An Evidence Accumulation Model of Acoustic Cue Weighting in Vowel Perception

Gabriel Tillman[1], Titia Benders[2], Scott D. Brown[1], and Don van Ravenzwaaij[1,3]

[1]School of Psychology, University of Newcastle, Australia

[2]ARC Centre or Excellence in Cognition and its Disorders; Department of Linguistics, Macquarie University, Australia.

[3]Faculty of Behavioural and Social Sciences, University of Groningen, Netherlands

Word Count: 11456

Author Note

Correspondence concerning this article should be addressed to Gabriel Tillman, School of Psychology, University of Newcastle, Callaghan NSW 2308, Australia.

Contact: gabriel.tillman@newcastle.edu.au

Abstract

Listeners rely on multiple acoustic cues to recognize any phoneme. The relative contribution of these cues to listeners' perception is typically inferred from listeners' categorization of sounds in a two-alternative forced-choice task. Here we advocate the use of an evidence accumulation model to analyze categorization as well as response time data from such cue weighting paradigms in terms of the processes that underlie the listeners' categorization. We tested 30 Dutch listeners on their categorization of speech sounds that varied between typical /ɑ/ and /aː/ in vowel quality (F1 and F2) and duration. Using the linear ballistic accumulator model, we found that the changes in spectral quality and duration lead to changes in the speed of information processing, and the effects were larger for spectral quality. In addition, for stimuli with atypical spectral information, listeners accumulate evidence faster for /ɑ/ compared to /aː/. Finally, longer durations of sounds did not produce longer estimates of perceptual encoding time. Our results demonstrate the utility of evidence accumulation models for learning about the latent processes that underlie phoneme categorization. The implications for current theory in speech perception as well as future directions for evidence accumulation models are discussed.

Keywords: Phoneme Categorization, Linear Ballistic Accumulator, Response Time

Phonemes are linguistic representations with an acoustic counterpart that can be characterized in a multidimensional acoustic space. Values along each acoustic dimension can serve as cues for listeners to recognize a speech sound as a particular phoneme. The cues, such as first (F1) and second (F2) formant frequency, duration, and fundamental frequency, are acoustic and continuous. Yet, these cues map onto phonological representations that may not be continuous, i.e., the phonemes. Phonemes can be viewed as clusters of exemplars in a multidimensional phonetic space (Pierrehumbert, 2001), or as abstract representations that are connected to a range of values along multiple phonetic dimensions (Boersma, 2007). Speech perception is the process of mapping the continuous acoustic information onto the phonological categories (Holt & Lotto, 2010).

Each phoneme correlates with multiple acoustic dimensions (Lisker, 1986) and multiple acoustic cues influence each phoneme categorization (Holt & Lotto, 2006). Some cues contribute strongly to a listener's decision and some cues contribute weakly to the decision – a phenomenon called *cue weighting*. Cue weighting in speech perception often reflects the reliability of the cues for the recognition of phonological categories in the ambient language (Holt & Lotto, 2010).

Researchers investigate cue weighting using a range of methods: computational statistical modeling (Toscano & McMurray, 2010; McMurray, Aslin, & Toscano, 2009), eye-tracking (Reinisch & Sjerps, 2013), neuro-physiological measurements (Lipski, Escudero, & Benders, 2012), in normal-hearing and hearing-impaired populations (Winn, Chatterjee, & Idsardi, 2012; Winn, Rhone, Chatterjee, & Idsardi, 2013), and most commonly, with behavioral data from phoneme categorization tasks (Repp, 1982). In the latter, researchers systematically vary the acoustic cue values of sounds that are

played to participants and observe the effects on phoneme categorization. Cue weighting is measured by how much each cue contributes to the categorization response and is therefore based on a measure at the end of processing and decision-making. To use categorization data to learn how acoustic cues are connected with phonological categories, we have to make the assumption that categorization data directly reflects the mapping of the experimentally manipulated cues onto the phonological categories. However, there are two fundamental issues with this assumption.

The first problem is that cue weighting is measured for a phoneme contrast and does not give us the association between cues and each category separately (i.e., the cue-to-*one*-phoneme mapping). For example, a cue that is strongly associated with one phoneme in the contrast and only loosely associated with the other phoneme can appear to be indiscriminately 'heavily weighted', because the cue contributes relatively strongly to the decision between these two phonemes. Given this confound, it is difficult to infer how much each acoustic cue contributes to each individual phoneme in the contrast.[1]

The second problem is that researchers only observe the association between experimentally manipulated cues and overt behavioral responses (i.e., the cue-to-response association), which means they need to assume that this association directly reflects the cue-to-phoneme mapping. Yet, a strong cue-to-phoneme mapping may not manifest as a strong cue-to-response association. One reason for a weak association between cues and responses despite a strong mapping could be that listeners do not have good access to the cue. Perhaps the cue is not always loud enough to be perceived or perhaps the cue appears late in the speech signal. Cues that appear later in the signal might be strongly associated with a phoneme, but may not appear as such

[1]We are interested in how much each cue contributes to each phoneme in the contrast, which is not the same thing as investigating how much an acoustic cue contributes to a particular phoneme outside the context of the contrast.

in a categorization task because earlier appearing cues have already been processed and potentially determined the response (cf. McMurray, Clayards, Tanenhaus, & Aslin, 2008; Reinisch & Sjerps, 2013). In order to address this issue, it is necessary to learn more about how listeners process the acoustic cues. For instance, is cue weighting as inferred from categorization data driven by differences in when cues are available in time, or by listeners processing one acoustic cue faster than another? In any case, researchers need a way to investigate such latent processes in order to derive more accurate conclusions about acoustic cue weighting in terms of cue-to-phoneme mapping.

Both problems limit our ability to use categorization data to learn about how listeners map acoustic information onto phonological categories. Therefore, we need a method to account for how acoustic cues are cognitively processed for each phoneme in the contrast. Below we discuss response times (RT) and eye-tracking, which are alternative measures to categorization data that give insight into the processing of acoustic information, but neither of these measures address both issues.

First, researchers can use the RT associated with phonological decisions to investigate phoneme perception. For example, researchers have investigated processing differences between non-identical and identical phonemes (Pisoni & Tash, 1974) and have determined that phoneme categorization decisions depend more on a phoneme's position in acoustic space than their perceived category goodness (Miller, 2001).

However, there are difficulties with analyzing either choice data or RT in isolation. We know that the accuracy of a decision depends on how fast the decision is made – in other words, a participant's speed-accuracy trade-off setting (e.g., Wickelgren, 1977; Luce, 1986; Heitz, 2014). Without any insight into the trade-off settings used by participants, researchers may draw incorrect conclusions from choice or RT data alone.

Furthermore, to analyze RT researchers typically average over all observations for each participant in order to subject the means to a statistical test, such as ANOVA. Analyzing the RT in this manner can lead to researchers drawing incorrect conclusions (e.g., Ashby, Maddox, & Lee, 1994; Curran & Hintzman, 1995; Heathcote, Brown, & Mewhort, 2000) and does not allow researchers to learn about the latent cognitive processes involved in speech perception. For example, an RT of 700ms on a given trial suggests that 700 ms was needed to perceptually encode the sound, decide what phoneme was heard, and execute a motor response. But, we cannot know how long each of these processes takes from analyzing mean RT with linear models. Given that RT is a measure at the end of processing, analyzing RTs alone only inform researchers about the cue-to-response association but not the cue-to-phoneme mapping.

Eye tracking is a another useful measure that is frequently used to observe how listeners process experimentally manipulated cues online (e.g., Allopenna, Magnuson, & Tanenhaus, 1998). For example, eye-tracking can be used to infer whether the order in which acoustic cues become available to listeners affects listeners' interpretation of the speech signal (McMurray et al., 2008; Reinisch & Sjerps, 2013). In fact, McMurray et al. (2008) showed that listeners do not wait for cues that are available later in a speech signal (e.g., vowel duration) to begin using earlier available cues (e.g., voice onset time). Moreover, Reinisch and Sjerps (2013) showed that listeners use vowel spectral cues before vowel duration cues, because listeners need to wait for the vowel offset before they have full information about the duration.

Eye-tracking data, like RTs, are typically averaged over all observations for each participant, meaning that the aforementioned objections against inferences from averaged data hold for eye-tracking data as well. Furthermore, eye-tracking data are

plagued by the first confound of categorization data discussed in detail above. That is, they can give insight into cue-weighting, but do not give the cue-to-one-phoneme mapping for phoneme contrasts.

Categorization, RT, and eye-tracking are all useful methods in speech perception research, but none of them address both the cue-to-one-phoneme mapping and the cue-to-phoneme mapping issues discussed above. In this paper, we advocate the simultaneous analysis of phoneme categorization data with their associated RTs using an evidence accumulation model (e.g., Ratcliff & McKoon, 2008; Usher & McClelland, 2001; Brown & Heathcote, 2008). The following section describes what evidence accumulation models are and what they can add to the current speech perception literature.

Evidence Accumulation Models

Since their advent (e.g., Stone, 1960), evidence accumulation models have been applied to many different fields (see Donkin & Brown, in press, for a review) – including recognition memory, brightness discrimination, lexical decision, consumer choice, workload capacity, optimal decision-making, implicit association, the effects of alcohol on decision-making, and the neural mechanisms of decision-making (Ratcliff, 1978; Ratcliff & Rouder, 1998; Wagenmakers, Ratcliff, Gomez, & McKoon, 2008; Hawkins et al., 2013; Eidels, Donkin, Brown, & Heathcote, 2010; Evans & Brown, 2016; van Ravenzwaaij, van der Maas, & Wagenmakers, 2011; van Ravenzwaaij, Dutilh, & Wagenmakers, 2012; Forstmann et al., 2008). These models posit that people continually sample evidence from the environment in order to choose between decision alternatives. The quality of this evidence governs the speed of processing and the amount of evidence needed governs the response caution. If people are biased they will

need different amounts of evidence before committing to each of the responses. These models also allow researchers to estimate the time needed for processes outside of the decision process, such as perceptual encoding and executing a motor response.

Evidence accumulation models have a track record for investigating latent cognitive processes. For example, it is typically found that as people age their RTs increase in cognitive tasks. For almost 20 years, the dominant theory of why performance declined with age was that aging resulted in a general slow-down (Salthouse, 1996). However, when different RT data sets that each exhibited this pattern were analyzed with an evidence accumulation model, researchers found that the locus of the slow-down in older people was higher response caution, not a lower processing speed (Ratcliff, Thapar, & McKoon, 2001, 2004; Thapar, Ratcliff, & McKoon, 2003; Ratcliff, Thapar, Gomez, & McKoon, 2004; Ratcliff, Thapar, & Mckoon, 2003).

