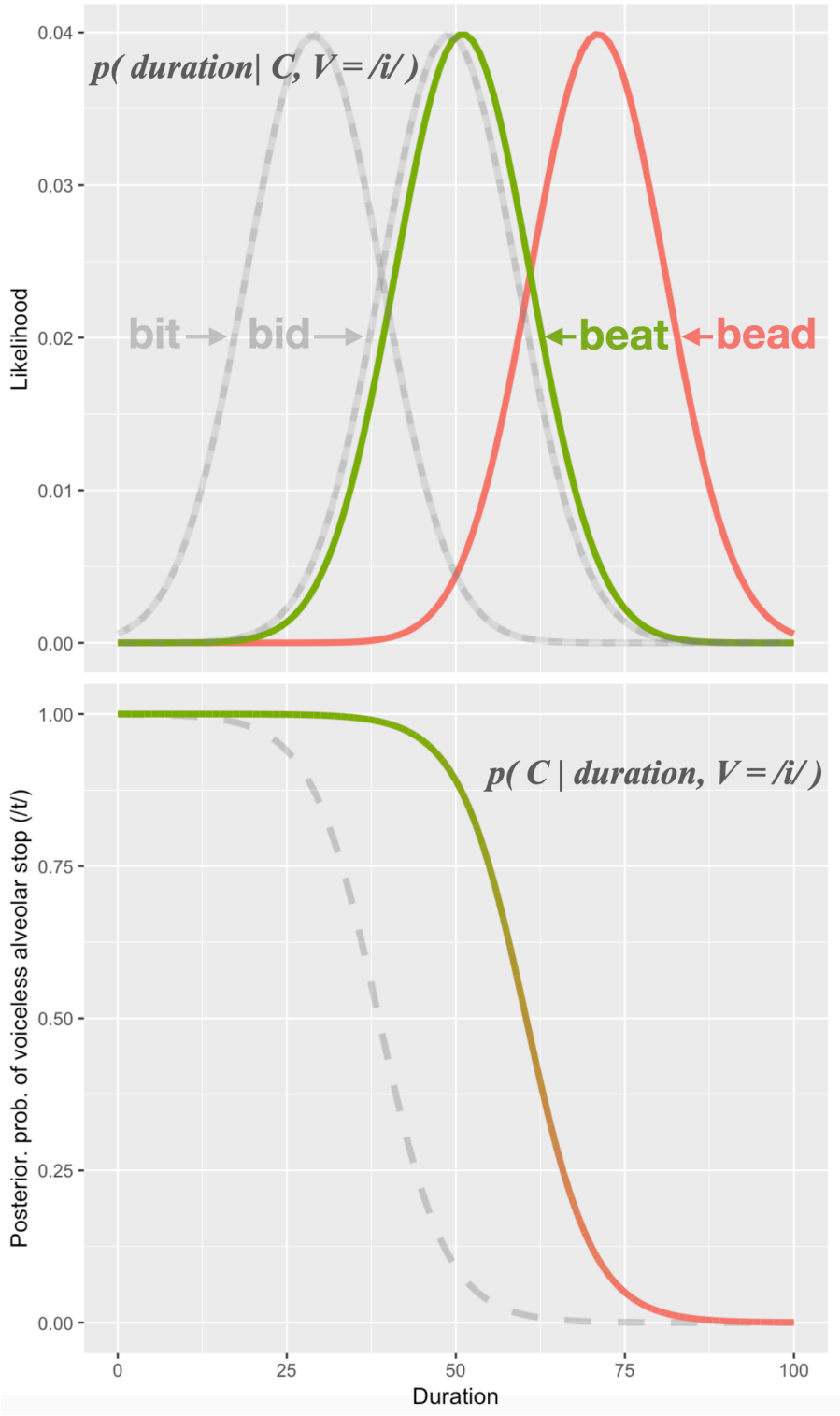


*Figure 3: T*he top panel illustration shows likelihood functions for when $V = /i/$. Shaded/ dashed lines show likelihood functions if $V = /I/$. The bottom panel shows both the classification function when $V = /i/$ and $V = /I/$ (shaded/dashed line).

Now let's assume the listener knows $V = /I/$, or that the vowel preceding the consonant is lax. In this context, all likelihood functions where $V = /i/$ are dropped, leaving only $p(duration \mid C, V = /I/)$. This is reflected in the posterior $p(C \mid duration, V = /I/)$ (Figure 4, bottom) where the optimal /t/-/d/ classification function falls to the left of the 50 unit durational midpoint. As such, the range of cue values that correlate with a high posterior probability of /t/ is *smaller* than the range of cue values that correlate with a low posterior probability of /t/ (or a high probability of /d/). This means that when operating in a "fixed" /ɪ/ context, the listener will more often favor a /d/ (voiced) interpretation of the alveolar stop.

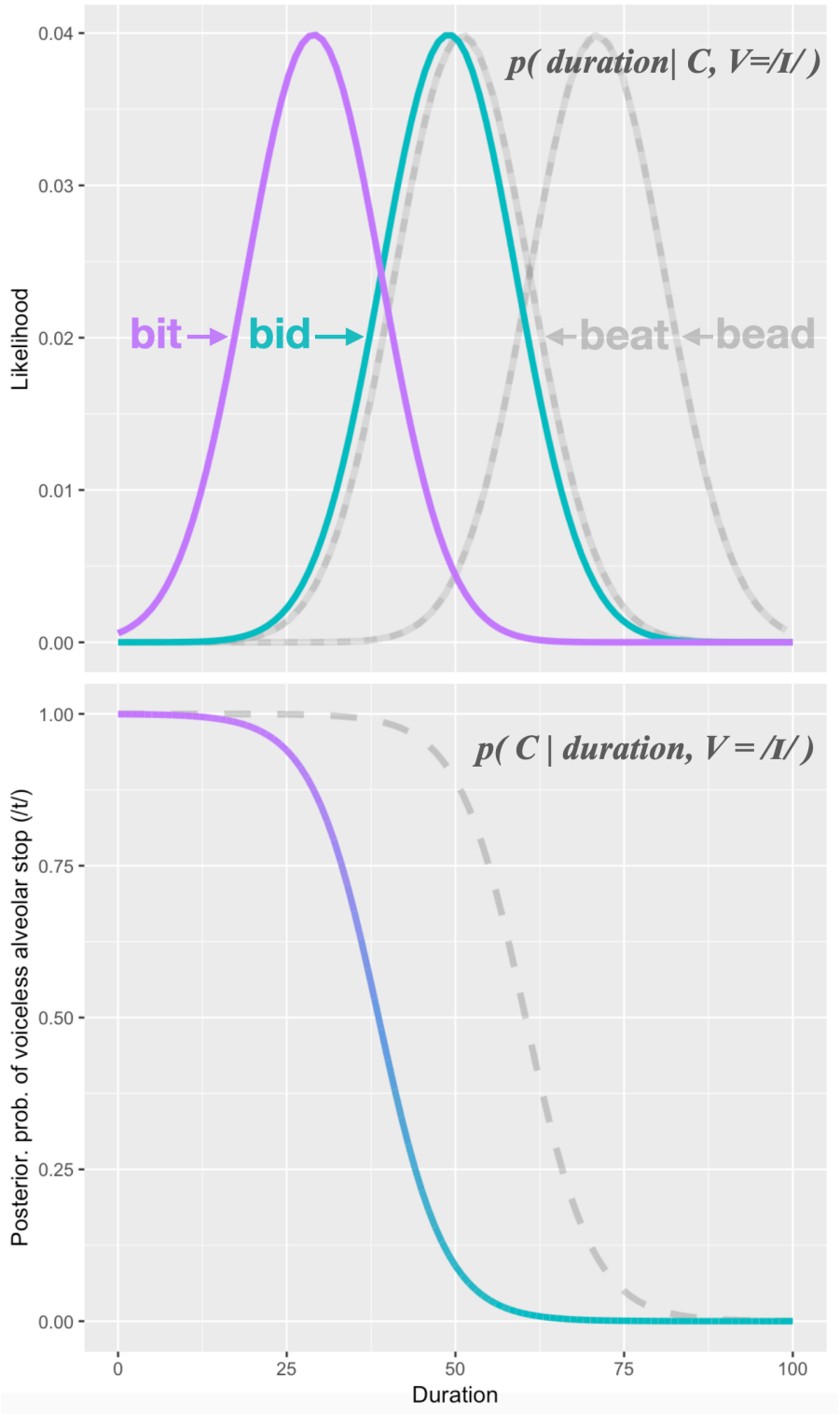


*Figure 4:* The top panel illustration shows likelihood functions for when $V = /I/$. Shaded/dashed lines show likelihood functions if $V = /i/$. The bottom panel shows both the classification function when $V = /I/$ and $V = /i/$ (shaded/dashed line).

Let's now circle back to the question of how the listener model uses contextual knowledge in determining the voicedness of the consonant. We now see that in an /i/ context, the model selects only the hypotheses compatible with $V = /i/$ (Figure 3, top) and in the /ɪ/ context, the model selects only the hypotheses compatible with $V = /I/$ (Figure 4, top). This carries through to the posterior classification function which is optimal given contextual information about $V$ (Figures 3 and 4, bottom). Thus the model effectively *compensates* for the influence of $V$ "for free" by selecting the optimal classification function for $C$ given the context: in the /i/ context, the classification function compensates by favoring a /t/ interpretation of the word-final alveolar stop; and in the /ɪ/ context, the classification function compensates by favoring a /d/ interpretation of the word-final alveolar stop. As a result of compensation for vowel tenseness, we encounter fewer "mismatches" between model-estimated and real-world likelihood values, increasing the model's classification accuracy.

