Psychonomic Bulletin and Review (2020) 27:640–676
https://doi.org/10.3758/s13423-019-01687-6

THEORETICAL REVIEW

When context is and isn't helpful: A corpus study of naturalistic speech

Kasia Hitczenko[1] · Reiko Mazuka[2] · Micha Elsner[3] · Naomi H. Feldman[4]

Published online: 12 March 2020

Abstract

Infants learn about the sounds of their language and adults process the sounds they hear, even though sound categories often overlap in their acoustics. Researchers have suggested that listeners rely on context for these tasks, and have proposed two main ways that context could be helpful: top-down information accounts, which argue that listeners use context to predict which sound will be produced, and normalization accounts, which argue that listeners compensate for the fact that the same sound is produced differently in different contexts by factoring out this systematic context-dependent variability from the acoustics. These ideas have been somewhat conflated in past research, and have rarely been tested on naturalistic speech. We implement top-down and normalization accounts separately and evaluate their relative efficacy on spontaneous speech, using the test case of Japanese vowels. We find that top-down information strategies are effective even on spontaneous speech. Surprisingly, we find that at least one common implementation of normalization is ineffective on spontaneous speech, in contrast to what has been found on lab speech. We provide analyses showing that when there are systematic regularities in which contexts different sounds occur in—which are common in naturalistic speech, but generally controlled for in lab speech—normalization can actually increase category overlap rather than decrease it. This work calls into question the usefulness of normalization in naturalistic listening tasks, and highlights the importance of applying ideas from carefully controlled lab speech to naturalistic, spontaneous speech.

Keywords Speech perception · Categorization · Category learning

Introduction

Listeners are exposed to highly variable, continuous speech and map it to discrete sound categories. To do so, they first learn as infants what the relevant sounds of their language are, and, subsequently, map incoming signal to learned categories. This is generally a robust process—infants learn about the sounds of their language as early as 6 months (Kuhl et al., 1992) and, for the most part, listeners process what they are hearing in an effortless manner. However, despite how seemingly easily listeners solve these tasks, they are computationally difficult problems. In fact, after decades of research in this area, researchers have not yet established a robust one-to-one mapping between signal and category that works to anywhere near the degree of success of human listeners.

The reason these tasks are so computationally difficult is because there is a large amount of variability in the speech signal, which can lead to acoustic overlap between different sound categories (Bion et al., 2013). One sound category can be acoustically realized in infinitely many ways, and two different sound categories can have identical acoustic realizations. This makes establishing a one-to-one mapping between speech and category difficult. Although we focus on speech perception in this paper, this problem is not unique to speech. For example, a particular visual

This work was supported by National Science Foundation grants #IIS-1421695, #IIS-1422987, #DGE-1449815, and #1713974.

✉ Kasia Hitczenko
kasia.hitczenko@northwestern.edu

[1] Department of Linguistics, Northwestern University, 2016 Sheridan Road, Evanston, IL 60208, USA

[2] RIKEN Center for Brain Science, Wako, Japan

[3] Department of Linguistics, The Ohio State University, Columbus, OH, USA

[4] Department of Linguistics and UMIACS, University of Maryland, College Park, MD, USA

stimulus may appear completely different across instances, due to, for example, lighting conditions, viewing angles, or occlusion. As in speech, two different objects can also have identical physical attributes, and, yet, for the most part, people effortlessly identify what they are seeing (Bar, 2004).

The basic problem is that absolute acoustic or other perceptual cues are insufficient to separate categories as well as humans do. This has led researchers to propose that listeners may be relying on context to help map from signal to categories. The role of context is widely studied in cognitive science, and fundamental to many cognitive theories, with most researchers largely agreeing that it is crucial in speech perception, language acquisition, object recognition, and visual perception, along with many other domains (e.g., Warren, 1970; Ganong, 1980; Port & Dalby, 1982; Mann & Repp, 1980; Bar & Ullman, 1996; Bar, 2004).

In the speech domain, researchers have identified two main, non-mutually exclusive ways that listeners could rely on context, based on two ways that context affects a speaker's production. The first is that context affects which sounds are likely to occur—e.g., /æ/ is much more likely than /ɛ/ to occur in the context th_t ('that' is a word, 'thet' is not), so listeners could be biased to perceive acoustics in that frame as /æ/ rather than /ɛ/. That is, top-down information could guide expectations about what category was likely to be heard. This type of information can supplement the acoustics, and we will refer to these as 'top-down information' accounts.[1]

The second is that context affects how sounds are produced. For example, who is speaking will significantly and systematically alter the acoustics of the signal. This leads to variability in how a particular sound is produced, and can lead to overlap between different sound categories (e.g., one speaker's /s/ could be another speaker's /ʃ/ as shown in Newman et al., 2001). Listeners could, thus, factor out systematic variability stemming from contextual factors like speaker (but also, speech rate, position in an utterance, neighboring sounds, and so forth) from their input. Removing variability may lead to less overlap between categories, and make the mapping from acoustics to categories clearer. In other words, context could be used to pre-process the acoustics that are used for categorization decisions. These types of accounts have generally been termed 'normalization' accounts.

The top-down information and normalization account examples provided above make use of two different contextual factors (i.e. neighboring sounds vs. speaker information), but many contextual factors can affect both stages of production (i.e. which category is produced and how it is produced). That is, the core difference between these two accounts is not which contextual factors are used, but rather how they are used. In this paper, we will broadly define context to include neighboring sounds, position in a word/utterance, part of speech of the word the sound was produced in, speech rate, speaker, as well as aspects of the sound itself that have already been processed. Many of these contextual factors could be useful in both top-down information and normalization accounts. For example, a particular phoneme may be a priori more likely to be produced word-finally, in which case a listener would benefit from a bias towards perceiving that phoneme word-finally, as in a top-down information strategy. At the same time, sounds are acoustically longer word-finally, so a listener would separately benefit from accounting for this difference in how the sound was produced, as in a normalization strategy.

These two ways of using context have both been studied extensively. There is a large body of experimental and computational work supporting the notions (i) that context does affect both which sound is produced and how it is produced, (ii) that listeners can make use of these strategies, and (iii) that listeners do make use of these strategies to help overcome the overlapping categories problem. Both ways of using context are relatively well accepted in the speech perception literature.