A parsimonious evidence accumulation model that retains all of the explanatory power of more complex models (e.g., Ratcliff & Rouder, 1998; Usher & McClelland, 2001), while having the advantage of being tractable, is the linear ballistic accumulator (LBA; Brown & Heathcote, 2008).[2] The LBA has been applied to a number of perceptual discrimination paradigms (e.g., Ho, Brown, & Serences, 2009; Forstmann, Brown, Dutilh, Neumann, & Wagenmakers, 2010; Forstmann et al., 2008; Cassey, Heathcote, & Brown, 2014; van Ravenzwaaij, Provost, & Brown, 2016) and has been fit to tasks where the responses are categories (e.g., Hawkins et al., 2014; Trueblood, Brown, & Heathcote, 2014), which is the same set-up as used in a phoneme categorization task. Therefore, the LBA can be feasibly extended to model phonological

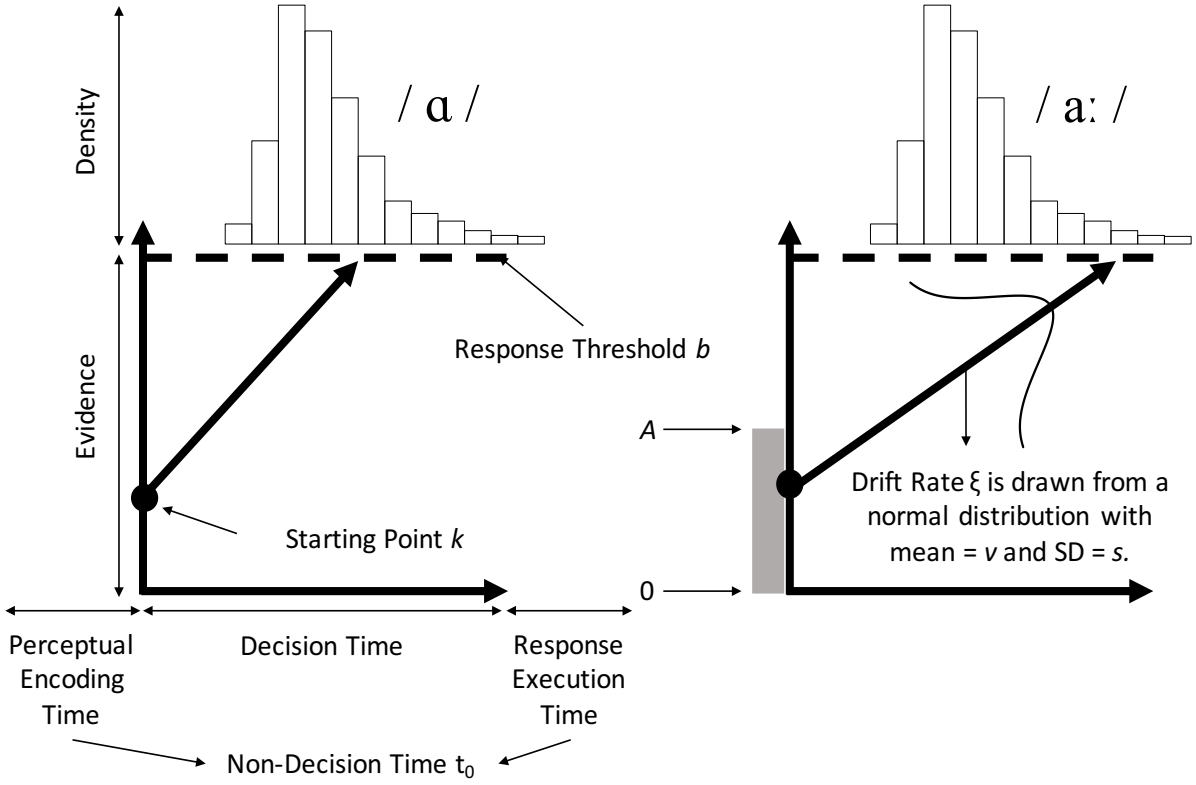
[2]There is a closed form expression for the likelihood function, which allows for relatively quick model fitting.

decisions in categorization tasks that yield choice data and RTs.

Suppose a participant needs to identify whether they have heard /ɑ/ or /aː/ – a schematic of the LBA explanation for this task is shown in Figure 1. At the beginning of a trial a participant hears a sound through a pair of headphones. It takes the listener time to perceptually encode the sound. After perceptual encoding, the evidence from the stimulus serves as input for the decision process. The LBA does not commit to what evidence is sampled from stimuli, only that the evidence leads to a response. In the discussion we suggest that one possible interpretation of the sampled evidence is the strength of the mapping between the acoustic information and the phoneme category associated with the response option. The evidence drives the decision process, which involves independent evidence accumulators for each response option. Each accumulator has a starting point of evidence accumulation k, which is a random value between 0 and A on each trial, a drift rate ξ that specifies the rate of evidence accumulation, and a response threshold b that specifies the amount of evidence needed to make a decision. From k, both accumulators independently race towards b at their respective drift rates. On each trial, drift rates are sampled from a normal distribution with mean drift rate v and standard deviation s.[3] The first accumulator to reach the response threshold determines the decision that is made. The time it takes for the LBA process to go from the starting point to the response threshold is the decision time. After a decision is made, the participant will need to overtly execute their response with a key press – the time needed for this overt response is the response execution time. The sum of the perceptual encoding time and response execution time makes up the non-decision time

[3]Human performance and the activity of neuron populations result in highly variable behavior in experiments, even when participants are presented with the same stimuli (see Usher & McClelland, 2001, for a discussion of sources of variability in evidence accumulation models). In addition, these variability parameters explain key aspects of RT data, such as slow error RT (Ratcliff, 1978) or fast error RT (Laming, 1968) relative to correct RT.

t_0. The total RT on a given trial is the sum of the decision time and the non-decision time. For convenience, an overview of the LBA parameters is provided in the table at the bottom of Figure 1.



Parameter labels and descriptions.

Symbol	Name	Psychological Interpretation
v	Mean Drift Rate	Speed of information processing
b	Response Threshold	Evidence required to make decision
A	Starting Point Variability	Variability in the initial evidence across trials
s	Drift Rate Variability	Variability in the speed of processing across trials
t_0	Non-Decision Time	Time for processes other than decision processing

Figure 1. The linear ballistic accumulator model and its account of choosing between /ɑ/ or /aː/. The top left panel shows the accumulator corresponding to the /ɑ/ response. The top right panel shows the accumulator corresponding to the /aː/ response. In the bottom panel we provide labels and descriptions for each LBA parameter.

Measuring Cognitive Processes From Behavioral Data

The LBA divides response times into decision times and non-decision times. The decision time is re-expressed as the amount of evidence needed for a response divided by the speed of evidence accumulation, which can be formally expressed as $\frac{b-k}{\xi}$. With the LBA we can go beyond the temporal aspects of the decision process by estimating drift rate. The drift rate is governed by the quality of evidence being sampled from the stimulus, with larger drift rates meaning faster speed of processing. We can also estimate non-decision time or response threshold and learn about the time needed for processes outside of the decision process or the response caution or bias associated with a participant's decision.

But how can we get unique estimates of non-decision time, drift rate, and response threshold from behavioral data? We can recover unique estimates because these parameters have unique behavioral signatures in response proportions and RT distributions (Ratcliff & McKoon, 2008). For example, the non-decision time parameter determines the smallest possible RT and also shifts the entire RT distribution, yet this parameter has no effect on the response proportions. Changes in mean drift rate cause small changes to the leading edge of the RT distribution – the fastest responses – and large changes to the tail – the slowest responses. In contrast, changes in response thresholds also shift the leading edge and cause relatively small changes to the tail. Higher mean drift rate leads to faster RTs in combination with a higher accuracy, whereas lower response thresholds lead to faster RTs with lower accuracy.

There are several benefits to using the LBA to analyze choice and RT data from a phoneme categorization task. First, the adjustment of response thresholds in the LBA is an intuitive account of the speed-accuracy trade-off, which cannot be addressed by

linear models of choice or RT. Second, rather than analyzing averaged data that can lead to researchers drawing incorrect conclusions (e.g., Ashby et al., 1994; Curran & Hintzman, 1995; Heathcote et al., 2000), the LBA can be used to analyze entire RT distributions for all responses. This analysis decomposes the behavioral data into their underlying constituent components of processing – such as drift rates, response thresholds, and non-decision times. Finally, the LBA reconciles the fact that continuous acoustic information leads to categorical decisions by explicitly describing phoneme decision-making as a result of an evidence accumulation process. Taken together, the LBA allows us to address both the cue-to-one-phoneme mapping and the cue-to-phoneme mapping issues outlined in the introduction. Specifically, we can learn about the cue-to-one-phoneme mapping by investigating the evidence accumulation dynamics in each accumulator, which represent processing for each phoneme response option. We learn about the cue-to-phoneme mapping by investigating how the parameters of the LBA model are affected by changes in acoustic information.

The Current Study

Spectral quality and duration are two of the acoustic cues that listeners can use to categorize vowels (e.g., Bohn & Flege, 1990; Flege, Bohn, & Jang, 1997; Adank, Van Hout, & Smits, 2004; Gerrits, 2001; Escudero, Benders, & Lipski, 2009; Reinisch & Sjerps, 2013). For example, first language (L1) English speakers weigh static spectral cues more than duration cues as they mostly use properties in the F1 and F2 to recognize the contrast between /iː/ and /ɪ/ and between /æ/ and /ɛ/ (Flege et al., 1997). On the other hand, second language (L2) speakers of English with L1 German mostly use the duration of the stimuli to recognize vowels in those same contrasts (Bohn & Flege, 1990). Similarly, L1 Dutch listeners weigh spectral cues heavier than

duration cues to distinguish between /ɑ/ and /aː/, whereas Turkish and Spanish L2 learners of Dutch weigh vowel duration heavier than spectral quality (Nooteboom & Cohen, 1984; van Heuven, Van Houten, & De Vries, 1986; Escudero et al., 2009; van der Feest & Swingley, 2011).