The sort of compensation we have described thus far is analogous to the pipeline approaches to compensation discussed earlier. These approaches also assume listeners know the specific context – tense /i/ or lax /ɪ/ – in which the alveolar stop was produced.

In this example we used contextual knowledge to compensate for vowel tenseness/laxness while classifying consonant voicedness simply by handing the model the preceding context (whether $V = /i/$ or $V = /I/$) and "selecting" the correct posterior function:

i) If we know the vowel is /i/, then $p(C \mid duration, V = /i/)$

ii) If we know the vowel is /I/, then $p(C \mid duration, V = /I/)$

So in either case, when the true context is provided, the model *necessarily* selects the optimal likelihood functions for that context. From these likelihoods, the model derives the posterior classification function for the /t/-/d/ voicing distinction while compensating for vowel tenseness.

However, real-life listeners are never simply handed complete information about vowel tenseness. There is no speech perception oracle who taps the listener on the shoulder and whispers "the preceding vowel is tense" so that the listener discards all /t/-/d/ likelihood functions where $V = /I/$. Instead, listener inferences are likely to be based on partial, uncertain and incomplete beliefs about the talker's message. So how might a listener compensate for the influence of *V* while classifying *C* when they are uncertain whether the vowel is tense or lax?

**Compensation and marginalization**

As in our fixed context model, our listener's goal is still to determine whether a talker intended on producing a /t/ or /d/. This means the listener again needs to find the optimal posterior function for stop consonant voicedness given segmental duration. However, in order to incorporate *uncertain information* about the influence of vowel tenseness into the model, our inferential starting point must be $p(C, V, duration)$ or the *joint probability distribution* over all variables. Applying Bayes rule we find the *joint posterior* over all model variables:

(4)

$$p(C, V \mid duration) \propto p(duration \mid C, V)p(C)p(V)$$

Furthermore, we model the listener's uncertain beliefs about *V* via *marginalization* or summing over the possible values of *V* to determine the marginal contribution of *C*. By calculating the marginal posterior – or summing over variables that are *not* the target of classification, in this case *V* – we are effectively averaging together the all model hypotheses, weighted by the listeners degree of belief regarding the true identity of *V*. Mathematically, this can be expressed as:

(5)

$$p(C \mid duration) = p(C, V = /i/ \mid duration) + p(C, V = /I/ \mid duration)$$

Figure 5 illustrates how uncertain listener beliefs about *V* influence the marginal posterior and thus the posterior classification function for *C* for five different values of $p(V = /i/)$. The different values for $p(V = /i/)$ in Figure 5 correlate to the listener's *degree of belief* regarding vowel tenseness. In this case, when $p(V = /i/)$ is *high*, the listener is more confident the vowel is tense and so leans toward an /i/ interpretation; when $p(V = /i/)$ is *low*, the listener is less certain the vowel is tense and so leans toward a lax /ɪ/ interpretation.

Now let's consider the two extremes of listener belief regarding $p(V = /i/)$: when $p(V = /i/) = 0$ and when $p(V = /i/) = 1.0$. Note that when $p(V = /i/) = 0$, the posterior function (Figure 5 bottom) is exactly the same as the posterior function in the fixed context model, $p(C \mid duration, V = /I/)$ favoring a /d/ interpretation of the final

stop consonant (Figure 4, bottom). This is because the listener's degree of belief regarding $p(V = /i/)$ is zero and thus the only interpretation of the vowel can be /ɪ/. In contrast, when $p(V = /i/) = 1.0$, the posterior function is exactly the same as the posterior function in the fixed model $p(C \mid duration, V = /i/)$ favoring a /t/ interpretation of the final stop consonant (Figure 3, bottom). This is because the listener's degree of belief regarding $p(V = /i/)$ is 1.0 and thus the only interpretation of the vowel can be /i/. So setting $p(V = /i/) = 0$ is equivalent to the fixed /ɪ/ context in the previous model and $p(V = /i/) = 1.0$ is equivalent to the fixed /i/ context.

However, similarities between this "uncertain" model and the previous "fixed" model end when $p(V = /i/)$ lies between 0 and 1.0. Let's first focus on when $p(V = /i/) = 0.8$, comparing this to the posterior function for $p(V = /i/) = 1.0$ (Figure 5, bottom).

At first glance, the function for $p(V = /i/) = 0.8$ is quite similar to $p(V = /i/) = 1.0$ (Figure 5, bottom). However, when $p(V = /i/) = 0.8$, we notice that the range of cue values that would indicate a prototypical /t/ interpretation in the $p(V = /i/) = 1.0$ model (roughly between 25 and 50 durational units) now corresponds to posteriors that favor the /t/ interpretation less. In other words, what was once a region of cue values that would deliver a strong voiceless interpretation of the consonant is now a region of uncertainty where the listener now must "hedge their bets" regarding consonant voicedness. This is why when we visually follow the classification function for $p(V = /i/) = 0.8$ from left to right, it seems to "droop" between the two model extremes

of $p(V = /i/) = 0$ and $p(V = /i/) = 1.0$ – right where we would expect uncertainty regarding vowel tenseness to affect our consonant voicedness classification.

So when the model is just a bit less certain about the tenseness of the vowel – the $p(V = /i/) = 0.8$ condition – the model is also slightly less confident in the voicelessness of the consonant. If we were to instead select a model where $p(V = /i/) = 1.0$ when the true value is $p(V = /i/) = 0.8$ our listener would be "overconfident" about the voicelessness of the alveolar stop. Yes, when $p(V = /i/) = 0.8$, the model does still favor a /t/ interpretation of the consonant as does the model where $p(V = /i/) = 1.0$, but to a degree commensurate with listener beliefs about the tenseness of the preceding vowel.

In sum, when we use marginalization to calculate the optimal posterior classification function for consonant voicedness given uncertain listener beliefs about vowel tenseness, it can be said that the listener model compensates for the influence of $V$ while simultaneously classifying the voicedness category $C$.

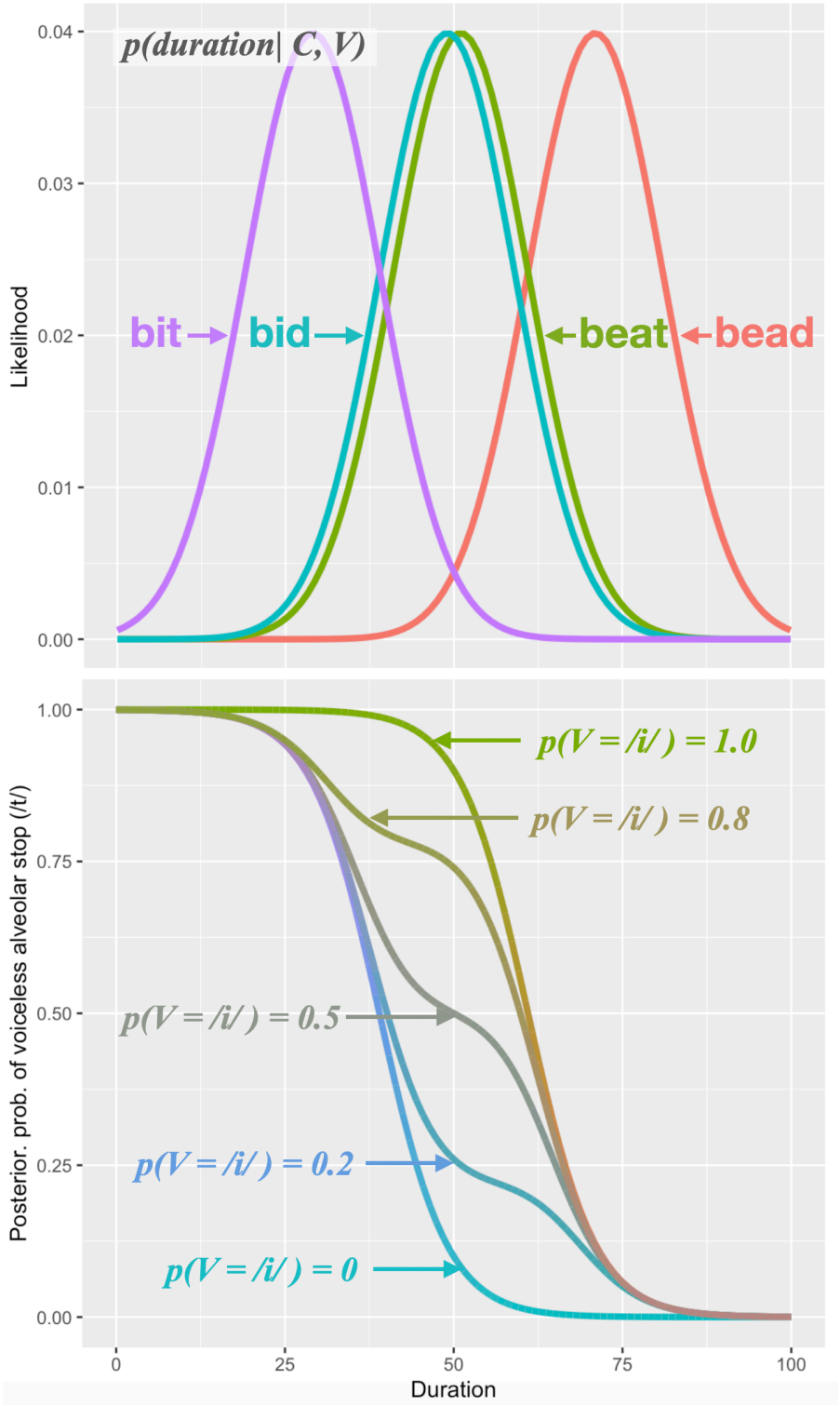


*Figure 5:* The top panel illustration shows likelihood functions for all combinations of consonant voicedness and vowel tenseness. The bottom panel illustrates posterior functions for multiple values of V (vowel tenseness).