However, there are two main limitations with previous work that warrant further study. First, these two ways of using context, although different, have been somewhat conflated in previous work, and have been difficult to dissociate experimentally. In particular, experiments that have been used to argue for one over the other generally show that an acoustic signal is perceived as one category in a particular context, but when the same signal is placed in a different context, it is perceived differently. This type of finding has been used to argue for both top-down information and normalization accounts, but depending on the specifics, merely shows that context is used, but not how. Therefore, it is not entirely clear whether listeners are using both of these strategies, and if not, which one they are using. This limitation requires separating these accounts, and testing them individually, which computational methods will allow us to do.

Second, these ideas have mostly been studied on synthetic or carefully controlled lab speech, which differs in important ways from the naturalistic and spontaneous speech that listeners actually learn from and process. It is not clear whether promising results from controlled lab speech generalize to more variable spontaneous speech; indeed, where tested, they have often not (e.g., Antetomaso

[1] While top-down information accounts sometimes refer exclusively to the lexical, syntactic, and semantic levels influencing lower levels of processing, we will also include influences by already-processed phonemic information under this umbrella term, as will be explained in more detail below.

et al., 2017). In addition, most of the debate so far has centered on whether listeners do or do not make use of these strategies, and has assumed that if listeners did use these strategies, doing so would help them process naturalistic speech. However, there is actually little to no evidence so far that these strategies are effective on naturalistic speech. Addressing this limitation requires applying these two strategies to naturalistic speech of the type that listeners are mostly exposed to, and testing whether they are effective in separating overlapping categories.

In this work, we study how context can be effectively used in speech perception, taking these two issues into account. We implement top-down information and normalization accounts separately and evaluate their relative contribution in the process of going from speech signal to categories—and we do so on spontaneous speech. We focus on the test case of Japanese vowel length, a test case with particularly overlapping categories that current computational models fail to learn. We find that top-down information is helpful in separating the sound categories, remaining robust even on spontaneously produced speech. However, contrary to expectations, we find that normalization is not helpful, at least as it has often been implemented in the cognitive literature. We then study why exactly the discrepancy between our results and previous findings occurs. We find that the discrepancy results from the difference between controlled lab speech and spontaneous speech, by showing that the exact same normalization process we use works if we apply it to lab speech that is more similar to the speech used in previous work. Simulations and a mathematical analysis reveal that one property of spontaneous speech that seems to play a particularly important role is the fact that categories do not occur uniformly across contexts in spontaneous speech, as they do in controlled lab speech. Imbalances in where categories occur—precisely the type of signal that is helpful in top-down information accounts—can hurt normalization. That is, this work not only dissociates two strategies that have often been conflated, but shows interesting interactions between them, such that properties of the input that make one of them effective can make the other ineffective.

Past research on these cognitive theories has tended to focus on whether listeners do or do not use these strategies, assuming that using them would actually solve the overlapping categories problem present in speech. While our results validate this assumption for top-down information accounts, our results show that in our case study, this assumption is wrong for a common implementation of normalization. It is possible that the theory about how listeners normalize could be repaired in light of these findings, as we will discuss, and this warrants further study. Overall, these results highlight the importance of studying speech perception using spontaneous speech, in addition to carefully controlled lab speech, as results from one do not necessarily generalize to the other.

Background

The Japanese vowel length contrast

This paper uses the Japanese vowel length contrast as a test case to compare the relative efficacy of top-down information and normalization strategies. In Japanese, there are two sound categories along the duration dimension—referred to as 'short' vowels and 'long' vowels (Vance, 1987). Which category is used can change the meaning of a word. For example, /biru/ with a short vowel means 'building,' while /biːru/ with a long vowel means 'beer'. Results from perception and production studies reveal that Japanese speakers differentiate short and long vowels: they produce short and long vowels differently and can identify which vowel length category a particular vowel belongs to (Chen et al., 2016; Hisagi et al., 2010; Mugitani et al., 2009; Werker et al., 2007). Based primarily on studies of controlled laboratory speech, vowel length is often thought to be signaled primarily by the vowel duration cue and to a lesser extent, by formant values (e.g., Arai et al., 1999; Kinoshita et al., 2002; Lehnert-LeHouillier, 2010). Some researchers have alternatively hypothesized that relativized vowel duration (the ratio between a vowel's duration and the duration of its neighboring sound or the word it is in) might be the primary cue to vowel length instead (Hirata, 2004). However, there is no conclusive evidence in either direction so, for a number of reasons, we follow a substantial body of work in using vowel duration as the cue to vowel length. On the one hand, absolute duration can be more easily and reliably measured in naturalistic speech, where, for example, vowels often occur in isolation without any neighboring sounds to relativize against. On the other hand, one of the theories we consider—normalization—has historically only operated over absolute cues. This is because it has been treated as an alternative, not a supplement, to relativizing cues: both are ways to transform the acoustics in such a way as to remove systematic contextual variability, and doing both could be redundant. Nonetheless, future work should study how these results generalize when using duration ratios, and we return to this issue in the General Discussion.

At this point, we wish to highlight an important terminological distinction between vowel length and vowel duration—and the corresponding two meanings that short/long can have in this context. Vowel length refers to the category status of a vowel—i.e., whether it is the vowel category that will result in /biru/ ('building') or /biːru/ ('beer'). Vowel duration refers to the acoustic property of a

vowel—i.e., how long it took the speaker to articulate the vowel—and is thought to be a cue to vowel length. Therefore, a vowel can be referred to as short (or long) if it belongs to the short (or long) category, but it can also be referred to short (or long) depending on its physical duration. In this paper, we will use 'phonologically short/long,' 'phonemically short/long,' or simply 'short/long' to refer to category status, and 'acoustically short/long' to refer to physical vowel duration.

This distinction is critical because a vowel's duration and length do not always line up. Recent work has shown that although short vowels and long vowels are different categories, the range of durations they can have overlap substantially (Bion et al., 2013). While long vowels are, on average, acoustically longer than short vowels, a particular production of a phonologically short vowel can be acoustically longer than a particular production of a phonologically long vowel. In fact, because only 9% of Japanese vowels are phonologically long, the combined distribution of all vowels is unimodal along the duration dimension (Fig. 1). Therefore, while vowel duration is thought to be the primary cue to vowel length, it is insufficient to completely separate short and long vowels in spontaneous productions. This is precisely what has led some researchers to instead consider relativized vowel duration as the primary cue to vowel length; however, that work has only considered controlled lab speech. On naturalistic speech, this problem persists, regardless of which type of cue is used (Bion et al., 2013).