Here, we analyze data from an experiment in which L1 Dutch listeners categorize synthetic vowels as the Dutch short and closed /ɑ/ and the long and open /aː/. /ɑ/ and /aː/ are a useful set of stimuli as these vowels are typically realized with both spectral and duration differences (Adank et al., 2004). These two vowels are the only two low vowels in Dutch, which are each other's closest neighbors in the acoustic vowel space defined by F1 and F2 and alternate in the singular–plural pairs of some nouns (e.g., /pɑd/ - /pɑːdə/) as well as verbs (/kʋɑm/ - /kʋaːmə/).

We will subject the categorization and RT data to Bayesian logistic regression and Bayesian ANOVA (Rouder, Morey, Speckman, & Province, 2012), respectively. These more traditional analyses will attempt to replicate the relatively heavier weighting of vowel quality compared to the duration of a vowel, which is typically observed in categorization tasks involving the Dutch contrast between /ɑ/ and /aː/ (e.g., Nooteboom & Cohen, 1984; van Heuven et al., 1986; Escudero et al., 2009; van der Feest & Swingley, 2011; Reinisch & Sjerps, 2013). Then, we will analyze both RT and accuracy simultaneously with the LBA. The LBA analysis will test if spectral quality and duration affect the drift rates of participants when they are categorizing /ɑ/ and /aː/. We will also test if the effects of the acoustic cues on drift rates as well as the participants' response thresholds are different across the two vowels. Finally, and in line with Reinisch and Sjerps (2013), we will test if longer vowel durations systematically increase the time needed for perceptual encoding, which will manifest as longer

non-decision times for sounds with longer durations.

Method

Participants

Thirty participants were tested (5 males and 25 females) at the Radboud University Nijmegen, the Netherlands. All participants had normal or corrected to normal vision and reported no hearing impairments. They were native Dutch speakers. All participants received monetary reimbursement for their participation.

Materials

One-hundred different vowel stimuli were used. They were synthetic isolated vowels covering a 10×10 matrix ranging from the typical /ɑ/, with low formants and a short duration, to a typical /aː/, with high formants and a long duration. The F1 and F2 values as well as the duration of each of the 10 steps are presented in Table 1. The spectral quality of the typical /ɑ/ and /aː/ were based on production data from 50 male speakers (Pols, Tromp, & Plomp, 1973), while the duration values were based on 10 male speakers (Adank et al., 2004), following the stimulus creation procedure in Escudero et al. (2009). The sounds had a falling fundamental frequency from 150Hz to 100Hz, to simulate male speech.

The 100 stimuli used in the experiment were synthesized in the computer program Praat (Boersma, 2002). The difference between two consecutive steps on the duration dimension equals one-tenth of the difference between the logged duration (ms) of the typical /ɑ/ and /aː/. The difference between two consecutive steps on the spectral quality dimension equals one-tenth of the difference between the typical /ɑ/ and /aː/ spectral quality in mel. By generating stimuli in this way we can approach equal psychoacoustic scaling across the steps within each dimension, although not necessarily

across dimensions. We can compare cue weighting across dimensions as the dimension endpoints were based on typical values of the vowels produced in language.

Table 1
F1, F2, and duration values, which were crossed factorially to generate all 100 stimuli.

Step	Spectral Quality (F1/F2 Hz)	Duration (ms)
1	687/1099	96
2	699/1121	104
3	711/1143	113
4	723/1165	123
5	736/1188	134
6	749/1211	146
7	762/1235	158
8	775/1259	172
9	788/1283	187
10	801/1308	203

The experiment was run in a sound attenuated quiet booth using Dell Precision T3600 computers, with an Intel Xeon Processor E5-1620 and a Sound Blaster ZX Gamer audio card. Stimuli were played over Sennheiser HD 215 MKII DJ headphones.

Procedure

On each trial, participants heard a sound over headphones and saw the orthographic symbols for /ɑ/ (“a”) and /aː/ (“aa”) on the left and right of the computer screen. The position of the symbols corresponded to the response key side, which was fixed for each participant and counterbalanced across participants. Participants were required to respond by pressing one of the two response keys within 2500ms after stimulus onset, otherwise they were presented with ‘Te Langzaam’ (Too Slow), which remained on screen for 3000ms. Once a response was made the next trial

started immediately. The 100 unique stimuli (10 formant steps × 10 duration steps) were played once per block in randomized order. Participants heard 5 blocks, for a total of 500 trials. Participants were allowed to take a short break after each block and could continue to the next block when ready. The experiment started with 10 practice trials that only presented the stimulus with the lowest F1 and shortest duration, a typical /ɑ/, and the stimulus with the highest F1 and longest duration, a typical /aː/. The experiment was run in the software PsychoPy (Peirce, 2007).

Behavioral Data Analysis

Before analyzing the response proportion and RT data together with the LBA, we ran the more traditional analyses on both dependent measures separately. The proportions of /aː/ responses were analyzed with a Bayesian logistic regression model and the RTs were analyzed with a two-way Bayesian ANOVA. Both analyses included subjects as a random effect, meaning that each subject had their own intercept, which was drawn from a normal distribution. Post-hoc Bayesian paired-samples t-tests were conducted where appropriate. The logistic regression was carried out using the R Stan package (Stan Development Team, 2016) in R (R Development Core Team, 2016) and the ANOVA analyses was carried out using JASP (The JASP Team, 2016; Morey, Rouder, & Jamil, 2014; Rouder et al., 2012).

The main reason for using Bayesian linear models instead of frequentist alternatives is to allow calculation of Bayes factors (BF), which we motivate below. In addition, Bayesian models also allow for researchers to calculate posterior distributions of parameters. Posterior distributions provide a range of plausible values as well as their corresponding probabilities. They have a similar purpose to standard errors of estimation, but do not make the assumption that estimation error is symmetrical and

normally distributed. This assumption is often incorrect, especially when dealing with data that are themselves not normally distributed.

For the behavioral data analysis (with the exception of the logistic regression), Bayes factors were used in place of conventional p values, as Bayes factors are arguably more appropriate for assessing statistical evidence (see Wagenmakers, 2007). We refer the interested readers to Kruschke (2011) and Lee and Wagenmakers (2013) for accessible introductions to Bayesian statistics for social scientists. Bayes factors represent "the primary tool used in Bayesian inference for hypothesis testing and model selection" (Berger, 2006, p. 378). They quantify evidence in favor of both the null hypothesis or the alternative hypothesis as a ratio. For example, when $BF_{10} = 5$ the observed data are 5 times more likely under the alternative hypothesis than under the null hypothesis. When $BF_{10} = .2$ the observed data are 5 times more likely under the null hypothesis than under the alternative hypothesis. To determine the evidence for a particular effect (e.g., spectral quality), we calculated an inclusion BF ($BF_{\text{Inclusion}}$). The $BF_{\text{Inclusion}}$ statistic represents the evidence in favor of models that include a particular effect in relation to models that do not include the effect. If we were testing whether spectral quality had an effect on RT, for instance, and we obtained $BF_{\text{Inclusion}} = 5$, then the data are 5 times more likely to come from a model with a spectral quality effect than a model without a spectral quality effect. Bayes factors greater than 3 or less than 1/3 will be considered positive evidence for the alternative and null hypothesis, respectively (Kass & Raftery, 1995). And Bayes factors less than 3 or greater than 1/3 will be considered inconclusive evidence for both hypotheses.

The LBA models, which we describe next, and the logistic regression models were evaluated on how well they predict future observations – the out-of-sample predictive

error. The gold standard for estimating the out-of-sample predictive error of a model is cross-validation (Geisser & Eddy, 1979). Cross-validation is computationally expensive and so we used a computationally faster approximation: the widely applicable information criterion (WAIC; Watanabe, 2010; Gelman, Hwang, & Vehtari, 2014). The WAIC balances goodness of fit against model complexity. This measure is calculated from a model fit value and a model complexity penalty value, which approximates the number of effective parameters of the model. In this sense, WAIC is similar to the BIC (Schwarz, 1978) and AIC (Akaike, 1974) measures, but WAIC extends these by quantifying model complexity as across-sample variability in model fit rather than simply counting up the number of free parameters. The method we used is described in detail by Vehtari, Gelman, and Gabry (2016). The model with the lower WAIC value has better out-of-sample predictive error and is therefore the preferred model.

Linear Ballistic Accumulator Analysis

We used hierarchical Bayesian methods to estimate the parameters of the LBA model. The model fitting details are outlined in the supplementary materials, which are available online at this paper's associated Open Science Framework page `https://osf.io/hp9xt/`.

Here we give a brief description of the different LBA models we implemented. Each LBA model had two accumulators, one corresponding to /ɑ/ and another corresponding to /aː/. Each model had the parameters mean drift rate v, response threshold b, starting point variability A, non-decision time t_0, and drift variability s. The s parameter serves as the scaling parameter, which we set to 1 for the /ɑ/ accumulator and estimated for the /aː/ accumulator. Having a scaling parameter allows all other parameter values to be identified, because their values are now relative to the

scaling parameter (Donkin, Brown, & Heathcote, 2009). All other parameters were estimated, but fixed across accumulators, with the exception of v and response threshold b.

To help describe all models considered, we present an LBA visualization in Figure 2. In the Figure, blue arrows represent changes in drift rate across the duration values, green arrows represent changes in drift rate across the spectral values, red arrows represent changes in non-decision time across the duration values, and grey arrows represent changes in response threshold across vowels.