# Full model and implementation

Thus far, we have sketched the foundation of our ideal compensator model on the *beat/ bit/bid/bead* example introduced at the outset. However, the model we have outlined thus far will easily generalize to *any* speech perception task where multiple non-/linguistic causes negatively impact the classification/categorization of an observed variable. To fully flesh out and implement our ideal compensator model, we now turn to a different speech perception task that also benefits greatly from compensation.

## Background

Previous research has shown that listeners are sensitive to subtle variations in segmental duration and that they adjust inferences regarding syntactic structure when duration is manipulated. In Beach (1991) listeners demonstrated the ability to use duration in an early part of a temporarily ambiguous sentence to predict upcoming sentence structure. The researchers constructed stimuli sentence pairs with identical beginnings that are resolved as either a direct object (DO) sentence or a sentence complement (SC) construction. For example, a stimuli sentence starting with *David's second wife claimed. . .* could be resolved as the DO sentence *. . .the entire family estate including the yacht* or the SC construction *. . . .[that] the entire family estate was rightfully hers.* Knowing that matrix verb stems in SC constructions feature durations that are longer than verb stem durations in DO constructions, the researchers electronically manipulated verb duration in the sentence onset to have longer or shorter duration. The sentences were then truncated before the disambiguating direct object or sentence complement structures, and presented to participants. Listeners were able to use durational patterns to identify

upcoming syntactic structure even without complete sentence information, classifying sentences as a direct object constructions at above-chance levels.

Stromswold et al. (2002) also explored how listeners might infer the syntactic outcome of a temporarily ambiguous sentence before explicit morphosyntactic cues (e.g. verbal inflection) are provided. However, rather than using direct object and sentence complement sentence constructions for stimuli, the researchers constructed active and passive sentence pairs with identical sentence onsets. For example, a stimuli sentence starting with *The girl was push. . .* could be resolved as the active sentence *-ing the boy* or the passive *-ed by the boy*. Using a 2AFC task in a visual world paradigm, listeners were asked to listen to the spoken stimuli sentence and identify which of two pictures on the screen best described the action being performed in the audio. An eye-tracking device captured listener eye movements. Eye-tracking data revealed that adult listeners started to determine the syntactic voice of a sentence at or before the verb stem in active sentences.

In a follow-up gating study, spoken active/passive sentences similar to the example above were truncated before the disambiguating verbal inflection (e.g., *The girl was push. . .*). Listeners were able to identify the correct syntactic outcome of the truncated stimuli with 83% accuracy, and a post-hoc analysis of verb stem duration revealed that passive verb stems were significantly longer than active ones (Stromswold, Kharkwal, and Sorkin, under review).

The data we used to evaluate our model were drawn from the set of spoken active/passive sentences recorded for another follow-up study to Stromswold et al. (2002) (Lai, 2015; see Appendix A for complete set of stimuli sentences used). Each of the eight, monolingual adult native American English speakers recorded for the study said 28 active/passive sentence pairs. All sentences follow either the form "The **NP1** was **VERB STEM** -ing the **NP2**" or "The **NP1** was **VERB STEM** -ed by the **NP2**." All sentence NPs are semantically reversible and appeared in both agent and patient positions in both active and passive syntactic forms. All sentences were initially segmented at the phonemic level using the Penn Phonetics Lab Forced Aligner (Yuan & Liberman, 2008). Boundaries for verb stem vowel segments were further adjusted by hand using Praat software (Boersma & Weenick, 2020) according the methods outlined in Francis, Ciocca and Yu (2003). Following segmentation, all 439 duration values were log transformed.

After segmenting the stimuli sentences, we focused on the verb stem vowel segment as it was the only segment to show a significant difference between active and passive distribution means. Visually inspecting the active/passive distributions for the verb stem vowel (see Figure 6) we see that despite finding a difference in means, there is a substantial amount of overlap in the active and passive distributions.

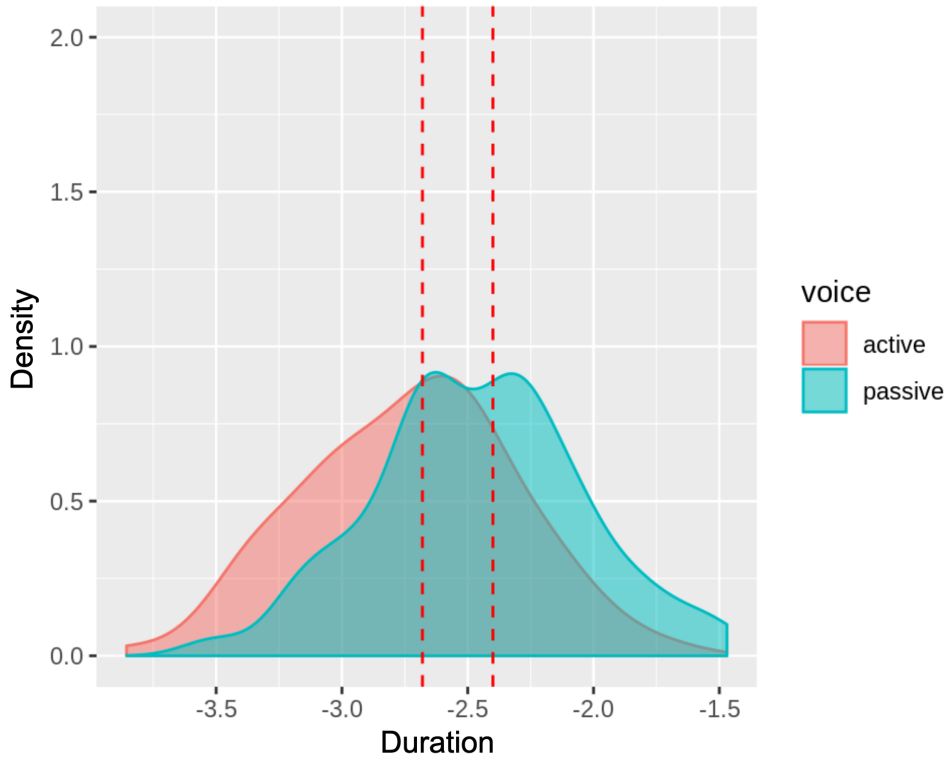


*Figure 6:* Active/passive distributions for verb stem vowel segment, all talkers and sentences. Note duration is in log scale.

The reason for this overlap is that active and passive categories are represented by distributions that contain variance due to multiple underlying causes. For our model implementation, we are focusing on just two such sources of variation: variation due to talker differences (*talker*) and category of the phoneme being uttered (*phoneme*). Thus, although means for passive/active distributions may be fixed, increasing variance in the distributions due to the influence of talker variation and phoneme category means an increase in overlap between distributions. In turn, such an overlap in distributions negatively impacts the accuracy with which verb stem vowel segment tokens are classified as active or passive. In fact, it seems unlikely that a listener would have much

success in trying to classify a sentence as active or passive using the distributions pictured in figure 2.[3]

**The *ideal compensator***

Our *ideal compensator* framework is founded on the assumption that our listener possesses a *generative model* of the talker they are listening to. The term *generative model* refers to a listener's knowledge of the processes with which a speaker generated the utterance heard. In this model implementation, we will be focusing on how the duration of a verb stem vowel segment is generated by the talker.

For example, if a talker wants to produce a duration for the verb stem vowel from the verb *kiss* in the sentence *The sheep was kissed by the pig*, they must choose the appropriate categories given the sentence they are trying to produce. In this instance, the talker is producing a segment where *syntax* is passive, and the *phoneme* is set to /IH/. This category information regarding *syntax* and *phoneme* is statistically encoded in segment durations produced by the talker.

Also encoded in this segment is indexical or *talker* information. This is non-linguistic information specific to the current talker, and could be anything from speaking style to regional dialect. In the end, a duration that reflects all of the above influences is selected and produced by the talker.

---

[3] Note that inference of syntactic voice may or may not be a direct one. It is possible that a listener may actually be inferring whether the verb is monosyllabic and therefore passive (e.g. "kicked") or polysyllabic and therefore active (e.g. "kicking"; cf. Rehrig, 2017)

It is important to note that the listener does not have complete or direct knowledge of the talker's generative model, but only uncertain beliefs about the model. They do not have any direct access to the category selections the talker has made. Speech sounds change depending on context and so the listener must allow some uncertainty regarding the statistical properties underlying speech sound categories. In fact, a successful listener must be open to integrating new observations with prior expectations so that the inferred generative model keeps pace with the changing statistics that have been encountered. Thus the listener infers the *underlying statistical properties* of an observed variable (in this case, duration) as opposed to inferring a single, fixed value.