Note that vowel length is not the only way that Japanese listeners need to categorize incoming vowels. There are ten total vowel categories in Japanese, a short and long version of five different vowel qualities (/a/, /e/, /i/, /o/, /u/), so Japanese listeners need to determine both the vowel length and the vowel quality of incoming vowels. However, the acquisition and processing of vowel quality and vowel length seem to be relatively independent processes. It is thought that Japanese infants learn the vowel length contrast at around 10 months of age, about 6 months after they have been argued to learn the vowel qualities (Sato et al., 2010). For the purpose of this paper, we simply consider how Japanese listeners may learn and process vowel length, treating vowel quality as something that is already known and can help in the categorization.

Japanese vowel length is just one instance of a commonly observed overlapping categories problem, both in speech perception (Allen et al., 2003; Hillenbrand et al., 1995; Hillenbrand et al., 2001; Narayan, 2013; 2008; Narayan et al., 2017; Newman et al., 2001; Swingley & Alarcon, 2018), and more generally (Adelson, 1993; Bar, 2004; Todorović, 2010), where the physical cues are insufficient on their own to explain human perception. The Japanese vowel length contrast is a good first test case to consider because (i) existing computational models fail to adequately learn and classify these vowels due to overlap between the categories, (ii) contextual information has been argued to play a role in processing and learning, making it a good test case for studying how context is helpful, and (iii) there exists a hand-annotated dataset consisting of both child- and adult-directed Japanese spontaneous speech.

However, one important question to consider is the extent to which our findings on vowel length will generalize to other contrasts. Our results are likely to be informative about other cases of overlapping categories, which may arise for similar reasons as in this test case. Nonetheless, the Japanese vowel length contrast has some unique properties that might make it different from some other contrasts. First, there are disproportionately many short vowels relative to long vowels, while other contrasts are often more balanced in numbers. In fact, the overlapping categories problem arises because of this imbalance: if short vowels and long vowels were equally frequent, then the distribution might be bimodal (as in controlled lab speech), which might make the contrast easier to learn. Second, the main acoustic cue to this contrast—duration—is particularly influenced by other linguistic and non-linguistic factors. This means that the underlying relationship between category and acoustic cue may be particularly difficult to recover for Japanese vowel length compared to a different contrast where the acoustic cue is less affected by other factors. Third, the Japanese

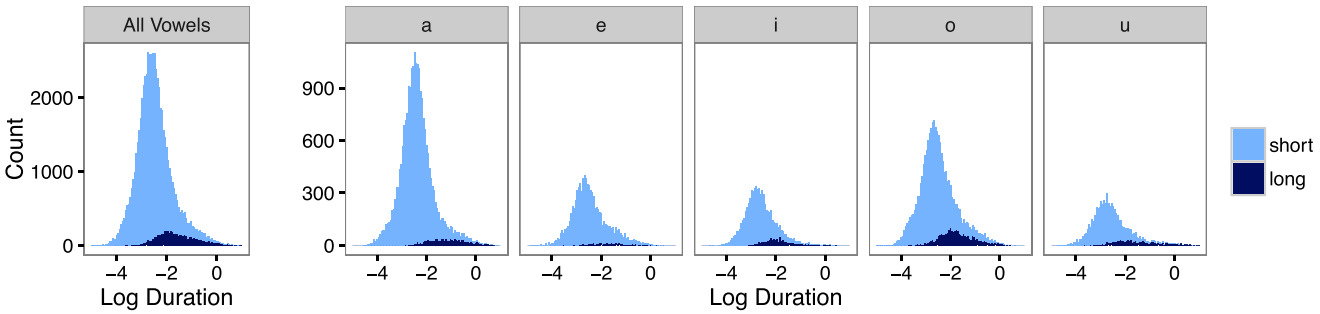


Fig. 1 Distribution of R-JMICC dataset vowels (by log-duration): all are unimodal distributions. Values displayed are logs of the vowel durations in seconds. As a result, log-durations will be negative whenever the vowel is less than a second long

vowel length contrast has relatively low functional load, because it does not distinguish many minimal pairs. As a result, it is possible that the acoustics are less important than in other contrasts, because which sound was produced might be more predictable. Finally, the Japanese vowel length is acquired relatively late compared to other contrasts (e.g., vowel quality contrasts) in both phonetic development (Sato et al., 2010) and phonological development (Mugitani et al., 2009). As a result, it could be that learners have access to different information or that a different learning mechanism is involved in learning this contrast. All of these properties suggest that Japanese vowel length may be a particularly overlapping and difficult contrast to learn, and results from these analyses could be illuminating for less extreme contrasts. While we will speculate on how our results might generalize to other types of contrasts through the paper, future work will need to investigate this issue more thoroughly.

Categorization - using unnormalized acoustic cues

Japanese listeners must first determine how many sounds there are along the duration dimension during acquisition and, once they have learned the language and its categories, they must decide which of the vowels they hear are short or long through a categorization process. We will test the usefulness of top-down information and normalization strategies by implementing them computationally and seeing how well they perform in categorizing Japanese vowels as short or long. We will compare their performance against a baseline model that categorizes exclusively based on unaltered, unnormalized acoustic cues.

All of the models we test are supervised and rely on already knowing the distinction between short and long vowels. As a result, these results are only directly applicable to adult speech perception, where the task is precisely to categorize vowels, and not to acquisition, where the task is to discover that there are two categories to begin with. Nonetheless, the results of this paper can provide some insight into acquisition, by pointing to promising directions to pursue in the future. Our categorization analyses reveal how well a strategy can, at its best, separate short vowels from long vowels. If a strategy cannot separate short vowels from long vowels in a supervised model, then it would be hard for an infant to use it to learn, and is less promising to pursue in the context of acquisition. A strategy that can separate short and long vowels in a supervised setting is a much more promising one to pursue in the unsupervised acquisition setting, even though the analyses in this paper can not make claims about how exactly infants learn these distinctions. In what follows, we lay out what this base categorization model looks like, before turning to a discussion of how context could be used in the process.

A categorization model can take many forms, but for the purpose of this paper, we model categorization using logistic regression, following previous work (McMurray & Jongman, 2011). Our logistic regression models will take as input a set of cues and map them to vowel category (either short or long). The baseline categorization model—argued to be insufficient in Bion et al. (2013) as described in the previous section—will take as input a vowel's acoustic cues—duration and formant values—and will categorize the vowel as short or long depending on those cues. It will do so by weighting each of the cues (in terms of how much they contribute to whether the vowel should be short or long), summing the weighted acoustic cues, and then transforming this value into a probability that represents the probability that this vowel is short versus long. That is, if we consider the acoustic cues, d, f_1, f_2, f_3, logistic regression takes the following form:

$$P(\text{long}|d, f_1, f_2, f_3) = \frac{1}{1 + e^{\beta_0+\beta_d d+\beta_{f_1} f_1+\beta_{f_2} f_2+\beta_{f_3} f_3}} \quad (1)$$

where the β terms are weights on the cues—duration, d, and formants, f_1-f_3. The probability that the vowel is short is $1-P(\text{long}|d, f_1, f_2, f_3)$. The model categorizes the sound as belonging to the category (short or long) that has the higher probability.