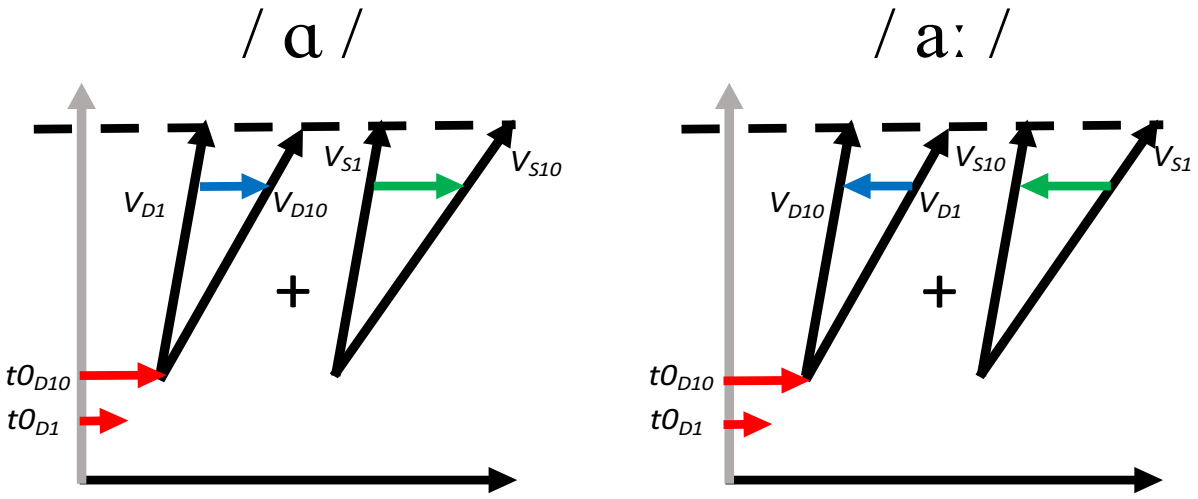


Figure 2. An example of a two-accumulator LBA model. The left panel shows the accumulator corresponding to the /ɑ/ response. The right panel shows the accumulator corresponding to the /aː/ response. In each accumulator we present 4 sloped lines that represent mean drift rates for the 1^{st} and 10^{th} duration values and 1^{st} and 10^{th} spectral values. The changes in slope correspond to changes in mean drift rate. The changes in mean drift that are induced by duration manipulations are depicted by the blue arrows, with arrows pointing in the direction of the change. Green arrows depict the spectral quality effects on mean drift rate. Red arrows show the changes in non-decision time that are due to different vowel durations – the further to the right the arrow extends the longer the non-decision time. Grey arrows show the response threshold heights for each accumulator.

The first model we tested, the equal drift model, allowed v to change across the spectral quality values and duration values. This model assumed that spectral quality and duration influenced the speed of information processing, but their effects were equal

for /ɑ/ and /aː/ responses. In Figure 2, we would see that blue arrows are of equal length, green arrows are of equal length, red arrows are of equal length, and grey arrows are of equal length.

The second model, the unequal drift model, allowed v to vary across responses in addition to spectral quality values, and duration values. Like the equal drift model, this model assumed that spectral quality and duration influence the speed of information processing, however this effect was not fixed to be equal across /ɑ/ and /aː/ responses. In Figure 2, we would see that blue arrows are not equal length and green arrows are not equal length, but red arrows and grey arrows are still equal length. This model is theoretically interesting because it tests how much each cue contributes to the recognition of /ɑ/ or /aː/ individually. In contrast, traditional cue weighting measures based on categorization data only allow researchers to infer about the contribution of duration or spectral quality cues to the overall vowel contrast, an approach that is captured by the equal drift model.

The third model we tested, the non-decision time model, was the same as the unequal drift model with one extension. The non-decision time model also had 10 separate t_0 parameters, one for each of the duration values. In Figure 2, we would see that blue arrows are not equal length, green arrows are not equal length, and the red arrows are not equal length, but grey arrows are of equal length. If longer stimulus durations delay the processing of duration information, since the participant must wait for the vowel to offset, then we should observe that longer stimulus durations induce systematic increases in non-decision processing time. Having 10 separate t_0 parameters for each duration value allows for longer durations to induce longer non-decision times. If we find evidence in favor of this model then this suggests that lengthening the vowel

may induce longer perceptual encoding times (Reinisch & Sjerps, 2013).[4] However, if we find evidence in favor of another model, without different t_0 parameters, then this suggests that longer durations do not reliably delay the processing of duration information.

The final model we tested, the response bias model, was the same as the unequal drift model, but it allowed response thresholds to change across accumulators. This allowed the model to account for potential response bias in the data, i.e., responding /ɑ/ more often than /aː/ overall. Response bias is typically captured in response thresholds, unless the locus of bias is related to the stimulus (e.g., perceptual decision criterion, see White & Poldrack, 2014). In Figure 2, we would see that blue arrows are not equal length, green arrows are not equal length, and grey arrows are not equal length, but red arrows are equal length.

Drift rates for each /aː/ response were defined as

$$v_{\mathrm{a\!:}SD} = v_{\mathrm{a\!:}} + \beta_{\mathrm{a\!:}S} X_S + \beta_{\mathrm{a\!:}D} X_D \tag{1}$$

where $v_{\mathrm{a\!:}}$ represents the base drift rate for the /aː/ response. X_S corresponds to the ${}_Sth$ value spectral quality and X_D corresponds to the ${}_Dth$ value duration. The symbols X_S and X_D denote the stimulus spectral quality and duration steps, respectively (i.e., not the raw values in Hz/ms, but ordinal scale values from 1-10 shown in column 1 of Table 1). $\beta_{\mathrm{a\!:}S}$ and $\beta_{\mathrm{a\!:}D}$ denote parameters that describe the effect of the spectral and duration changes on drift rate for the /aː/ response, respectively.

Drift rates for each /ɑ/ response were defined as

[4]Longer non-decision times could also mean longer motor response times, but we cannot disentangle effects of perceptual encoding time and motor response time. Here, we assume that longer durations do not have systematic effects on motor response time.

$$v_{\text{ɑ}SD} = v_{\text{ɑ}} + \beta_{\text{ɑ}S}(11 - X_S) + \beta_{\text{ɑ}D}(11 - X_D) \quad (2)$$

where $v_{\text{ɑ}}$ represents the base drift rate for the /ɑ/ response. $\beta_{\text{ɑ}S}$ and $\beta_{\text{ɑ}D}$ denote parameters that describe the effect of the spectral and duration changes on drift rate for the /ɑ/ response, respectively. Note that for the equal drift model, the $\beta_{\text{ɑ}S}$ and $\beta_{\text{ɑ}D}$ are the same as the $\beta_{\text{aː}S}$ and $\beta_{\text{aː}D}$ estimates. The terms $(11 - X_S)$ and $(11 - X_D)$ ensure that higher spectral quality and duration conditions, which correspond to atypical values for ɑ, produce smaller drift rates compared to lower spectral quality and duration conditions.

By estimating drift rate via Equations 1 and 2 we obtain four drift rate coefficient estimates – $\beta_{\text{aː}S}$, $\beta_{\text{aː}D}$, $\beta_{\text{ɑ}S}$ and $\beta_{\text{ɑ}D}$ – which represent the effect that our experimental cue manipulations have on the drift rate parameter. Note that Equations 1 and 2 produce linear effects on drift rate, but linear increases in drift rate also allow for non-linear effects on behavioral data (see e.g., van Ravenzwaaij, Brown, & Wagenmakers, 2011, Fig. 2).

We compared four versions of our LBA model: the equal drift model, the unequal drift model, the non-decision time model, and the response bias model using WAIC. In addition, we used Bayesian predictive p-values to statistically test for differences between the posterior distributions of the drift coefficient parameters (Meng, 1994). We used a criterion of .05 to determine if two posterior distributions are overlapping. Specifically, we calculated the difference between the two spectral quality posteriors, the two duration posteriors, the spectral /aː/ and the duration /aː/ posteriors, and the spectral /ɑ/ and the duration /ɑ/ posteriors. We calculated the probability (p-value) that the resulting difference distributions were equal to or less than 0. A p-value "is a

measure of discrepancy between the observed data and the posited assumptions, among which the hypothesis being tested is only a part" (Meng, 1994, p. 1144). Thus, similar to the traditional p-value, a low predictive p-value indicates a low probability of observing this or more extreme data if the null hypothesis were true.

Results

The R code for all analyses and experimental data are available online at this paper's associated Open Science Framework page `https://osf.io/hp9xt/`.

Behavioral Data Results

We measured the choice proportion of participants for the two typical vowels: the minimum-duration step 1 and minimum-spectral step 1 stimuli (expected response was /ɑ/) as well as the maximum-duration step 10 and maximum-spectral step 10 stimuli (expected response was /aː/). Overall, the expected response was made 97% of the time, which suggests there was high consensus for the extreme stimuli. The percentage of expected responses ranged from 70% to 100% across participants. No participants were excluded from the analysis. Overall, there was a response bias as participants responded /ɑ/ 55% of the time, which we explore further in the LBA analysis.

Figure 3 shows the effects of both the duration and spectral manipulations as a heat map. The top right corner shows stimuli with typical /aː/ cue information (maximum-duration 10 and maximum-spectral 10) and the bottom left corner shows stimuli with typical /ɑ/ cue information (minimum-duration step 1 and minimum-spectral step 1). This plot shows the rate of change from responding /ɑ/ to responding /aː/ as a function of both the duration and spectral quality manipulations. As we move along the y-axis we can observe the change across spectral quality and as we move along the x-axis we can see the change across duration. The diagonal of the

graphic visualizes the listeners' perceptual boundary.

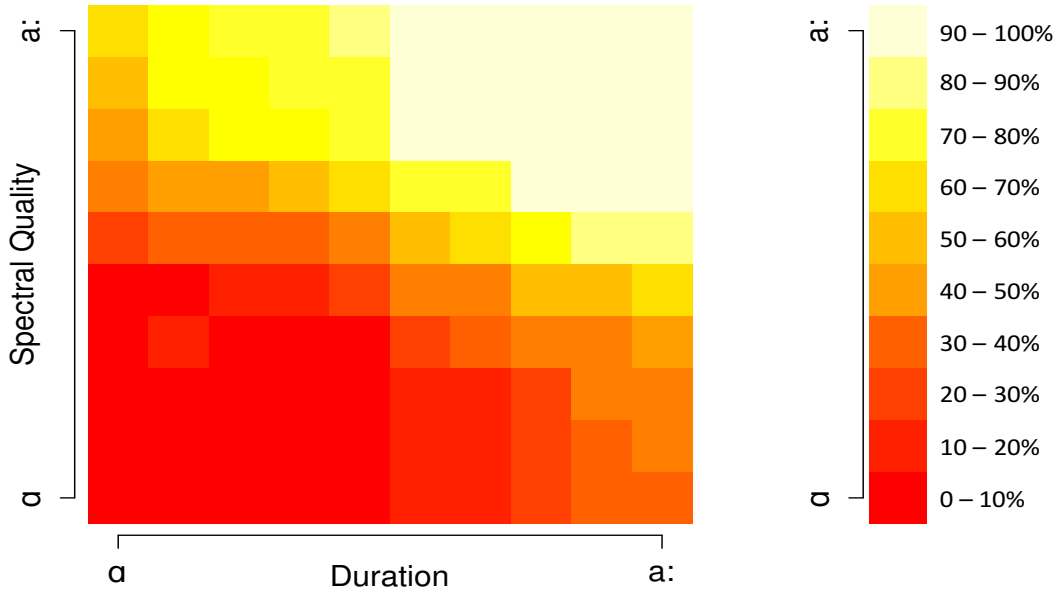


Figure 3. A heat map showing the overall effects of duration and spectral quality on categorization. The y-axis displays the spectral steps and the x-axis displays the duration steps, where each square represents a step. The text in the legend of this figure contains the percentage that participants responded /aː/ overall. The top right corner shows that participants predominantly respond /aː/ to stimuli with typical /aː/ values. The bottom corner shows that participants predominately response /ɑ/ to stimuli with typical /ɑ/ values.