Given an observed duration and the causal variables highlighted above – $talker, phoneme$ and $syntax$ – we can formally specify the generative model of a talker in the language of bayesian statistics as:

(6)

$$p(syntax, talker, phoneme, duration) = p(duration \mid talker, phoneme, syntax) \cdot p(talker) \cdot p(phoneme) \cdot p(syntax)$$

Since the generative model includes multiple underlying causes, listener inferences regarding a single, target cause (i.e., syntax) relevant to the perceptual task at hand (i.e., classifying active/passive) given a particular durational token cannot occur in an inferential vacuum; they must also account for all other causes (i.e., phoneme, talker). So that the listener might infer the targeted underlying causes (or classify active/passive syntax) we invert the generative model and use Bayes rule:

(7)

$$p(syntax \mid duration, phoneme, talker) \propto p(duration \mid syntax, phoneme, talker) \cdot p(syntax)$$

The left-hand side of (7) formally expresses the posterior probability of a segment being active or passive conditioned on listener knowledge of durations, phonemes and talkers. This posterior is proportional to the likelihood of a duration given syntax, phoneme and talker, multiplied by the prior for syntax (i.e., the prior probability of a sentence being active or passive).

Note that while the listener has no knowledge of whether a given duration has been drawn from an active or passive sentence, they do know who the talker is and what phoneme they are producing. However, knowing who is talking and what phoneme that talker has spoken does not mean the listener can be *absolutely certain* of what effect or influence that particular combination of talker/phoneme has on the observed duration. As mentioned earlier, listeners do not have complete or direct knowledge of a talker's generative model, but only uncertain beliefs about the model. As such, we want to explicitly represent this uncertainty about the current generative model in our formalization. More specifically, we want to capture the model's uncertainty about the relationship between talkers, phonemes, syntax and the observed stimuli. Thus we need to take into account the parameters of the model, $params$, or *the mean and variance of the likelihood for each combination of speaker, phoneme and syntax.* Formally, this can be expressed as:

(8)

$$p(params, syntax \mid duration, phoneme, talker) \propto p(duration \mid params, syntax, phoneme, talker) \cdot p(params) \cdot p(syntax)$$

Our inverted generative model now allows us to infer the underlying syntactic target cause and the statistical relationship between each combination of causes and the observed duration.

However, the addition of $params$ to the model raises two additional issues. First, if we considered the myriad of variables and categories that are brought to bear on segmental duration in everyday speech, the number of possible combinations our listener would need to consider would become intractable. In fact, the combinations would scale multiplicatively with every additional factor introduced. As far as this particular model is concerned, capturing the statistical relationship between every combination talker, phoneme, and active and passive voicing is not untractable since the variables and categories have been artificially constrained by our dataset labels. With 8 talkers, 7 phonemes and 2 syntactic voices we would expect to estimate means and variances for 112 likelihood distributions.

This brings us to the second issue we need to address concerning the addition of $params$ to our model: the limitations of a relatively small dataset. Even at only 112 combinations of talker, phoneme and voice, our statistical "pie" has to be "sliced" so many ways that with only 439 data points, there is simply not enough data to provide reasonable estimates of all likelihood means and variances. Thus it is reasonable to make some additional assumptions regarding how our model $params$ are structured.

## Model parameters and their arrangement

As was outlined in previous sections, compensation/classification is successful only when we compute or select likelihood functions that are specific to and fully reflect the particular set of causes that have given rise to the observed stimuli. In this case, we want to compute the likelihood function $p(duration \mid params, syntax, phoneme, talker)$ for all relevant categories (active/passive syntax) and combinations of causes (phoneme and talker). However, as described in the previous section, we must do so in a way that recognizes the limitations of inferring model parameters from a relatively small dataset and in a way that addresses the explosion of causal factor combinations inherent to such a model.

So instead of calculating likelihoods for every combination of causes "from scratch", we use a factorization approach to describe and additively combine our model parameters. In this approach, compensated active/passive likelihood distributions are constructed by *summing* the mean values of individual factors. This simplification treats the influence of each additional cause as *independent* of the others, but allows the model to consider more causes given a fixed amount of data. In the end, the composed likelihoods for each combination of causes assume a normal distribution and are uniquely defined by a mean and variance.

There are two chief components in our factorization. First, the model parameter `mu` governs the distance between the active/passive category means of $syntax$, centering active/passive distributions around zero. For clarity's sake, we will refer to mean active/passive durations defined by `mu` as a syntactic *base* duration to be shared across all

talkers. Second, the model parameter `offset` captures the statistical relationship between individual talkers and the phonemes that each of these talkers produce. We can refer to durations defined by the `offset` parameter as talker-phoneme *offsets*.[4] With these two model parameters now defined, we can now compose likelihood distributions over all model variables by summing the inferred base and offset durations.

How these model parameters are leveraged during model testing and training in our model implementation also bears mention. We implemented the ideal compensator model in Stan statistical software with the `rstan` interface (Stan Development Team, 2020; see Appendix B for Stan model code) and our dataset of 439 active/passive verb stem vowel segment tokens. For the training talkers, active and passive $syntax$ labels for all durational tokens were made available to the model in order to fix the syntactic "polarity" of the grand, bimodal active/passive distribution. For testing talkers, $talker$ and $phoneme$ category labels are known while active/passive $syntax$ labels remain unknown. The fully implemented model jointly and simultaneously infers model base and offset parameters, classifying active/passive syntax while compensating for talker and phoneme.

## Model variants and comparisons

For the sake of comparison, we implemented two additional models in Stan: a *non-compensated* pipeline-style version of the model that only infers syntactic voice; and a

---

[4] The labels *offset* and *base* highlight the fact that there are two separate parameter components that are additively combined in the model. The labels offset and base themselves are completely arbitrary.

*pipeline*-style compensation version of our model. In the pipeline model, verb stem vowel segment tokens are first centered by talker and phoneme and then fed into the Stan implementation for classification (without inferred compensation). We also carried out four types of model comparison to evaluate the ideal compensator model:

***Mean classifier accuracy.*** For a simple and direct comparison and evaluation of the models' performance we use mean accuracy. To determine classification accuracy, we thresholded posteriors for each model at $p = 0.5$ and calculate mean accuracy using a winner-takes-all approach.

We fully expect that both ideal compensator and pipeline models will see an *improvement* in classification accuracy over the non-compensated model variant. However, we have no reason to expect that the ideal compensator model will necessarily *outperform* the pipeline model, as this was never the goal of the model in the first place. Again, the ideal compensator model addresses *conceptual* issues that arise from pipeline models, not *performance* issues.

***Model posteriors.*** The output of all three Stan model implementations are sampled posterior values. Paralleling our expectations for mean accuracy, we also anticipate that ideal compensator and pipeline models will see an *increase* in mean posterior probability of the true category over the non-compensated model. Again, the ideal model does not need to outperform the pipeline model in this respect.

***Variance reduction.*** One of the main goals of pipeline compensation schemes is to reduce variance in the relevant category distributions of the dataset to increase classifier

accuracy. So as far as pipeline models are concerned, we can evaluate their effectiveness by comparing compensated data distributions with non-compensated or "raw" data distributions. Comparing pipeline and non-compensated distributions alone, we expect to see a *decrease* in category distribution variance.

However, this comparison is only possible – or relevant, even – due to the inherent nature of pipeline models. As discussed earlier, pipeline models assume that raw cue values contain unwanted variation that must be removed before classification. Thus they invoke a preprocessing stage where some sort of compensation scheme is employed (e.g., centering). Once this preprocessing stage is complete, the compensated data can be evaluated as to whether there has been a reduction of variance in category distributions in comparison the raw, unprocessed values.

In contrast, the ideal compensator model does not have a separate, preprocessing stage for compensation. This is because, as outlined earlier, the model dynamically and simultaneously infers linguistic category and how to compensate for talker and phoneme. Both conceptually and implementationally speaking, there is nowhere we can "stop" the model to locate a set of preprocessed "compensated" data points. Doing so would be akin to stopping an elevator between floors: there is nothing of particular value in accessing a "half floor" and stopping midfloor simply brings the model implementation to a crashing halt.

However, we can reconstruct a set of compensated data points post hoc using sampled offsets and the original raw data points. We can then make a side-by-side

comparison of category distributions to evaluate whether the reduction in variance is qualitatively similar between the two models. We can also use this reconstruction to visualize the degree of correlation between compensated token values from each model.

***Ideal offsets and pipeline point estimates.*** In the preprocessing stage of the pipeline model, all data points are centered by speaker and phoneme. Centering is accomplished by grouping the dataset by speaker and phoneme, and subtracting mean durations for speaker-phoneme pairs from the individual tokens.

The correlate of these point estimates in our ideal compensator model are our `offset` parameter distributions. Comparing point estimates with `offset` distributions, we expect to see the pipeline point estimates fall on or around where we would expect to find the distribution means for the respective `offset` speaker-phoneme combination.

# Results

***Offsets and centered durations.*** In our model, the `offset` parameter estimates the influence of each *phoneme* by *talker*. Figure 7 shows density plots of `offset` samples generated by the Stan implementation. For clarity's sake, we have plotted the offset distributions for the verb stem vowel by phoneme for talker 205 only.

In our ideal compensator model the sampled offset for a given talker/phoneme combination is combined with the observed durational token for the corresponding talker/phoneme. The variance we see in these offset sample distributions represents *listener uncertainty* regarding the amount of influence talker and phoneme have on the observed durations.