Learning this function involves learning an intercept (β_0), as well as a weight for each cue ($\beta_d ... \beta_{f_3}$). The model is trained on data that consist of the unnormalized acoustic cues of a vowel, labeled with the category that vowel belongs to, and weights are learned so as to optimally separate the short vowels from the long vowels. Once we learn this function, we can take any new vowel and calculate the probability that that vowel is long (or short). The only information that this model has access to is the acoustics of the vowel, so it will be insufficient when categories overlap in acoustic cues, but incorporating context could help.

How context could be used

There are two ways that listeners could use context, which we define broadly to include neighboring sounds, speaker, position in a word/utterance, speech rate, already processed aspects of the sound itself (like vowel quality), and so forth (see full list in the first column of Table 1). To illustrate the two ways, it is helpful to consider the production process.

When a speaker produces a vowel, they first decide which category to produce (short or long vowel) depending on what word they are producing, and then utter a particular acoustic value for that vowel based on the vowel category they are producing. Both of these components of the production process are affected by the context of the vowel, but this is ignored in the base categorization model. The

Table 1 The full set of contextual factors available for each dataset, with factors that were included in the normalization upper-bound shown in bold (as described in the sections on normalization methods). In the case of the R-JMICC corpus, these are taken from the linear regression normalization method, which outperformed the neural network normalization method

R-JMICC Spontaneous	Werker Read	Werker Spontaneous
Vowel Quality	**Vowel Quality**	**Vowel Quality**
Speaker	**Speaker**	**Speaker**
Previous Sound	**Previous Sound**	**Previous Sound**
Following Sound	Following Sound	Following Sound
Prosodic Position of Word (2 factors)	**Prosodic Position**	**Prosodic Position**
Prosodic Position of Vowel (12 factors)	F0	**F0**
Accented		
Previous Sound Duration (Speech Rate)		
Following Sound Duration (Speech Rate)		
Condition (Toys or Books)		
Part of Speech		

following two sections introduce the two ways context affects a speaker's sound production and, consequently, the two ways that context could be used to improve categorization.

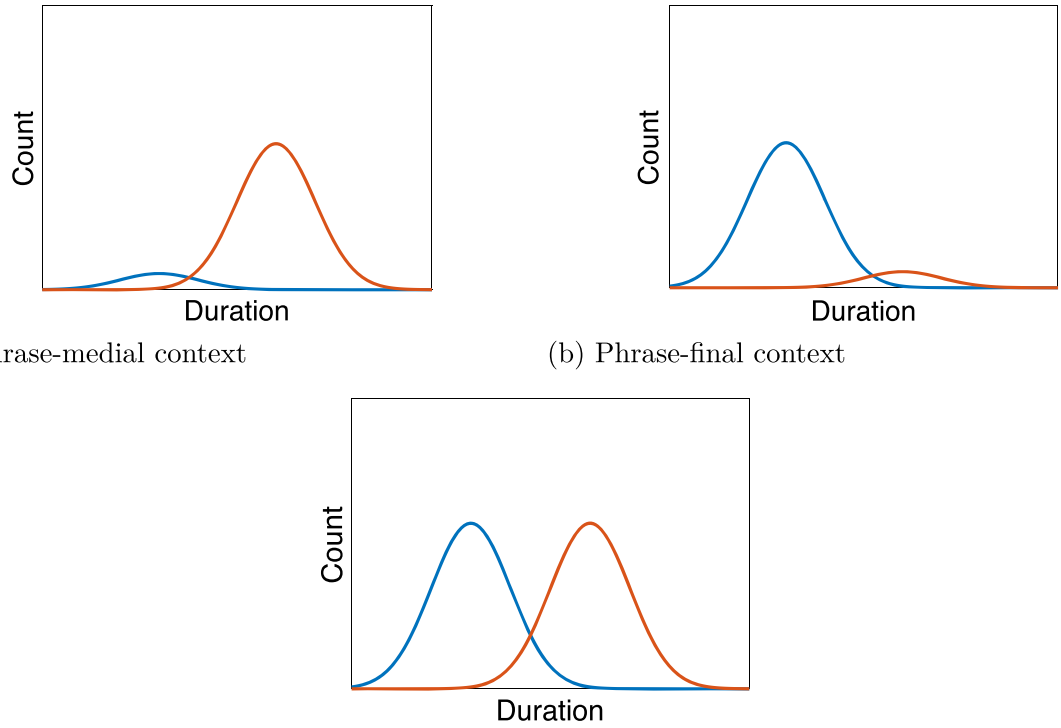
Top-down information accounts

First, the context of a sound directly relates to which vowel category is more or less likely to be produced. An English speaker is much more likely to produce an /æ/ vowel (as in 'mat') than an /ɛ/ vowel (as in 'met') in the context th_t (the word 'that' exists, but 'thet' does not), 'but the opposite holds when the context is w_t instead (the word 'wet' exists, but 'wat' does not)'. In Japanese, a speaker is relatively more likely to produce a long vowel if they are saying an /o/ vowel than if they are saying an /a/ vowel, as can be seen in Fig. 1. A listener could benefit from taking this type of prior knowledge into account, and indeed listeners' perception appears to be biased by which sound was a priori more likely to occur.

This type of strategy is often referred to as a 'top down information' account, as it makes use of listeners' prior knowledge of which sounds are likely to occur in which contexts, in addition to the sounds' bottom-up acoustic cues. It can also be thought of a 'predictive' strategy, in the sense that context is used to directly predict which sound occurred.

In this paper, we will use the term 'top-down information' account to refer to the use of any prior knowledge—including information at the phonemic level—to directly bias perception. We wish to make explicit that we are using the term more broadly than it sometimes is used in the literature. It is sometimes used to refer only to lexical, syntactic, or semantic factors influencing speech perception. However, we use it in the sense of any non-acoustic information directly biasing speech perception, which can include information at the phonemic level (e.g., vowel quality or neighboring sounds).