Figure 4 shows the effects of spectral quality and duration on mean RT as a heat map. The top right corner shows stimuli that have typical /aː/ cue information and the bottom left corner shows stimuli that have typical /ɑ/ cue information.

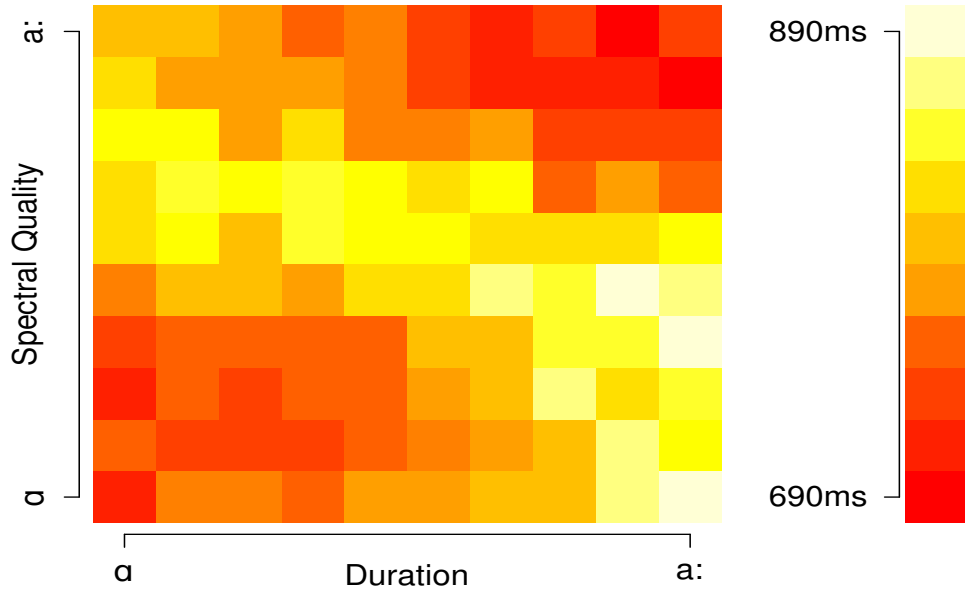


Figure 4. A heat map showing the effects of cue manipulation on RT, which has been collapsed over response. The y-axis displays the different spectral steps and the x-axis displays the different duration steps. The top right corner shows fast RTs for stimuli with typical /aː/ values. The bottom left corner shows fast RTs for stimuli with typical /ɑ/ values.

We determined the effect of spectral cues and duration cues on categorization by subjecting the proportion of /aː/ responses to a Bayesian logistic regression. One coefficient of the regression corresponds to the spectral changes (β_{Spectral}) and another to the duration changes (β_{Duration}), and these represent each cue's influence on categorization. We regressed choice proportion on duration, spectral quality, and the interaction between the two. The model with main effects for both duration and spectral quality and an interaction between the two had the lowest WAIC (12965.2). The null model with no effects, the duration only, spectral quality only, and the model with both main effects only had WAIC values of 20548.4, 19210.4, 15059.9, and 12970.5, respectively.

The spectral manipulation had a larger effect on categorization than the duration manipulation. Each step increase in duration increased the log odds of responding /aː/ by $\beta_{\text{Duration}} = .403$ [95% Credible Interval: 0.389, 0.449]. The percentage of /aː/

responses for the 1^{st} duration step was 22.9% and this increased to 66.4% for the 10^{th} duration step (t-test: $BF_{10} = 1.24\times10^{7}$). Each step increase in spectral quality increased the log odds of responding /aː/ by $\beta_{\text{Spectral}} = .664$ [95% Credible Interval: 0.650, 0.709]. The percentage of /aː/ responses for the 1^{st} spectral quality step was 11.7% and this increased to 84% for the 10^{th} spectral quality step (t-test: $BF_{10} = 2.66\times10^{11}$). The interaction between spectral quality and duration showed that for the 1^{st} duration step, the percentage of /aː/ responses increased from 20% to 58% from the 1^{st} to the 10^{th} spectral quality step (t-test: $BF_{10} = 1.20\times10^{7}$). For the 10^{th} duration step the percentage of /aː/ responses increased from 28% to 96% from the 1^{st} to the 10^{th} spectral quality step (t-test: $BF_{10} = 2.95\times10^{7}$).

We regressed RTs on duration, spectral quality, response, and all interactions and found main effects for response ($BF_{\text{Inclusion}} > 10^{15}$), duration ($BF_{\text{Inclusion}} = 2.42\times10^{11}$), and spectral quality ($BF_{\text{Inclusion}} > 10^{15}$). Moreover, the model included an interaction between spectral quality and response ($BF_{\text{Inclusion}} > 10^{15}$), duration and response ($BF_{\text{Inclusion}} = 303$), and spectral quality and duration ($BF_{\text{Inclusion}} = 112$). RTs for the 1^{st} spectral quality step were faster than RTs for ambiguous spectral quality steps, such as 5 (t-test: $BF_{10} = 2.16$) or 6 (t-test: $BF_{10} = 10.99$), but the difference in RTs between the 1^{st} and 5^{th} spectral quality step is inconclusive. RTs for the 10^{th} spectral quality step were faster than RTs for the 5^{th} (t-test: $BF_{10} = 42.98$) or the 6^{th} spectral quality step (t-test: $BF_{10} = 165.95$). RTs for the 1^{st} duration step were equal to RTs for ambiguous duration steps, such as 5 (t-test: $BF_{10} = .203$) or 6 (t-test: $BF_{10} = .226$). RTs for the 10^{th} duration were slower than RTs for the 5^{th} (t-test: $BF_{10} = 1.88$) and 6^{th} (t-test: $BF_{10} = .80$) duration step, but the evidence for these differences was inconclusive.

As shown in the left panel of Figure 5, there was a crossover interaction between

spectral quality and response. The evidence for differences in RT between the /ɑ/ and /aː/ responses for the 1^{st} spectral quality was inconclusive (t-test: $BF_{10} = 1.05$). But, RTs for /aː/ responses were faster than /ɑ/ responses for the 10^{th} spectral quality (t-test: $BF_{10} = 64.47$). As shown in the right panel of Figure 5, RTs for /ɑ/ responses were faster than RTs for /aː/ responses for the 1^{st} duration step, but the evidence was inconclusive (t-test: $BF_{10} = 2.40$). In contrast, RTs for /ɑ/ responses were slower than RTs for /aː/ responses for the 10^{th} duration step (t-test: $BF_{10} = 14.63$). RTs for /aː/ responses were slower for the 5^{th} duration step (t-test: $BF_{10} = 0.949$) and faster for the 6^{th} duration step (t-test: $BF_{10} = 0.899$), but the evidence was inconclusive for both. There were no differences in RT between the /aː/ and /ɑ/ responses for the 5^{th} (t-test: $BF_{10} = 0.19$) and 6^{th} (t-test: $BF_{10} = 0.29$) duration steps. The absolute difference in RTs for the 1^{st} and 10^{th} spectral qualities was larger for stimuli with longer duration values (114ms) than shorter duration values (108 ms), but the evidence was ambiguous (t-test: $BF_{10} = 1.38$).

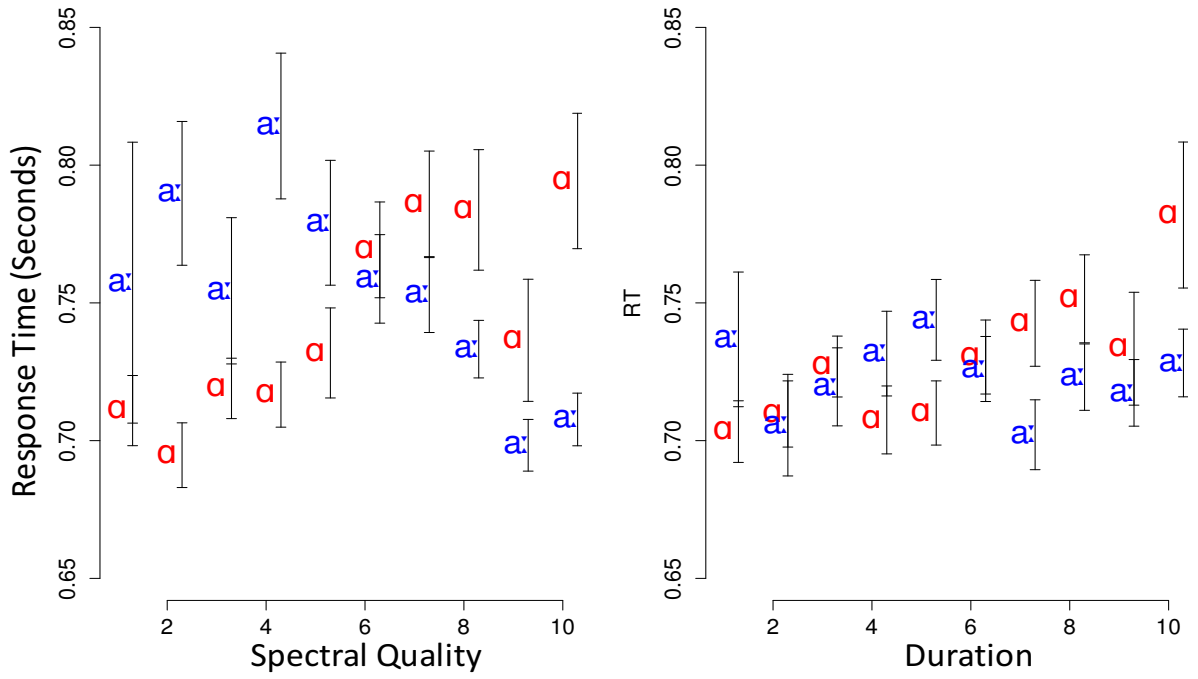


Figure 5. The interaction between the spectral quality and response (left) and duration and response (right). The /ɑ/ responses are plotted in red and the /aː/ responses are plotted in blue. Interval bars represent 1 standard error of the mean.