In contrast, the vertical dashed lines represent the corresponding non-bayesian point estimate of the offset calculated by finding the mean log duration by talker and phoneme. These point estimates are the exact amounts that were subtracted from corresponding talker/phoneme durations during the preprocessing (centering) stage of the pipeline-style model.

Visually, we can see that the point estimates for each talker/phoneme combination roughly correlate with the means for each offset distribution. However, we can also see that for certain phonemes (namely /ɑ/,/æ/, /eɪ/ and /ʊ/) point estimates fall noticeably to the left or right of the inferred distribution means. This disparity is due to the fact that there are fewer data points for these particular phonemes, as is illustrated in Figure 8. Also in Figure 8, we see scatterplots visually demonstrating the correlation between mean compensated durations over all samples from the ideal compensator model, and centered token durations from the pipeline preprocessing (centering) stage.

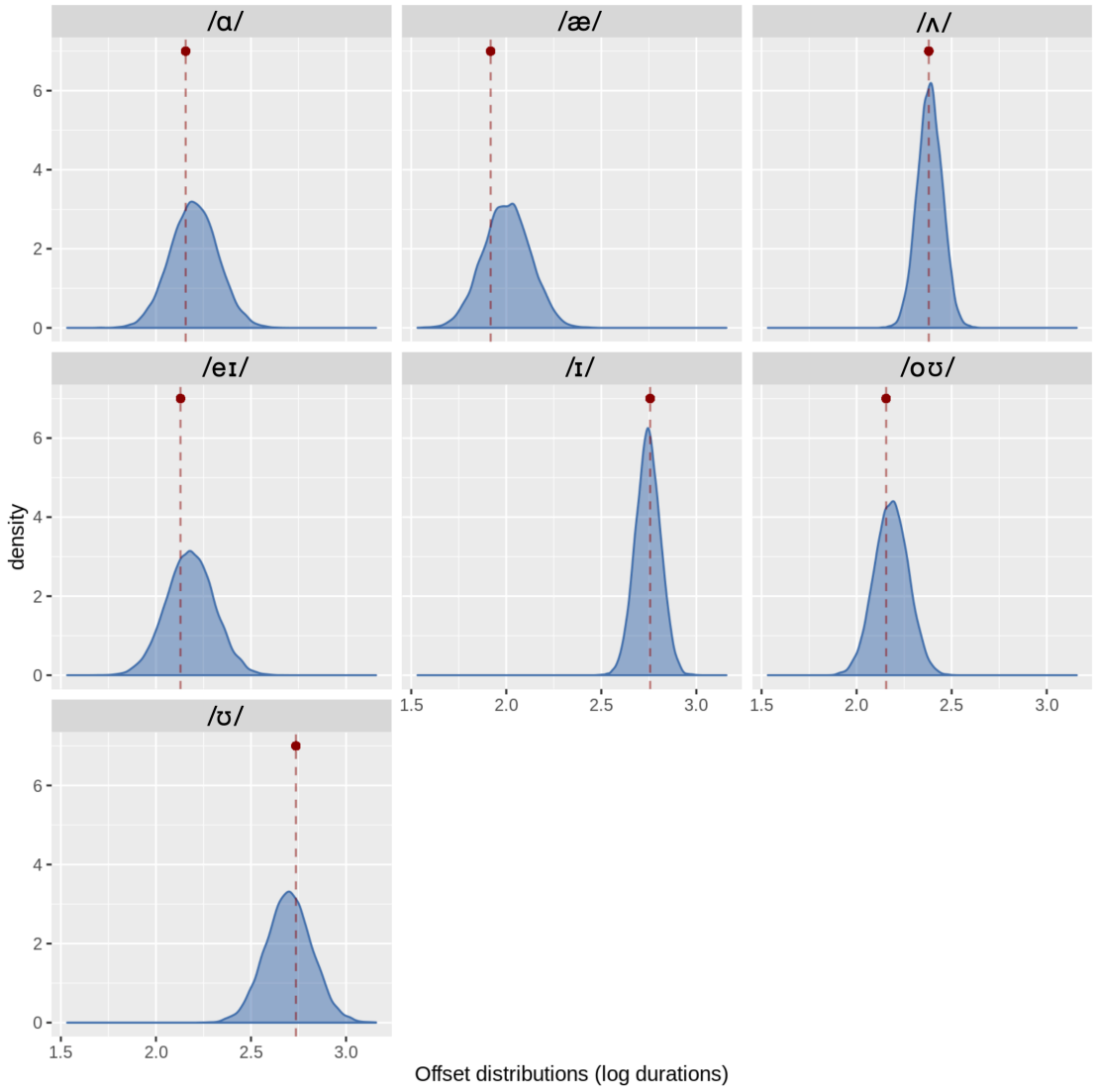


*Figure 7:* Density plots of sampled offsets for talker 205, verb stem vowel. Each pane shows the distribution of offsets for one of the 7 verb stem vowel segments. Red dashed lines represent the corresponding "pipeline" point estimate for talker/phoneme offset.

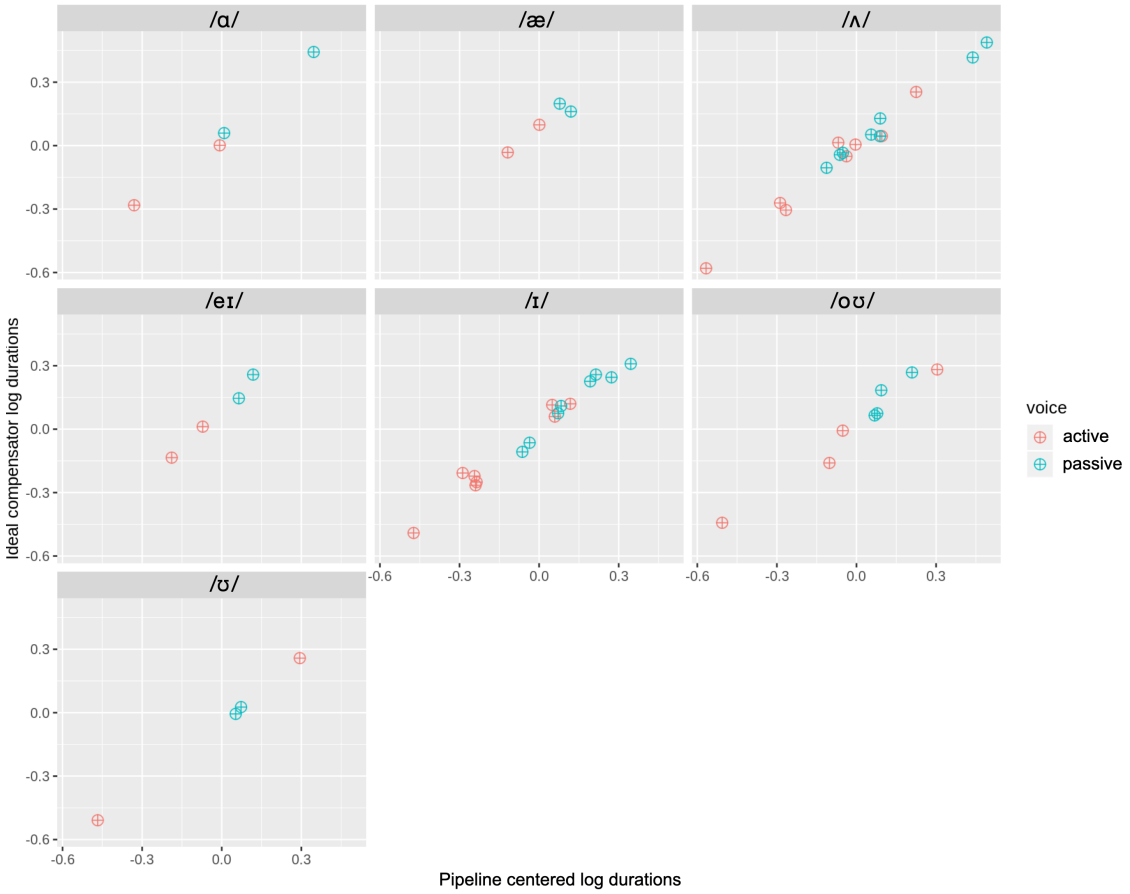


*Figure 8:* Correlations between ideal compensator and pipeline centered durations by phoneme for talker 205. Red data points are active, blue are passive.

***Verb stem vowel density plots.*** Figure 9 shows a triptych of verb stem vowel duration density plots: the leftmost panel shows active/passive distributions of raw, non-compensated durations; the center panel shows our reconstruction of compensated data distributions from the ideal compensator model; the rightmost panel shows distributions obtained after centering (non-bayesian compensations). Visually inspecting Figure 9 we see that both ideal compensator and pipeline-style compensation reduce variance in active/passive distributions in comparison to non-compensated token distributions.