The categorization model presented in the previous section does not, in its current form, take this type of information into account: it only takes into account whether one of the vowel categories is more likely to occur overall, not whether vowel categories are more likely to occur in particular contexts. To illustrate why this is problematic, consider the toy case shown in Fig. 2, in which there are two categories (short and long), and only two contexts (let's say phrase-medial vowels and phrase-final vowels). Overall, phonemically short and long vowels occur with identical frequency; however, the phonemically short category is much more likely to occur in phrase-final position and the phonemically long category is much more likely to occur in phrase-medial position. The base categorization model will simply place the category boundary halfway between the short and long vowel means in (c), when really this category boundary should be at a shorter duration for vowels that occur phrase-medially and at a longer duration for vowels that occur phrase-finally. This means that the base categorization model will overclassify phrase-medial short vowels (i.e., short vowels in contexts where long vowels are much more likely to occur) as long and will overclassify phrase-final long vowels as short. However, taking into account context as a top-down influence can help correct this problem. In particular, if the model or listener takes into account expectations about which vowel is more likely



(a) Phrase-medial context (b) Phrase-final context

(c) Overall

Fig. 2 A toy example demonstrating how using contextual information as top-down information can be helpful. Although short vowels and long vowels are equally common overall, short vowels are much more common phrase-finally, and the opposite holds phrase-medially. Our baseline categorization model will not be able to take this into account and will miscategorize some vowels as long in phrase-final position and miscategorize some vowels as short in phrase-medial position

to occur in the context heard, then they will, all else being equal, be biased towards categorizing vowels as short in contexts where short vowels are more likely to occur, and biased towards categorizing vowels as long in contexts where long vowels are more likely to occur.

The base categorization model can be augmented to take this into account, in order to reflect what listeners are thought to do. In the logistic regression, this could be accomplished by adding the contexts as independent predictors. For example, in our Japanese example, if we added the vowel quality, q, of the vowel as an independent predictor, this could encode the fact that vowels that are /o/ are relatively more likely to be long than /a/ vowels:

$$P(\text{long}|d,\, f_1,\, f_2,\, f_3,\, q) = \frac{1}{1 + e^{\beta_0+\beta_1 d+\beta_2 f_1+\beta_3 f_2+\beta_4 f_3+\beta_5 q}} \tag{2}$$

Essentially, this means that in addition to the acoustic cues affecting the relative probabilities of the vowel being short or long, the quality of vowel can also affect the categorization decision. Additional terms could be added depending on what other contextual factors are thought to predict category membership.

The effect is that instead of having one categorization boundary overall, the boundary between categorizing a vowel as short and categorizing it as long will shift depending on the context, and how likely short vs. long vowels are to occur in that context. If phonemically long vowels are relatively more likely to occur in a particular context than short vowels, then the boundary between short and long vowels will shift towards vowels of shorter durations, such that more vowels are classified as long, and the opposite holds in contexts where phonemically short vowels are relatively more likely.

Crucially, this model assumes that the acoustics of a sound will be the same regardless of the context it was produced in, and so it cannot take into account the fact that vowel durations may systematically vary between different contexts (due to e.g., acoustic lengthening effects).

Normalization accounts

In the previous section, we saw that context can affect which category a speaker is likely to produce.

The next and final component of the speaker's production process is to actually produce an acoustic value for the vowel category they have chosen. This portion of the

production process is also affected by context, as context systematically and predictably affects how a particular sound category is acoustically realized. As an example, vowels uttered in fast speech are, all else being equal, acoustically shorter than vowels uttered in slow speech. Similarly, vowels uttered phrase-finally are, all else being equal, acoustically longer than vowels uttered phrase-medially.

This can introduce variability and overlap between short vowels and long vowels into the overall distribution—and is problematic for a categorization model simply relying on absolute acoustic cues. Consider the toy case in Fig. 3. Here again, there are two vowel categories (short vowels and long vowels) and there are two contexts (let's say phrase-medial and phrase-final). In phrase-medial position, short vowels are produced with an average acoustic duration of 150 ms and long vowels are produced with an average acoustic duration of 300 ms. In phrase-final position, vowels are systematically acoustically lengthened by 100 ms. This scenario is problematic for the base categorization model because the overall distribution will reveal a lot of overlap between sound categories. In particular, long vowels in phrase-medial position will overlap with short vowels in phrase-final position. The baseline categorization model presented previously learns a categorization boundary between short vowels and long vowels, which is the same for all vowels, regardless of context. This will cause the model to overclassify vowels occurring in lengthening contexts as phonemically long, and overclassify vowels occurring in acoustically short contexts as phonemically short.

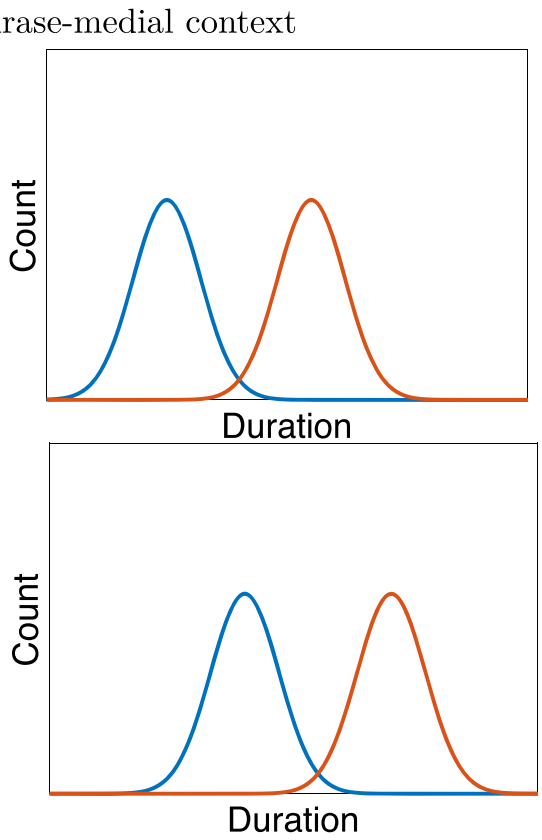


Fig. 3 A toy example demonstrating how using contextual information to normalize acoustics can be helpful. Phrase-final vowels are systematically acoustically lengthened, which introduces overlap in the overall distribution of short vowels and long vowels. However, a listener who knows that phrase-final vowels are systematically acoustically lengthened could normalize for this acoustic lengthening, and reduce the overall overlap between short vowels and long vowels in their input

However, the shifts in acoustic cue values are systematic and predictable once the context is known, so using contextual information can help overcome these problems—and listeners have been argued to do so in listening situations. There are various ways this problem could be overcome, and corresponding ways the baseline logistic regression model could be augmented. One is that listeners might build a separate mapping between acoustics and category membership for each context they encounter, such that lengthening contexts will have a boundary between short/long vowels at a higher duration, and vice versa for shortening contexts. This idea is referred to as adaptation (Kleinschmidt & Jaeger, 2015), and we will return to it in later discussion, but do not directly study it in this paper. Instead, we focus on a second idea, referred to as normalization.