Linear Ballistic Accumulator Results

We found that drift rates (speed of processing) were affected more by spectral quality than by duration. We also found that the effects of spectral quality on speed of processing were not equal for both for phoneme responses. Given the findings of McMurray et al. (2008) and Reinisch and Sjerps (2013), who found that cues that appeared later in the signal affected eye-tracking behavior later in a trial, we expected longer non-decision times for stimuli with longer vowel durations – in particular, because participants were assumed to not be processing duration information until the vowel offset. However, our analysis showed that increased duration of vowels did not produce any systematic increases in non-decision processing time.

Specifically, to compare the equal drift, unequal drift, and non-decision time models we assessed which model was selected by WAIC. The unequal drift model (WAIC = 629.2) was preferred over the equal drift model (WAIC = 918.7), the

non-decision time model (WAIC = 1083.2), and the response bias model (WAIC = 946.9). The unequal drift model also provided good fits to the empirical data, which are presented in the supplementary materials.

Figure 6 shows the group level mean posteriors of spectral and duration drift coefficients from the unequal drift model. Bayesian p-values suggest that changes in spectral cues induced larger changes in drift rate than duration. The effects of spectral quality on drift rates are not identical for both vowels, where changes in spectral quality affect drift rates more for /aː/ than /ɑ/ responses. The effects of duration on drift rate are equal for both vowels.

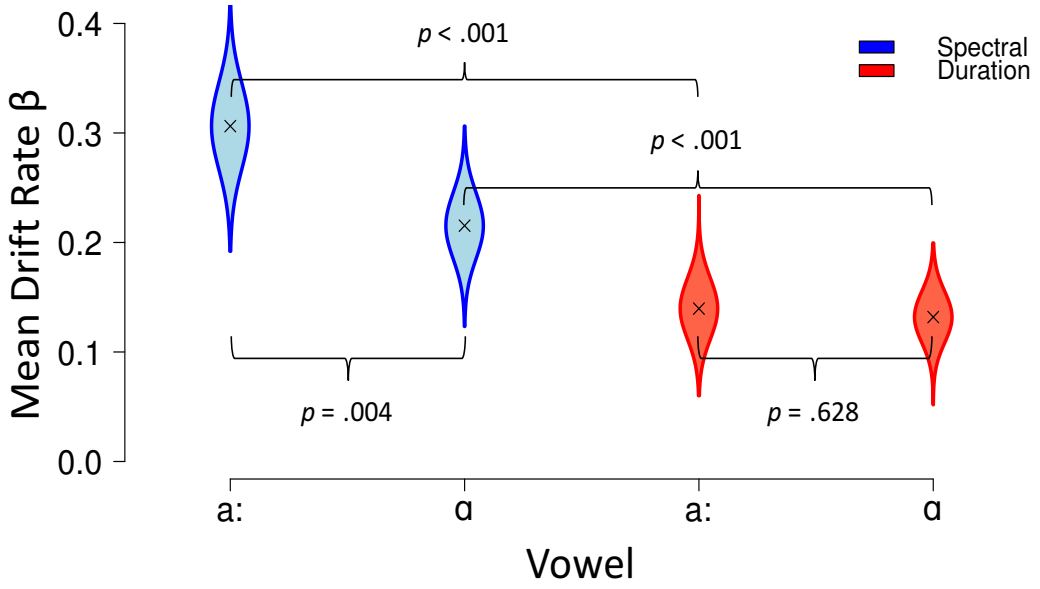


Figure 6. Group-level mean posterior distributions for mean drift rate β for both the /ɑ/ and /aː/ responses.

The group-level mean posteriors for all other LBA parameters are shown in Table 2. The base drift rate for /ɑ/ is higher than for /aː/. This explains the overall bias we observed, where participants responded /ɑ/ 55% of the time. We tested whether the bias is in response thresholds, instead of drift rate, by fitting the response bias model with separate thresholds for each vowel. The response bias model performed worse compared to the unequal drift model. Therefore the bias is better captured in drift rate.

Table 2
Group-level Mean Posteriors.

Parameter	Median	Credible Interval (2.5%, 97.5%)
b	2.125	(2.417, 2.720)
A	0.953	(0.841, 1.062)
s	0.911	(0.866, 0.958)
t_0	153ms	(61ms, 192ms)
$v_{\text{ɑ}}$	2.70	(2.41, 2.98)
$v_{\text{aː}}$	2.37	(2.10, 2.65)
$\beta_{\text{aː}S}$	0.306	(.252, .360)
$\beta_{\text{aː}D}$	.140	(.103, .176)
$\beta_{\text{ɑ}S}$	.215	(.174, .256)
$\beta_{\text{ɑ}D}$	.132	(.100, 1.64)

The combined effect of base drift rate and spectral quality drift coefficients can be better understood by looking at the mean drift rates in each of the 10 spectral quality steps. To calculate these mean drift rates we inserted the group-level mean posterior median (see Table 2) for the base drift rate and drift coefficient parameters into Equations 1 and 2. This resulted in the mean drift rates in each of the 200 conditions. To arrive at the mean drift rates in each of the 10 spectral quality steps we averaged over the duration conditions. Zooming in on the different effects of drift rate across spectral quality for /ɑ/ and /aː/ (cf. blue posterior distributions in Figure 6), Figure 7 shows that for stimuli with atypical spectral quality values, which is the 1^{st} step for /aː/ and the 10^{th} step for /ɑ/, participants had higher drift rates for /ɑ/ compared to /aː/. The difference between mean drift rates between the two vowels decreases as the stimuli approach the typical spectral quality.

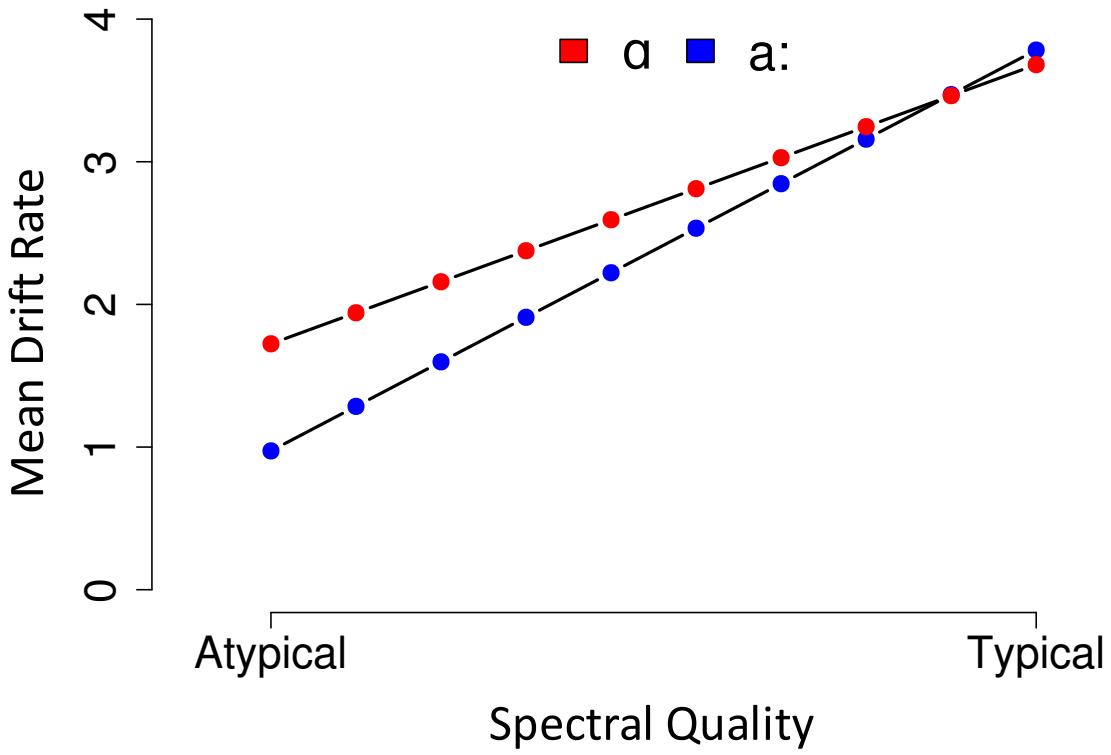


Figure 7. Mean drift rates across the 10 spectral qualities for /ɑ/ (red) and for /aː/ (blue). Mean drift rates are shown along the y-axis. Along the x-axis are the 10 different spectral quality values ranging from the atypical values, which is the 1^{st} value for /aː/ and the 10^{th} value for /ɑ/, to the typical values, which is the 10^{th} value for /aː/ and the 1^{st} value for /ɑ/.

Discussion

In this study, Dutch listeners categorized sounds as /ɑ/ and /aː/ in an experiment in which we manipulated the spectral and duration properties of the sounds. Both the categorization and the response times (RTs) of participants were recorded and independently analyzed using Bayesian linear models. We then applied a mathematical model of decision-making – the linear ballistic accumulator model (LBA; Brown & Heathcote, 2008) – which analyzed both streams of data simultaneously. The model allowed us to investigate how changes in acoustic cues affect latent cognitive processes that underpin phoneme decisions.

In terms of behavioral data, we were able to replicate the relatively heavy spectral

cue weighting compared to duration that is typically observed in vowel categorization tasks involving the Dutch contrast between /ɑ/ and /aː/ (Nooteboom & Cohen, 1984; van Heuven et al., 1986; Escudero et al., 2009; van der Feest & Swingley, 2011; Reinisch & Sjerps, 2013). The LBA analysis addressed fundamental issues that arise when using categorization data to learn about the mapping between acoustic cues and phonemes. The model posits an evidence accumulation process, which helps explain how continuous acoustic information can result in discrete phoneme decisions. With the LBA we observed that changes in spectral and duration cues lead to changes in drift rates (i.e., speed of information processing) for both /ɑ/ and /aː/ responses. In addition, changes in spectral quality had larger effects on the behavioral data than changes in duration and this was driven by differences in speed of information processing. Furthermore, when the stimuli had more atypical spectral qualities, which is a short duration and darker spectral quality for /aː/ or a long duration and clearer spectral quality for /ɑ/, listeners accumulate evidence faster for /ɑ/ responses compared to /aː/. This asymmetry in processing speed explains why listeners' responded /ɑ/ 55% of the time.