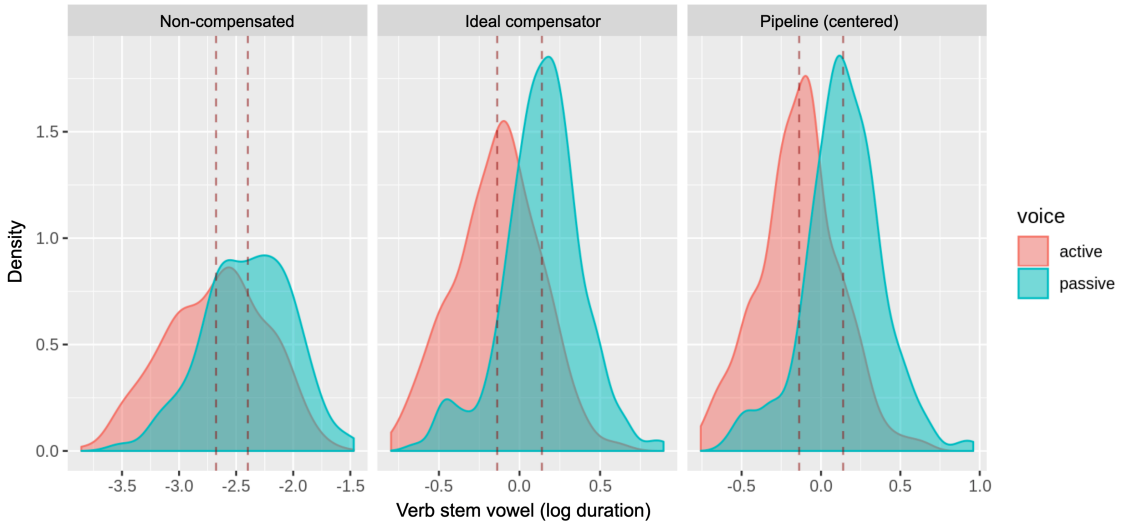


*Figure 9.* Density plots of verb stem vowel durations by syntactic voice (active/passive), all talkers and sentences. Left: raw durations. Center: ideal compensator reconstructed durations. Right: 'pipeline' centered data.

***Token posteriors.*** The following three beeswarm plots (figures 10a, 10b, and 10c) show token posteriors for both non-compensated and ideal compensator implementations. All token posterior values reflect the posterior probability of the actual or true structure. The advantage of this sort of visualization is that it allows us to visually contrast the behavior of our ideal compensator model posteriors with individual posteriors from the non-compensated model.

In Figure 10a we see that while the non-compensated posteriors tend to cluster around the dashed line at $p = 0.5$ (active/passive syntax is completely ambiguous), the ideal compensator model posteriors seem to shift and spread themselves well above the $p = 0.5$ mark. This difference between ideal compensator and non-compensated posteriors is summarized in the mean posterior probabilities for each group, represented

by red dots and error bars. The mean posterior probability for the non-compensated implementation is 0.552.

In terms of listener belief (i.e., how strongly does the listener believe a given token has been drawn from and active or passive sentence) this non-compensated mean is just slightly higher than $p = 0.5$, or chance levels for active/passive. In contrast, the mean posterior probability for the ideal compensator implementation is 0.602. This is just slightly less than the 0.619 mean posterior probability for the pipeline-style model (not visualized for clarity's sake). As we will see in a moment, this increase in the overall degree of listener belief translates into a higher mean accuracy score in classification over the non-compensated model variant.

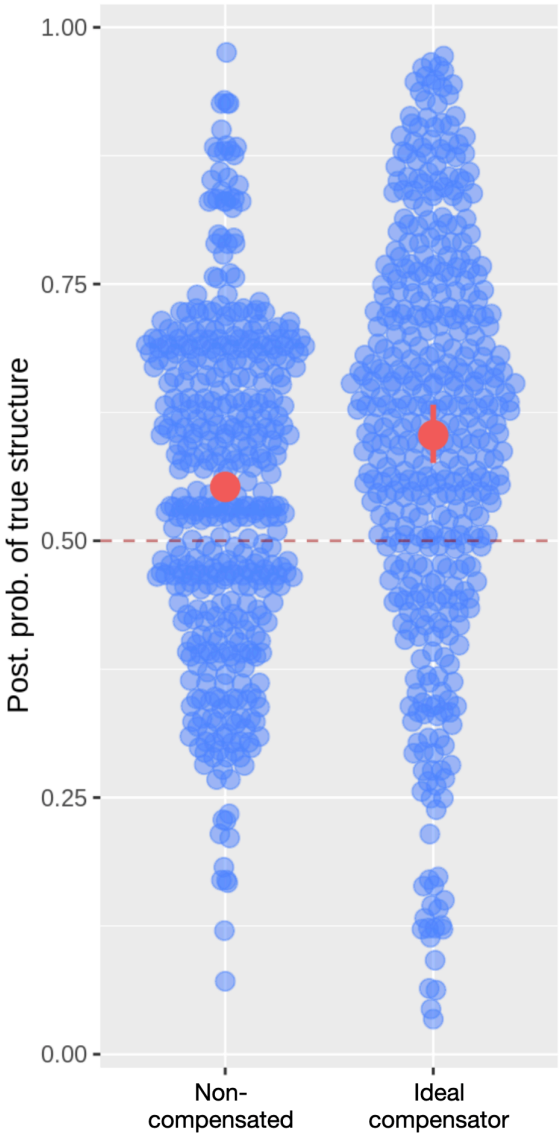


*Figure 10a.* Posterior values for all talkers and sentences, non-compensated and ideal compensator models. Right side (ideal compensator): each blue dot represents the mean token posterior over all posterior samples. The red dot represents the mean posterior over all tokens and samples, and error bars indicate the 95% credible interval over all samples. Left side (non-compensated): each blue dot represents a token posterior, the red dot the mean posterior over all tokens, and error bars indicate standard error.

The beeswarm plots in Figure 10b again show token posteriors from non-compensated and ideal compensator implementations. However, we can now see the "behavior" of individual tokens where we have compensated for talker and phoneme.

Visually, we can see that while some posteriors moved below the $p = 0.5$ threshold (negative change in probability indicated by red dots/lines), many more posteriors either changed very little (indicated by white dots/lines) or increased in probability (positive change indicated by blue dots/lines). In the end, 290 of 439 posteriors increased in probability under our ideal compensator implementation (293 posteriors increased under the pipeline model, not shown). The maximum increase in probability was 0.473 and the maximum decrease was 0.420.

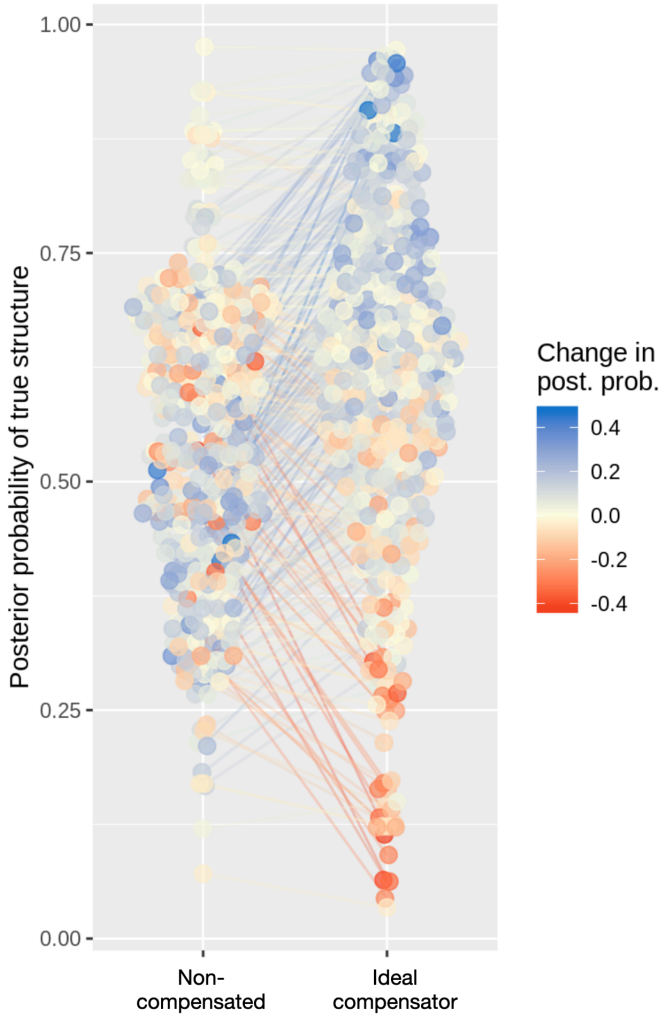


*Figure 10b.* Posterior values for all talkers and sentences, non-compensated and ideal compensator models. Each dot represents a token posterior. Lines between corresponding non-/ideally compensated posterior pairs show the trajectory of posterior movement. Red lines/dots indicate decreased probability between non-/compensated posteriors; blue lines/dots indicate increased probability.The degree of change in posterior probability is represented by the slope of the line and the darkness of the color. Steeper slopes and darker colors indicate greater change in probability.

In order to determine classification accuracy of each model variant, we thresholded our posteriors at $p = 0.5$ and calculated mean accuracy using a winner-takes-all approach. We found that mean accuracy for the non-compensated model was 61%

while mean accuracy for the ideal compensator model was 73.6%, an increase of 12.6%. Mean accuracy for the pipeline-style model (not shown) was 75.2%.

Figure 10c visually illustrates how our ideal compensator implementation affected individual posterior tokens with respect to overall classification accuracy. By and large, compensation did not affect the classification accuracy of most tokens (332, yellow dots). However, there were more than three times as many tokens that improved overall accuracy (81, green dots) than negatively impacted it (26, red dots).