The idea behind normalization is that instead of creating a different acoustic boundary between short/long vowels for every context encountered, all acoustics are mapped to the same context-independent acoustic space and then one boundary is estimated in this context-independent space. This is done by estimating how much any particular context lengthens or shortens the vowels, and then undoing all lengthening or shortening processes. Returning to our example, normalization would work by estimating that the vowels in phrase-final context are on average 100 ms longer than vowels in phrase-medial context, and then essentially shifting the distributions to compensate for this lengthening. Another way to think about it is that each vowel is represented relative to the mean duration of vowels that occurred in the same context. Acoustic cues that have been mapped to this context-independent space are referred to as normalized cues.

In the top-down information accounts, the logistic regression in Eq. 1 was augmented by adding additional predictors based on the context of the sound in question (e.g., the vowel quality, q). In normalization accounts, the logistic regression in Eq. 1 is changed by performing a

preprocessing step (which will be described in detail below), and inputting normalized cues (d^{norm}, f_1^{norm}, f_2^{norm}, f_3^{norm}) into the logistic regression, instead of unnormalized cues as before (d^{unnorm}, f_1^{unnorm}, f_2^{unnorm}, f_3^{unnorm}).

This means that while information about the context a sound occurs in is a direct input to the logistic regression categorization model in top-down information accounts, it is not in normalization accounts. Rather, in normalization accounts, the contextual information is used to obtain normalized acoustic cues, which are ultimately the only input to the categorization model (e.g., Cole et al., 2010; McMurray & Jongman, 2011).

The normalized cues of a sound are obtained by predicting its expected cue values based on the context it occurs in, and then comparing these expected cue values against its actual cue values.

$$cue^{norm} = cue^{unnorm} - cue^{expected} \tag{3}$$

The expected cues can be calculated from a sound's contextual information using various methods, and we make use of two such methods. First, we follow past work, and train a linear regression to predict a sound's acoustic cues from the context the sound occurs in (Cole et al., 2010; McMurray & Jongman, 2011).

The second method involves training a neural network to predict a sound's acoustic cues from the context the sound occurs in. The benefit of this method is that it allows for more powerful, non-linear normalization functions to be learned. Once the pre-processing step is complete and we have normalized all of the acoustic cues relative to context, we can then replace the unnormalized cues with normalized cues in the logistic regression categorization model:

$$P(\text{long}|d^{norm}, f_1^{norm}, f_2^{norm}, f_3^{norm}) = \frac{1}{1 + e^{\beta_0 + \beta_1 d^{norm} + \beta_2 f_1^{norm} + \beta_3 f_2^{norm} + \beta_4 f_3^{norm}}} \tag{4}$$

Normalizing has the effect of shifting where the boundary between short and long vowels falls. In particular, considering the example in Fig. 3, the vowels in phrase-final position are perfectly shifted from those in phrase-medial position. Ignoring context will cause there to be a huge degree of variability and overlap between short and long vowel acoustics in the overall distributions. However, normalizing out this variability by shifting the two contexts so that they line up will help. In particular, vowels that are acoustically quite long will be readily classified as short because listeners may be accounting for the fact that these vowels were lengthened and undoing this effect. That is, a long acoustic duration presented in phrase-medial context may be perceived as long; however, when placed in phrase-final context, that same vowel with the same acoustics may now be perceived as phonemically short because it may be relatively short relative to other vowels that occur in that same lengthening context.

There are other implementations of normalization, including z-scoring, vocal tract normalization, relativizing cues, as well as proposals by Dillon et al. (2013) that we do not test in this work. We return to the question of how our results generalize to other normalization methods in the General Discussion, but future work should investigate this question more thoroughly.

Adaptation accounts

Another idea that has been proposed is that of adaptation (Kleinschmidt & Jaeger, 2015). Under 'adaptation' accounts, listeners build a separate model for each context they encounter, so they have a different mapping between acoustic space to categories for each context a sound occurs in. For example, a listener using an adaptation strategy would build a separate model for utterance-medial /o/ vowels, utterance-final /a/ vowels, etc. (see Kleinschmidt & Jaeger, 2015 for a more thorough explanation of adaptation). In doing so, adaptation allows listeners to ignore systematic acoustic variability that stems from the context a sound occurs in. These models would encode the fact that a shorter absolute duration is required to classify a vowel as long in utterance-medial position than in utterance-final position, without transforming the vowels' acoustic cues as is done in normalization.

While both normalization and adaptation aim to explain how listeners account for systematic acoustic variability, they do so in different ways. In particular, under normalization accounts, all acoustics are mapped to one context-independent acoustic space using an explicit normalization function, and listeners only estimate one boundary between short and long vowels in the context-independent acoustic space. Under adaptation accounts, listeners estimate a different acoustic boundary between short and long vowels for each context they encounter, without considering data from other contexts. This means that in addition to accounting for systematic acoustic variability, adaptation can also encode top-down information. Building a separate model for each encountered context necessarily encodes relative frequency of occurrence of different sound categories across different contexts, and this could bias perception. Therefore, adaptation can take advantage both of factoring out systematic variability and using top-down information. Because we wish to disentangle the relative contribution of these two ideas, we do not study the efficacy of adaptation strategies here, but we return to the idea of adaptation in the General Discussion.

Crucially, we have seen that these two distinct theories about how listeners could use context in categorization produce similar changes in categorization, which has led

researcher to sometimes conflate them in the literature. In what follows, for each of the two strategies, we first review the evidence that has been used to argue in favor of listeners using them, and then review evidence that both ways of using context could potentially be helpful in Japanese.

Top-down information

Evidence for top-down information accounts

Experimental and computational work suggests that people can and do make use of higher-level linguistic information in a top-down fashion—at least on synthesized or carefully controlled laboratory speech. In various experiments, researchers have presented participants with stimuli that have portions of removed or degraded acoustic information, and shown that participants make use of contextual information to compensate. Warren (1970) showed that when adult participants are played a sentence with a single phone (and its transitional cues) completely removed and replaced with a cough, they nonetheless report hearing the sound, suggesting that linguistic context affects speech perception.