We also argued that if duration processing can only start at vowel offest, as suggested by Reinisch and Sjerps (2013), we would observe relatively longer perceptual encoding times for longer vowel durations. Non-decision time in the LBA is made up of the time needed to perceptually encode the stimuli and execute a motor response. If we assume that longer sound durations have no systematic effects on motor response times, then changes in non-decision time would be due to differences in perceptual encoding time. However, we found that different vowel durations did not affect listener's non-decision processing time.

What Precedes an Evidence Accumulation Process?

With the LBA we provide part of the picture of how continuous acoustic information leads to discrete phoneme decisions. In the past, researchers used categorization data to look at associations between cues before-processing and overt responses, which overlooks the decision processes required for categorization. With the LBA we specifically investigate the decision processes, showing that the relatively heavy weighting of spectral quality compared to duration is not merely a result of the timing of cue availability, but a property of the decision-making process.

While the LBA analysis presented here is a considerable advance on traditional separate analyses of response proportion and RT, it does not explicitly explain how continuous acoustic information becomes evidence that drives the decision process. Recall that the drift rate of the decision process is considered a measure of the quality of the evidence. But what is the evidence to this accumulation process? One possibility is to complement the LBA with a formal model of a cue mapping process that takes the continuous acoustic information and maps it onto exemplar clusters or discrete representations of phonetic categories. The output of this cue mapping model could serve as evidence that inputs into the LBA processes (see Ratcliff, Gomez, & McKoon, 2004, for a similar concept in lexical decision-making). The drift rate in the LBA would then be a measure of the speed, efficiency, or certainty with which a certain acoustic cue is mapped onto a phoneme category. In summary, the continuous acoustic information enters the cognitive system, the acoustic information then maps onto a phonetic category, the mapping processes outputs evidence that enters an LBA process, which accumulates evidence until a response threshold is reached and an overt response is made.

But what could cause differences in speed of information processing? In our case, we found behavioral evidence that spectral quality is weighed heavier than duration for the /aː/ and /ɑ/ contrast. The LBA analysis showed that changes in spectral quality cause larger changes in speed of processing than changes in duration. Perhaps the differences in drift rate were due to spectral quality cues being mapped more efficiently or more certainly onto the categories /aː/ and /ɑ/ compared to duration cues. Similarly, the higher spectral quality drift rate for /aː/ than for /ɑ/ suggests that the mapping between spectral quality information and /aː/ provides better evidence than the mapping between spectral quality and /ɑ/. Finally, the higher drift rate for atypical values of /ɑ/ than for atypical values of /aː/ suggests that the mapping between atypical acoustic cues and categories is more efficient or more certain for /ɑ/ than it is for /aː/. Asymmetric mappings between cue values and categories could be instrumental in explaining asymmetries in vowel perception (Polka & Werker, 1994), as well as provide a phonetic basis for the notion of phonological under-specification (Lahiri & Reetz, 2010).

Future Directions

In this study, we investigated duration and spectral quality, which are cues that are processed separately and by different pathways in the brain (e.g., Zatorre & Belin, 2001). The LBA model is not constrained by this independence of acoustic cues. For instance, researchers could examine phoneme decisions about sounds that are cued by two spectral dimensions (F1 and F2) rather than one spectral and one temporal dimension.

Moreover, the LBA can be used to model how listeners deal with correlated acoustic dimensions, such as F1 and inherent vowel duration (House & Fairbanks, 1953;

Peterson & Lehiste, 1960; Lehiste & Lass, 1976) or F2 and spectral tilt for /i/ and /u/ (Ito, Tsuchida, & Yano, 2001). These correlations can be modeled by the LBA by changing the way parameters vary across conditions. For instance, if spectral quality and duration were positively correlated, Equation 1 and 2, which estimate drift rates for each condition, could be extended by including an interaction term. The mean drift rates for each /aː/ response could be defined as

$$v_{\text{aː}SD} = v_{\text{aː}} + \beta_{\text{aː}S} X_S + \beta_{\text{aː}D} X_D + \beta_{\text{aː}SD} X_S X_D \tag{3}$$

where $\beta_{\text{aː}SD}$ denotes the parameter that describes the interaction effect of the spectral and duration changes on mean drift rate for the /aː/ response. A formal model comparison (i.e., comparing a model fit with and without the interaction terms in their ability to fit the data) can shed light on the way drift rates change with both dimensions. Note that this approach allows one to investigate the effect of both dimensions independently, even though they covary in practice. The LBA analysis with the added interaction terms can be incorporated for phonetic distinctions cued by one dimension, two dimensions, three dimensions, or more by augmenting Equation 3 with the relevant terms.

The LBA could also be used to address other long standing questions in speech perception. For example, the literature so far does not contain a definitive explanation of an asymmetry first observed by Nooteboom and Doodeman (1980, p. 277): reducing the duration of /aː/ can lead to listeners responding 100% /ɑ/; yet, increasing the duration of /ɑ/ does not lead to a consistent /aː/ response. Van der Feest and Swingley (2011) proposed two possible explanations of this phenomenon. Firstly, lengthening sounds might show weaker effects because it occurs in natural language as prosodic

effects, such as the application of emphatic stress (Ko, Soderstrom, & Morgan, 2009). Secondly, lengthening sounds may facilitate perceptual access to vowel quality, whereas the spectral quality in shortened vowels may be harder to evaluate, leading to reliance on duration for short vowels. The LBA could model the second explanation by letting the drift rate for spectral quality change within a trial as the stimulus duration increases. Fortunately, a non-constant drift rate over the course of the trial is something that is possible to explore in the LBA (Holmes, Trueblood, & Heathcote, 2016).

Allowing non-constant drift rate within a trial opens the door to investigating how time-variant acoustic cues cause changes in LBA parameters that also change over time. For example, the cues to pre-voiced /b/ become available in a sequence: first pre-voicing, which provides information about phonological voicing, then the burst, which provides information about the place of articulation and voicing, and then the formant trajectories into the vowel, which listeners rely on most to determine place of articulation. Another example of a time-variant acoustic cue is vowel-inherent spectral change (Morrison, 2013). These dynamic cues can by mapped onto latent cognitive processes by linking them to dynamic LBA parameters. Dynamic LBA parameters may be able to disentangle the effects of cues that are available earlier (in time) in the speech from cues that are simply mapped more strongly to phonetic categories.

Overall, our study demonstrates a successful and novel application of evidence accumulation models to phoneme categorization tasks. These models allow researchers to investigate latent cognitive processes by analyzing behavioral data. Given the merit of using evidence accumulation models, we hope to see them applied to other research questions embedded in the speech perception literature, some of which we have outlined here.

References

Adank, P., Van Hout, R., & Smits, R. (2004). An acoustic description of the vowels of northern and southern standard dutch. *The Journal of the Acoustical society of America*, *116*(3), 1729–1738.

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, *19*(6), 716–723.

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of memory and language*, *38*(4), 419–439.

Ashby, F. G., Maddox, W. T., & Lee, W. W. (1994). On the dangers of averaging across subjects when using multidimensional scaling or the similarity-choice model. *Psychological Science*, *5*(3), 144–151.

Berger, J. O. (2006). Bayes factors. *Encyclopedia of statistical sciences*.

Boersma, P. (2002). Praat, a system for doing phonetics by computer. *Glot international*, *5*(9/10), 341–345.

Boersma, P. (2007). Some listener-oriented accounts of h-aspiré in french. *Lingua*, *117*(12), 1989–2054.

Bohn, O.-S., & Flege, J. E. (1990). Interlingual identification and the role of foreign language experience in l2 vowel perception. *Applied Psycholinguistics*, *11*(03), 303–328.

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: linear ballistic accumulation. *Cognitive psychology*, *57*(3), 153–178.

Cassey, P., Heathcote, A., & Brown, S. D. (2014). Brain and behavior in decision-making. *PLoS Computational Biology*, *10*(7), 1-8.

Curran, T., & Hintzman, D. L. (1995). Violations of the independence assumption in process dissociation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(3), 531-547.

Donkin, C., & Brown, S. D. (in press). Response time modeling. In T. Wixted & E.-J. Wagenmakers (Eds.), *The stevens' handbook of experimental psychology and cognitive neuroscience* (4th ed., Vol. 5). Wiley.

Donkin, C., Brown, S. D., & Heathcote, A. (2009). The overconstraint of response time models: Rethinking the scaling problem. *Psychonomic Bulletin & Review*, *16*(6), 1129–1135.

Eidels, A., Donkin, C., Brown, S. D., & Heathcote, A. (2010). Converging measures of workload capacity. *Psychonomic Bulletin & Review*, *17*(6), 763–771.

Escudero, P., Benders, T., & Lipski, S. C. (2009). Native, non-native and l2 perceptual cue weighting for dutch vowels: The case of dutch, german, and spanish listeners. *Journal of Phonetics*, *37*(4), 452–465.

Evans, N. J., & Brown, S. D. (2016). People adopt optimal policies in simple decision-making, after practice and guidance. *Psychonomic Bulletin & Review*, 1–10.

Flege, J. E., Bohn, O.-S., & Jang, S. (1997). Effects of experience on non-native speakers' production and perception of english vowels. *Journal of phonetics*, *25*(4), 437–470.

Forstmann, B. U., Brown, S. D., Dutilh, G., Neumann, J., & Wagenmakers, E.-J. (2010). The neural substrate of prior information in perceptual decision making: a model-based analysis. *Frontiers in Human Neuroscience*, *4*(40), 1-12.

Forstmann, B. U., Dutilh, G., Brown, S., Neumann, J., Von Cramon, D. Y., Ridderinkhof, K. R., & Wagenmakers, E.-J. (2008). Striatum and pre-sma facilitate decision-making under time pressure. *Proceedings of the National Academy of Sciences*, *105*(45), 17538–17542.

Geisser, S., & Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, *74*(365), 153–160.

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for bayesian models. *Statistics and Computing*, *24*(6), 997–1016.

Gerrits, E. (2001). *The categorisation of speech sounds by adults and children: a study of the categorical perception hypothesis and the development weighting of acoustic speech cues* (Unpublished doctoral dissertation). Utrecht University.

Hawkins, G. E., Marley, A., Heathcote, A., Flynn, T. N., Louviere, J. J., & Brown, S. D. (2013). Integrating cognitive process and descriptive models of attitudes and preferences. *Cognitive science*, *38*(4), 701-35.