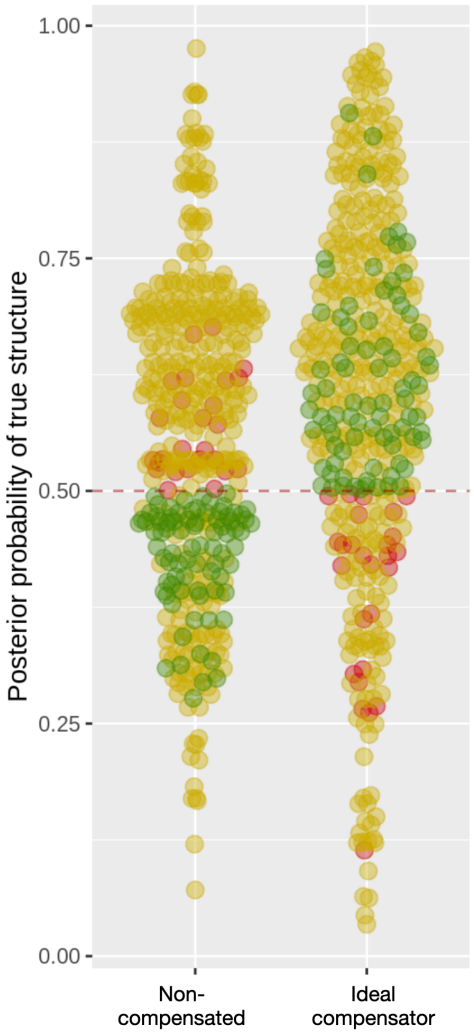


*Figure 10c.* Posterior values for all talkers and sentences, non-compensated and ideal compensator models. Token posteriors are coded according to whether they increased (green) decreased (red) or had no effect on classifier accuracy (yellow).

# Discussion

In this work we have outlined a new computational framework for speech perception and compensation, *the ideal compensator*. As with previous models of compensation, one of the primary goals of this model is to overcome the variability inherent to acoustic speech cues in order to accurately infer the linguistic category intended by the speaker.

In contrast to most compensation procedures, however, this model does not adopt a *pipeline* approach to compensation. This is because instead of starting from the perspective that acoustic cues contain a jumble of informative *breadcrumbs* and misleading *red herrings*, we start with the assumption that cues are *inherently ambiguous* and thus only *partially informative.* As such, speech perception is a matter of *inference under uncertainty* and the goal of our listener model is to optimally infer the linguistic intent of the speaker based on the observed cue. Because our listener model actually *infers how to compensate* based on a speaker's *generative model* while also *simultaneously* inferring linguistic category, we believe our approach is novel when compared to other compensation models and procedures.

While the aim of developing this new computational model of compensation was to address *conceptual* issues in pipeline-style models and not *performance* issues, the ideal compensator model did indeed perform quite well in our model comparison. With a mean accuracy of 73.6%, our model increased classifier accuracy by 12.6% over the non-compensated model, just short of the 75.2% accuracy score achieved by the pipeline-style model.

In future work, we would also like to see how well our model results predict behavioral data. To do so, we intend on collecting behavioral data based on the exact same active/passive sentence pairs used in our model implementation and comparing these behavioral results with model results. While we would expect similar *overall* mean behavioral and model accuracy results in such a comparison, it would be interesting to

see whether model accuracy results could predict behavioral results *by the sentence* as well.

Although we feel that our model provides a better conceptual framework for compensation than pipeline models, from a practical standpoint there is a small "cost" to the researcher associated with implementing fully bayesian compensation. In comparison with the relatively simple mathematical procedures associated with pipeline-style compensation schemes, (such as the centering procedure we used in our pipeline comparison model) implementing a fully bayesian compensation model in an existing software framework (such as Stan) requires quite a bit more work both in terms of model refactoring and programming effort. So while a fully bayesian implementation allows us to execute a conceptually rigorous computational model, the relative simplicity and ease of computation that is part and parcel for pipeline-style procedures may remain an attractive choice for areas of research where the conceptual stakes surrounding compensation are low. However, as advanced software programs like Stan continue to become more powerful, accessible, flexible and computationally efficient, we are likely to see more researchers embrace computational models of perception that are both conceptually complete and implementationally expedient.

# Appendix A

List of recorded sentences.

28 Active Sentences:
The bear was licking the dog.
The bear was punching the dog.
The cat was pushing the mouse.
The cat was touching the rhino.
The cow was poking the zebra.
The dog was licking the bear.
The dog was punching the bear.
The duck was washing the rabbit.
The elephant was kicking the kangaroo.
The fox was combing the lion.
The frog was shoving the monkey.
The frog was trapping the monkey.
The hippo was chasing the turtle.
The kangaroo was kicking the elephant.
The lion was combing the fox.
The monkey was pinching the rabbit.
The monkey was shoving the frog.
The monkey was trapping the frog.
The mouse was pushing the cat.
The pig was kissing the sheep.
The pig was scrubbing the sheep.
The rabbit was pinching the monkey.
The rabbit was washing the duck.
The rhino was touching the cat.
The sheep was kissing the pig.
The sheep was scrubbing the pig.
The turtle was chasing the hippo.
The zebra was poking the cow.

28 Passive Sentences:
The bear was licked by the dog.
The bear was punched by the dog.
The cat was pushed by the mouse.
The cat was touched by the rhino.
The cow was poked by the zebra.
The dog was licked by the bear.
The dog was punched by the bear.
The duck was washed by the rabbit.

The elephant was kicked by the kangaroo.
The fox was combed by the lion.
The frog was shoved by the monkey.
The frog was trapped by the monkey.
The hippo was chased by the turtle.
The kangaroo was kicked by the elephant.
The lion was combed by the fox.
The monkey was pinched by the rabbit.
The monkey was shoved by the frog.
The monkey was trapped by the frog.
The mouse was pushed by the cat.
The pig was kissed by the sheep.
The pig was scrubbed by the sheep.
The rabbit was pinched by the monkey.
The rabbit was washed by the duck.
The rhino was touched by the cat.
The sheep was kissed by the pig.
The sheep was scrubbed by the pig.
The turtle was chased by the hippo.
The zebra was poked by the cow.

# Appendix B

Stan code for ideal compensator model.

```
data {
  int n;

  int n_voice;
  int n_syl_label;
  int n_participant_num;

  int voice[n];
  int syl_label[n];
  int participant_num[n];
  int istest[n];
  real duration[n];

}

parameters {

  real anti_mu_raw[n_participant_num];
  real anti_mu_mu;
  real<lower=0> anti_mu_sigma;
  real<lower=0> sigma[n_voice];
  real offset[n_participant_num, n_syl_label];
  real offset_mean;
```

```
  real<lower=0> offset_scale;

}

transformed parameters {

  real mu[n_voice, n_participant_num];
  real duration_shifted[n];

  for(j in 1:n_participant_num){
    real anti_mu;
    anti_mu = anti_mu_mu + anti_mu_sigma * anti_mu_raw[j];
    mu[1,j] = anti_mu;
    mu[2,j] = -anti_mu;
  }


  for (i in 1:n) {
    duration_shifted[i] = duration[i] + offset[participant_num[i],
syl_label[i]];
  }
}

model {

  anti_mu_mu ~ normal(0,1);
  anti_mu_sigma ~ cauchy(0, 1);
  sigma ~ cauchy(0, 1);


  for(j in 1:n_participant_num){
    anti_mu_raw[j] ~ std_normal();
  }


  offset_mean ~ normal(0,1);
  offset_scale ~ cauchy(0,1);

  for (i in 1:n_participant_num) {
    for (j in 1:n_syl_label) {
      offset[i,j] ~ normal(offset_mean, offset_scale);
    }
  }

  for (i in 1:n) {
    if (istest[i]) {
      real l_lhoods[n_voice];
      for (j in 1:n_voice){
        l_lhoods[j] = normal_lpdf(duration_shifted[i] |
mu[j,participant_num[i]], sigma[j]) + log(0.5);
      }
      target += log_sum_exp(l_lhoods);
    }else{
      duration_shifted[i] ~ normal(mu[voice[i], participant_num[i]],
sigma[voice[i]]);
    }
  }
}
```

```
generated quantities{

  real log_lh[sum(istest), n_voice];
  int k;
  k = 1;

  for (i in 1:n){
    if(istest[i]){
      for (j in 1:n_voice){
        log_lh[k,j] = normal_lpdf(duration_shifted[i] |
mu[j,participant_num[i]], sigma[j]);
      }
      k=k+1;
    }
  }
}

// end Stan code
```

# Bibliography


Adank, P., Smits, R., van Hout, R. (2004). "A Comparison of Vowel Normalization Procedures for Language Variation Research." *The Journal of the Acoustical Society of America* 116 (5): 3099–3107. doi:10.1121/1.1795335.

Ainsworth, W. A. (1972). "Duration as a Cue in the Recognition of Synthetic Vowels." *The Journal of the Acoustical Society of America* 51 (2B): 648–51. doi:10.1121/1.1912889.

Aylett, M., & Turk, A. (2004). "The Smooth Signal Redundancy Hypothesis: A Functional Explanation for Relationships Between Redundancy, Prosodic Prominence, and Duration in Spontaneous Speech." *Language and Speech* 47 (1): 31–56. doi:10.1177/00238309040470010201.

Beach, C. M. (1991). "The Interpretation of Prosodic Patterns at Points of Syntactic Structure Ambiguity: Evidence for Cue Trading Relations." *Journal of Memory and Language* 30 (6): 644–63. doi:10.1016/0749-596X(91)90030-N.

Bell, A., Gregory, M. L., Brenier, J. M., Jurafsky, D., Ikeno, A., & Girand, C. (2002). Which predictability measures affect content word durations?. In *ISCA Tutorial and Research Workshop (ITRW) on Pronunciation Modeling and Lexicon Adaptation for Spoken Language Technology*.