This is true even when full acoustic information is available. In a classic study, Ganong (1980) played participants acoustic continuua that ranged between a non-word and a word (i.e. from dask to task, or from dash to tash), and showed that participants were biased towards categorizing the initial sound in a way that resulted in a word. They were more likely to classify a given sound as /t/ for dask-task, but as /d/ for dash-tash, suggesting that listeners use context (in this case, lexical information), in addition to acoustics, to constrain their categorization decisions. Similar work has shown that phonotactic constraints also affect categorization decisions, such that listeners are more likely to classify a particular sound in a way that adheres to, rather than violates, the phonotactics of their language (e.g., Brown and Hildum, 1956; Massaro & Cohen, 1983).

Particularly relevant to our test case, there is experimental evidence from Moreton and Amano (1999) that Japanese speakers may make use of higher-level contextual information to make decisions about vowel length. Words in Japanese fall into four main groups based on their historical origin (e.g., Foreign words, Sino-Japanese words, etc.) and these word groups differ in their properties. For example, long /a/ occurs in Foreign words, but not Sino-Japanese words, and Sino-Japanese and Foreign words have different frequency distributions over consonants (e.g., /p/ is frequent in Foreign words, but very rare in Sino-Japanese words, and vice versa for /hy/). Taken together, this means that, for example, an /a/ vowel that co-occurs with a /hy/ vowel is almost certainly phonemically short, while an /a/ vowel that co-occurs in a word with a /p/ could also be long. In a series of experiments, Moreton and Amano (1999) showed that Japanese listeners make use of these regularities when identifying vowels: the other consonants a particular vowel token co-occurred with affected whether participants categorized it as short or long, again showing that top-down information affects adults' categorization.

Children also seem to use top-down information to guide acquisition and processing. A number of studies have shown that both adults and infants use lexical context while acquiring sound categories (Thiessen, 2007; Swingley, 2009; Feldman et al., 2013b). For example, Feldman et al. (2013b) showed that adults and infants were more likely to assign acoustically similar vowels (/ɔ/ vs. /ɑ/) to different sound categories when they were not exposed to minimal pairs between them (i.e., when they did not occur in the same phonetic contexts) than when they were exposed to minimal pairs (i.e., when the vowels occurred in identical phonetic contexts). In addition, Feldman et al. (2013a) showed that a computational model that made use of information about the word frames that sounds occurred in resulted in an improvement in phonetic category learning over models that did not incorporate lexical information.

The idea that higher-level information influences speech perception and language acquisition has been replicated many times over, and is mostly accepted in the field. Most of the support for this idea, however, comes from work on simplified speech data. Furthermore, the model from Feldman et al. (2013a) was recently applied to the problem of Japanese vowel length we study here, and was found to be ineffective on spontaneous speech (Antetomaso et al., 2017). Therefore, there is some recent doubt that this strategy could be helpful on spontaneous speech. However, phonemically short vowels and phonemically long vowels have been shown to differ in the contexts that they are likely to occur in Japanese, so there is potentially signal that would be helpful to a listener relying on such a strategy. We discuss this evidence in the following section.

Evidence that there is top-down information in Japanese

With the exception of Moreton and Amano (1999) and Antetomaso et al. (2017), there has not been much work on studying the role of top-down information in the acquisition and processing of Japanese vowel length. However, there is independent evidence that there are systematic differences between short and long vowels in the types of contexts/environments they occur in that listeners could make use of.

First, different vowel qualities have different relative probabilities of short and long vowels, as seen in Fig. 1. In particular, long vowels make up a greater proportion of /o/ vowels than /a/ vowels.

Short and long vowels also differ in the types of sounds they co-occur with, due, for example, to properties of various subsets of the Japanese lexicon as seen in the previous section (Moreton & Amano, 1999). Similarly, in some dialects of Japanese, long vowels do not occur before nasals, due to phonotactic constraints. Vowels also tend to be phonologically short when adjacent to long consonants. Therefore, the adjacent sounds of a vowel could potentially provide useful, disambiguating information about the length status of a target vowel (Isei-Jaakkola, 2004).

Finally, prosodic position could also be helpful, as phonemically long vowels are less likely to occur domain-finally (e.g., Kubozono, 2002). As a result, listeners could exploit the prosodic position of the vowel to help determine the length of a vowel: they could be biased towards classifying a domain-final (e.g., word-final) vowel as short.

Overall, there are various patterns due to phonological, historical, or lexical reasons that result in differences in how likely short versus long vowels are to occur in particular contexts. Listeners could exploit this information in a top-down fashion to categorize and learn the vowel length contrast. We test how effective this strategy could be by applying it to the Japanese vowel length contrast.

Normalization

Evidence for normalization

A body of experimental work has been used to argue that listeners can and do normalize when making categorization decisions—at least on the carefully controlled laboratory speech or synthetic speech that is typically studied (but see Johnson, 1997, 2006; Pierrehumbert, 2002, which argue against normalization). This work generally shows that listeners' perception of a particular sound can change by modifying the context it appears in. As we saw, modifying the context can also change listeners' perception if they are relying on a top-down information strategy. Therefore, this evidence is insufficient to argue uniquely for normalization as a useful strategy when the contextual factor being normalized out could also prove helpful in a top-down information account (e.g., neighboring sounds, prosodic position).

However, for contextual factors that do not influence which category is more likely to be produced (e.g., speech rate and speaker), there is extensive evidence that listeners factor out systematic variability from the acoustics of lab speech, though these studies do not necessarily pinpoint normalization as the involved mechanism—as opposed to adaptation, for example (Kleinschmidt & Jaeger, 2015). In this section, we review the literature that has been taken as support for normalization in the field, even if it could also be used to argue for adaptation or top-down accounts, but we return to the issue of how to properly dissociate these accounts, and what evidence could be taken as unequivocal support for one of these theories, in the General Discussion.

Nearey (1978) studied synthetic speech and showed that listeners factor out systematic variability stemming from speaker. His study showed that listeners' category boundaries were shifted upward in F1 and F2 when a target sound followed a vowel that sounded like a child produced it instead of a man. This type of result has been repeatedly reported (e.g., Strand and Johnson, 1996).

Mann and Repp (1980) studied synthetic speech and argued that listeners also take into account coarticulatory influences. They played participants a fricative from the /ʃ/ to /s/ continuum, followed by either the rounded vowel /u/ or the unrounded vowel /a/. They found that participants were more likely to identify the fricative as /s/ when it was followed by /u/ than when it was followed by /a/. Fujisaki and Kunisaki (1978) found a similar effect with Japanese speakers.

Various studies have also shown that listeners take into account the influence of speech rate. These findings are particularly relevant to the Japanese vowel length case, because they offer evidence that participants using durational cues also take into account systematic variability due to context. Using synthesized speech, Fujisaki et al. (1975) studied Japanese listeners' perception of the contrast between short and long consonants as a function of contextual speech rate. They played participants synthesized syllables ranging from /ise/ to /isse/ and found that the absolute duration at which participants' percept changed from a short consonant to a long consonant was affected by the speech rate of the utterance.