Hawkins, G. E., Marley, A., Heathcote, A., Flynn, T. N., Louviere, J. J., & Brown, S. D. (2014). Integrating cognitive process and descriptive models of attitudes and preferences. *Cognitive science*, *38*(4), 701–735.

Heathcote, A., Brown, S. D., & Mewhort, D. J. K. (2000). The power law repealed: The case for an

exponential law of practice. *Psychonomic Bulletin and Review*, *7*, 185-207.

Heitz, R. P. (2014). The speed-accuracy tradeoff: history, physiology, methodology, and behavior. *Frontiers in neuroscience*, *8*, 150.

Ho, T. C., Brown, S. D., & Serences, J. T. (2009). Domain general mechanisms of perceptual decision making in human cortex. *The Journal of Neuroscience*, *29*(27), 8675–8687.

Holmes, W. R., Trueblood, J. S., & Heathcote, A. (2016). A new framework for modeling decisions about changing information: The piecewise linear ballistic accumulator model. *Cognitive psychology*, *85*, 1–29.

Holt, L. L., & Lotto, A. J. (2006). Cue weighting in auditory categorization: Implications for first and second language acquisitiona. *The Journal of the Acoustical Society of America*, *119*(5), 3059–3071.

Holt, L. L., & Lotto, A. J. (2010). Speech perception as categorization. *Attention, Perception, & Psychophysics*, *72*(5), 1218–1227.

House, A. S., & Fairbanks, G. (1953). The influence of consonant environment upon the secondary acoustical characteristics of vowels. *The Journal of the Acoustical Society of America*, *25*(1), 105–113.

Ito, M., Tsuchida, J., & Yano, M. (2001). On the effectiveness of whole spectral shape for vowel perception. *The Journal of the Acoustical Society of America*, *110*(2), 1141–1149.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, *90*(430), 773–795.

Ko, E.-S., Soderstrom, M., & Morgan, J. (2009). Development of perceptual sensitivity to extrinsic vowel duration in infants learning american english. *The Journal of the Acoustical Society of America*, *126*(5), EL134–EL139.

Kruschke, J. K. (2011). *Doing Bayesian analysis: A tutorial with R and BUGS*. Academic Press.

Lahiri, A., & Reetz, H. (2010). Distinctive features: Phonological underspecification in representation and processing. *Journal of Phonetics*, *38*(1), 44–59.

Laming, D. R. J. (1968). Information theory of choice-reaction times.

Lee, M., & Wagenmakers, E. (2013). *Bayesian modeling for cognitive science: A practical course*. Cambridge: University Press.

Lehiste, I., & Lass, N. J. (1976). Suprasegmental features of speech. *Contemporary issues in experimental phonetics*, *225*, 239.

Lipski, S. C., Escudero, P., & Benders, T. (2012). Language experience modulates weighting of acoustic cues for vowel perception: An event-related potential study. *Psychophysiology*, *49*(5), 638–650.

Lisker, L. (1986). "voicing" in english: A catalogue of acoustic features signaling/b/versus/p/in trochees. *Language and speech*, *29*(1), 3–11.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization* (No. 8). Oxford University Press.

McMurray, B., Aslin, R. N., & Toscano, J. C. (2009). Statistical learning of phonetic categories: insights from a computational approach. *Developmental science*, *12*(3), 369–378.

McMurray, B., Clayards, M. A., Tanenhaus, M. K., & Aslin, R. N. (2008). Tracking the time course of phonetic cue integration during spoken word recognition. *Psychonomic bulletin & review*, *15*(6), 1064–1071.

Meng, X.-L. (1994). Posterior predictive p-values. *The Annals of Statistics*, 1142–1160.

Miller, J. L. (2001). Mapping from acoustic signal to phonetic category: Internal category structure, context effects and speeded categorisation. *Language and Cognitive Processes*, *16*(5-6), 683–690.

Morey, R., Rouder, J., & Jamil, T. (2014). Bayesfactor: Computation of bayes factors for common designs. *R package version 0.9*, *8*.

Morrison, G. S. (2013). Theories of vowel inherent spectral change. In *Vowel inherent spectral change* (pp. 31–47). Springer.

Nooteboom, S. G., & Cohen, A. (1984). *Het proces van spreken en verstaan, een nieuwe inleidingin de experimentele fonetiek.* The Netherlands: Van Gorcum: Assen.

Nooteboom, S. G., & Doodeman, G. (1980). Production and perception of vowel length in spoken sentences. *The Journal of the Acoustical Society of America*, *67*(1), 276–287.

Peirce, J. W. (2007). Psychopy—psychophysics software in python. *Journal of Neuroscience Methods*, *162*(1), 8–13.

Peterson, G. E., & Lehiste, I. (1960). Duration of syllable nuclei in english. *The Journal of the Acoustical Society of America*, *32*(6), 693–703.

Pierrehumbert, J. (2001). Why phonological constraints are so coarse-grained. *Language and Cognitive*

Processes, *16*(5-6), 691–698.

Pisoni, D. B., & Tash, J. (1974). Reaction times to comparisons within and across phonetic categories. *Perception & Psychophysics*, *15*(2), 285–290.

Polka, L., & Werker, J. F. (1994). Developmental changes in perception of nonnative vowel contrasts. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(2), 421-435.

Pols, L. C., Tromp, H. R., & Plomp, R. (1973). Frequency analysis of dutch vowels from 50 male speakers. *The journal of the Acoustical Society of America*, *53*(4), 1093–1101.

R Development Core Team. (2016). The r project for statistical computing [Computer software manual]. Vienna, Austria.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59-108.

Ratcliff, R., Gomez, P., & McKoon, G. M. (2004). A diffusion model account of the lexical decision task. *Psychological Review*, *111*, 159-182.

Ratcliff, R., & McKoon, G. M. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*, 873–922.

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, *9*(5), 347–356.

Ratcliff, R., Thapar, A., Gomez, P., & McKoon, G. (2004). A diffusion model analysis of the effects of aging in the lexical-decision task. *Psychology and aging*, *19*(2), 278-289.

Ratcliff, R., Thapar, A., & McKoon, G. (2001). The effects of aging on reaction time in a signal detection task. *Psychology and aging*, *16*(2), 323-341.

Ratcliff, R., Thapar, A., & Mckoon, G. (2003). A diffusion model analysis of the effects of aging on brightness discrimination. *Perception & psychophysics*, *65*(4), 523–535.

Ratcliff, R., Thapar, A., & McKoon, G. (2004). A diffusion model analysis of the effects of aging on recognition memory. *Journal of Memory and Language*, *50*(4), 408–424.

Reinisch, E., & Sjerps, M. J. (2013). The uptake of spectral and temporal cues in vowel perception is rapidly influenced by context. *Journal of Phonetics*, *41*(2), 101–116.

Repp, B. H. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychological bulletin*, *92*(1), 81 - 110.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for

ANOVA designs. *Journal of Mathematical Psychology*, *56*, 356-374.

Salthouse, T. A. (1996). The processing-speed theory of adult age differences in cognition. *Psychological review*, *103*(3), 403-428.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, *6*(2), 461–464.

Stan Development Team. (2016). R stan: the r interface to stan, version 2.10.1. [Computer software manual]. Retrieved from `http://mc-stan.org`

Stone, M. (1960). Models for choice-reaction time. *Psychometrika*, *25*(3), 251–260.

Thapar, A., Ratcliff, R., & McKoon, G. (2003). A diffusion model analysis of the effects of aging on letter discrimination. *Psychology and Aging*, *18*(3), 415-429.

The JASP Team. (2016). *Jasp (version 0.7.5)[computer software].* https://jasp-stats.org/.

Toscano, J. C., & McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive science*, *34*(3), 434–464.

Trueblood, J. S., Brown, S. D., & Heathcote, A. (2014). The multiattribute linear ballistic accumulator model of context effects in multialternative choice. *Psychological review*, *121*(2), 179 -205.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky competing accumulator model. *Psychological Review*, *108*, 550-592.

van der Feest, S. V. H., & Swingley, D. (2011, Mar). Dutch and english listeners' interpretation of vowel duration. *J Acoust Soc Am*, *129*(3), 57-63.

van Heuven, V., Van Houten, J., & De Vries, J. (1986). De perceptie van nederlandse klinkers door turken. *Spektator*, *15*, 225–238.

van Ravenzwaaij, D., Brown, S. D., & Wagenmakers, E.-J. (2011). An integrated perspective on the relation between response speed and intelligence. *Cognition*, *119*(3), 381–393.

van Ravenzwaaij, D., Dutilh, G., & Wagenmakers, E.-J. (2012). A diffusion model decomposition of the effects of alcohol on perceptual decision making. *Psychopharmacology*, *219*(4), 1017–1025.

van Ravenzwaaij, D., Provost, A., & Brown, S. D. (2016). A confirmatory approach for integrating neural and behavioral data into a single model. *Journal of Mathematical Psychology*.

van Ravenzwaaij, D., van der Maas, H. L. J., & Wagenmakers, E.-J. (2011). Does the name-race implicit association test measure racial prejudice? *Experimental Psychology*, *58*(4), 271-7.

Vehtari, A., Gelman, A., & Gabry, J. (2016). *Practical bayesian model evaluation using leave-one-out cross-validation and waic.* Retrieved from `http://arxiv.org/abs/1507.04544`

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problem of p values. *Psychonomic Bulletin and Review*, *14*, 779-804.

Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & McKoon, G. (2008). A diffusion model account of criterion shifts in the lexical decision task. *Journal of Memory and Language*, *58*(1), 140–159.

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *The Journal of Machine Learning Research*, *11*, 3571–3594.

White, C. N., & Poldrack, R. A. (2014). Decomposing bias in different types of simple decisions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(2), 385 - 398.

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta psychologica*, *41*(1), 67–85.

Winn, M. B., Chatterjee, M., & Idsardi, W. J. (2012). The use of acoustic cues for phonetic identification: Effects of spectral degradation and electric hearing. *The Journal of the Acoustical Society of America*, *131*(2), 1465–1479.

Winn, M. B., Rhone, A. E., Chatterjee, M., & Idsardi, W. J. (2013). The use of auditory and visual context in speech perception by listeners with normal hearing and listeners with cochlear implants. *Frontiers in psychology*, *4*.

Zatorre, R. J., & Belin, P. (2001). Spectral and temporal processing in human auditory cortex. *Cerebral cortex*, *11*(10), 946–953.