Boersma, P. & Weenink, D. (2020). Praat: doing phonetics by computer [Computer program]. Version 6.1.37, retrieved 16 December 2020 from http://www.praat.org/

Brown, M., Salverda, A. P., Dilley, L.C., Tanenhaus, M. K. (2011). "Expectations from Preceding Prosody Influence Segmentation in Online Sentence Processing." *Psychonomic Bulletin & Review* 18 (6): 1189–96.

Clayards, M., Tanenhaus, M. K., Aslin, R. M., Jacobs, R. A. (2008). "Perception of Speech Reflects Optimal Use of Probabilistic Speech Cues." *Cognition* 108 (3): 804–9. doi:10.1016/j.cognition.2008.04.004.

Clopper, C. G. (2009). "Computational Methods for Normalizing Acoustic Vowel Data for Talker Differences." *Language and Linguistics Compass* 3 (6): 1430–42. ` doi:10.1111/j.1749-818X.2009.00165.x.

Crystal, T. H. & House, A. S. (1990). "Articulation Rate and the Duration of Syllables and Stress Groups in Connected Speech." *The Journal of the Acoustical Society of America* 88 (1): 101–12. doi:10.1121/1.399955.

Delattre, P., Liberman, A. M., Cooper, F. S., Gerstman, L. J. (1952). "An Experimental Study of the Acoustic Determinants of Vowel Color; Observations on One- and Two-Formant Vowels Synthesized from Spectrographic Patterns." WORD 8 (3): 195–210. doi:10.1080/00437956.1952.11659431.

Disner, S. F.. (1980). "Evaluation of Vowel Normalization Procedures." *The Journal of the Acoustical Society of America* 67 (1): 253–61. doi:10.1121/1.383734.

Fabricius, A. H., Watt, D., Johnson, D. E. (2009). "A Comparison of Three Speaker-Intrinsic Vowel Formant Frequency Normalization Algorithms for Sociophonetics." *Language Variation and Change* 21 (3): 413–35.

Flynn, N., & Foulkes, P. (2011, August). Comparing Vowel Formant Normalization Methods. In *ICPhS* (pp. 683-686).

Fowler, C. A., & Housum, J. (1987). "Talkers' Signaling of 'New' and 'Old' Words in Speech and Listeners' Perception and Use of the Distinction." *Journal of Memory and Language* 26 (5): 489–504. doi:10.1016/0749-596X(87)90136-7.

Francis, A. L., Ciocca, V., & Ching Yu, J. M. (2003). Accuracy and variability of acoustic measures of voicing onset. *The Journal of the Acoustical Society of America*, *113*(2), 1025-1032.

Hillenbrand, J. M., Clark, M. J., Houde, R. A. (2000). "Some Effects of Duration on Vowel Recognition." *The Journal of the Acoustical Society of America* 108 (6): 3013–22. doi:10.1121/1.1323463.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical society of America*, *97*(5), 3099-3111.

Hogan, J. T., & Rozsypal, A. J. (1980). "Evaluation of Vowel Duration as a Cue for the Voicing Distinction in the Following Word−final Consonant." *The Journal of the Acoustical Society of America* 67 (5): 1764–71. doi:10.1121/1.384304.

Holt, Y. F., Jacewicz, E., Fox, R. A. (2015). "Variation in Vowel Duration Among Southern African American English Speakers." *American Journal of Speech-Language Pathology* 24 (3): 460–69. doi:10.1044/2015_AJSLP-14-0186.

Jacewicz, E., Fox, R. A., Salmons, J. (2007). "Vowel Duration in Three American English Dialects." *American Speech* 82 (4): 367–85. doi:10.1215/00031283-2007-024.

Kessinger, R. H., & Blumstein, S. (1998). "Effects of Speaking Rate on Voice-Onset Time and Vowel Production: Some Implications for Perception Studies." *Journal of Phonetics* 26 (2): 117–28. doi:10.1006/jpho.1997.0069.

Kim, M., & Stoel-Gammon, C. (2010). "Segmental Timing of Young Children and Adults." *International Journal of Speech-Language Pathology* 12 (3): 221–29. doi:10.3109/17549500903477363.

Kondaurova, M. V., & Francis. A. (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 (6): 3959–71. doi:10.1121/1.2999341.

Lai, M. K. (2015). Kicking or Being Kicked? Using Acoustics as Unconscious Cues to the Passive. Interdisciplinary Honors Thesis, Department of Linguistics and Department of Psychology School of Arts and Sciences, Rutgers University)

Lehiste, I. (1972). "The Timing of Utterances and Linguistic Boundaries." *The Journal of the Acoustical Society of America* 51 (6B): 2018–24. doi:10.1121/1.1913062.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., Studdert-Kennedy, M. (1967). "Perception of the Speech Code." *Psychological Review* 74 (6): 431–61. doi:10.1037/h0020279.

Lobanov, B. M. (1971). "Classification of Russian Vowels Spoken by Different Speakers." *The Journal of the Acoustical Society of America* 49 (2B): 606–8. doi:10.1121/1.1912396.

McMurray, B., and Jongman, A. (2011). "What Information Is Necessary for Speech Categorization? Harnessing Variability in the Speech Signal by Integrating Cues Computed Relative to Expectations." *Psychological Review* 118 (2): 219–46. doi:10.1037/a0022325.

Mermelstein, P. (1978). “On the Relationship Between Vowel and Consonant Identification When Cued by the Same Acoustic Information.” *Perception & Psychophysics* 23 (4): 331–36. doi:10.3758/BF03199717.

Miller, J. D. (1989). “Auditory−perceptual Interpretation of the Vowel.” *J. Acoust.Soc. Am.* 85 (5): 22.

Picheny, M. A., Durlach, N. I., Braida, L. D. (1986). “Speaking Clearly for the Hard of Hearing II: Acoustic Characteristics of Clear and Conversational Speech.” *Journal of Speech, Language, and Hearing Research.* 29 (4): 434–46.

Raphael, L. J. (1972). “Preceding Vowel Duration as a Cue to the Perception of the Voicing Characteristic of Word−Final Consonants in American English.” *The Journal of the Acoustical Society of America* 51 (4B): 1296–1303. doi:10.1121/1.1912974.

Rehrig, G. L. (2017). *Acoustic correlates of syntax in sentence production and comprehension* (Doctoral dissertation, Rutgers University-School of Graduate Studies).

Sawusch, J. R., & Pisoni D. (1974). “On the Identification of Place and Voicing Features in Synthetic Stop Consonants.” *Journal of Phonetics* 2 (3): 181–94.

Sorensen, J. M., Cooper,W. E., Paccia, J. (1978). “Speech Timing of Grammatical Categories.” *Cognition* 6 (2): 135–53. doi:10.1016/0010-0277(78)90019-7.

Stan Development Team. (2020). Stan Modeling Language Users Guide and Reference Manual, 2.18.0. https://mc-stan.org

Strand, E. A., & Johnson, K. (1996). “2. Gradient and Visual Speaker Normalization in the Perception of Fricatives.” In *Natural Language Processing and Speech Technology*, edited by Dafydd Gibbon, 14–26. Berlin, Boston: De Gruyter. doi:10.1515/9783110821895-003.

Stromswold, K., Eisenband, J., Norland, E., & Ratzan, J. (2002, March). Tracking the acquisition and processing of English passives: Using acoustic cues to disambiguate actives and passives. In *CUNY conference on sentence processing* (Vol. 2123). New York: NY

Stromswold, K., Kharkwal, G., Sorkin, J. (*in review*). “Tracking the Elusive Passive: The Processing of Spoken Passives.”

Syrdal, A. K., & Gopal, H. (1986). “A Perceptual Model of Vowel Recognition Based on the Auditory Representation of American English Vowels.” *The Journal of the Acoustical Society of America* 79 (4): 1086–1100. doi:10.1121/1.393381.

Toda, M. (2007). "Speaker Normalization of Fricative Noise: Considerations on Language-Specific Contrast," 16th International Congress on Phonetic Sciences, Saarbrucken, Germany: 825-828

Tsao, Y. C., & Weismer, G. (1997). "Interspeaker Variation in Habitual Speaking Rate: Evidence for a Neuromuscular Component." *Journal of Speech, Language, and Hearing Research.* 40 (4): 858.

Turk, A. (2010). "Does Prosodic Constituency Signal Relative Predictability? A Smooth Signal Redundancy Hypothesis." *Laboratory Phonology* 1 (2).

Turk, A., & Shattuck-Hufnagel, S. (2007). "Multiple Targets of Phrase-Final Lengthening in American English Words." *Journal of Phonetics* 35 (4): 445–72.

Umeda, N. (1975). "Vowel Duration in American English." *The Journal of the Acoustical Society of America* 58 (2): 434–45. doi:10.1121/1.380688.

Whalen, D. H. (1989). Vowel and consonant judgments are not independent when cued by the same information. *Perception & Psychophysics*, *46*(3), 284-292

Wright, C. E. (1979). "Duration Differences Between Rare and Common Words and Their Implications for the Interpretation of Word Frequency Effects." *Memory & Cognition* 7 (6): 411–19. doi:10.3758/BF03198257.

Yuan, J., & Liberman, M. (2008). Speaker identification on the SCOTUS corpus. Proceedings of Acoustics '08, The Penn Phonetics Lab Forced Aligner (P2FA).