Analogous effects have been found for English vowel and voicing categorization, as well as /b/-/w/ distinctions, and recent work has even suggested that changing the speech rate of neighboring consonants can cause listeners to not hear or insert entire function words (e.g., Ainsworth, 1974; Verbrugge et al., 1976; Ainsworth, 1973; Dilley and Pitt, 2010; Miller & Liberman, 1979; Minifie et al., 1977 Summerfield, 1981). Overall, the general finding that listeners' perceptions of a sound (or even a word) change as a function of the context it occurs in has been replicated many times over (e.g., Crystal & House, 1990; Miller, 1981; Miller et al., 1984; Miller et al., 1997; Newman & Sawusch, 1996; Pickett & Decker, 1960; Sawusch & Newman, 2000; Wayland et al., 1992, 1994) and has often since been taken as evidence for normalization.

However, as mentioned above, recent work has also suggested that some of the experimental findings that have been taken as evidence for factoring out systematic variability may actually be support for participants making use of top-down information. In a classic study, Port

and Dalby (1982) argued that listeners use durations of neighboring sounds, in addition to utterance speech rate, to calibrate (or normalize) the durational cues of the target sound. They ran several experiments studying English listeners' voicing judgments in synthesized minimal pairs like *rapid* versus *rabid*. They showed that the duration of a vowel neighboring a stop could affect listeners' perception of whether that stop was voiced or voiceless (Port & Dalby, 1982), and similar findings have been reported in other research as well (e.g., Boucher, 2002; Summerfield, 1981). These findings have classically been interpreted as evidence that listeners factor out the effect of speech rate, and use the relative duration of the stop's closure duration to the neighboring vowel to do so. However, Toscano and McMurray (2012) argued that these same findings were consistent with the alternative idea that listeners are using neighboring vowel duration as a direct cue to the voicing of the target stop (parallel to closure duration or VOT), rather than normalizing for it. Although this reinterpretation has been discussed with reference to a particular set of studies (Boucher, 2002; Port & Dalby, 1982; Summerfield, 1981), it raises the interesting possibility that other studies arguing for normalization could also be used as evidence for a top-down information account, rather than for normalization. In particular, this holds true for all studies where the contextual factor that is normalized out could also prove helpful in a top-down information account—for example, neighboring sounds.

Experimental findings in support of normalization have been supplemented by recent computational work, which has generally found that models that normalize for systematic variability achieve better sound category identification results, and better match human performance than models that do not.

McMurray and Jongman (2011) showed that a model that normalized for multiple contextual factors better matched human behavior than a model that did not. They took lab recordings of the 8 English fricatives /f, v, θ, ð, s, z, ʃ, ʒ/ produced in the initial position of a CVC syllable, where the vowel was one of six vowels, and the final consonant was always /p/. They had measurements of 24 cues from these tokens (Jongman et al., 2000). They presented a subset of these recordings to listeners and asked them to identify the syllable-initial fricative. They then used a method from Nearey (1990) and Cole et al. (2010), that we also make use of in this paper, to compare whether normalized or unnormalized cues led to more human-like identification in their model. They found that the version that normalized for speaker and neighboring vowel yielded a better match to human categorization than the version that used unnormalized cues. This finding has been replicated many times, sometimes with different normalization implementations (Apfelbaum & McMurray, 2015; Cole et al., 2010; Richter et al., 2017); however, these models have, for the most part, only been applied to controlled and well-enunciated lab speech.

There has also been some work looking at normalization in acquisition. Dillon et al. (2013) considered the problem of learning the phonological system of Inuktitut, using elicited speech. Inuktitut has three vowels (/i/, /u/, /a/), but these vowels are lowered when followed by uvular consonants. The researchers found that a computational model that learned from the unnormalized vowel formants failed to learn the correct sound categories of Inuktitut (learning six categories instead), but when they subtracted out the influence of the neighboring uvular and used these normalized vowel formants as input to the model, it was able to learn the correct three categories of Inuktitut, just as infants do, suggesting that normalization is a possible strategy that infants could be using in learning the sounds of their language.

Because most cognitive research has focused on carefully controlled laboratory research or synthesized speech, and because many of the empirical studies supporting normalization could also be in support of top-down information accounts, it is hard to draw strong conclusions about the efficacy of normalization in naturalistic listening environments. This paper further tests its efficacy in naturalistic listening situations.

Evidence that factoring out systematic variability might be useful in Japanese

Factors other than phonological length influence the duration of Japanese vowels, and could cause the overlap between short and long vowels. This is variability that normalization could, in principle, help reduce.

First, the quality of a vowel systematically affects its duration. Hirata (2004) had Japanese participants produce disyllabic non-words in a carrier phrase and found that the vowel /e/ tended to be acoustically longer than /o/ and /u/. In addition, Bion et al. (2013) analyzed a corpus of spontaneously produced infant-directed speech and found that low vowels were acoustically longer than high vowels.

Japanese vowels are also acoustically shorter in fast speech than slow speech, all else being equal. Hirata (2004) had participants produce Japanese sentences (including non-words) at three different speech rates—slow, normal, and fast speech—and found that as the speech rate quickened, the vowels became acoustically shorter.

There is evidence that the prosodic position of a sound influences the duration of a vowel, as well. There are various prosodic phrase types in Japanese—utterances are made up of intonational phrases (IPs), which are, in turn, made up of accentual phrases (APs)—and a vowel's position relative to these phrasal units affects its duration. Bion et al. (2013)

found that in spontaneous infant-directed speech, vowels are acoustically longer when followed by an intonational phrase boundary, but acoustically shorter when followed by a word boundary that is not an intonational phrase boundary. Martin et al. (2016) calculated the average mora duration in various prosodic positions in spontaneously produced adult- and infant-directed speech. They found that the average mora duration increases, moving from more phrase-medial to more phrase-final positions (from phrase-medial, to AP-final, to IP-final, to utterance-final position), which suggests that segments are acoustically lengthened phrase-finally.

Some work has also shown that neighboring sounds can influence the duration of a vowel. For example, several studies have found that vowels tend to be acoustically longer before a geminate than a singleton consonant (Fukui, 1978; Han, 1994; Kawahara, 2006). Other work has suggested that accented vowels tend to be acoustically longer than unaccented vowels (Hirata, 2004).

Finally, although these factors have not been studied in Japanese, work in other languages suggests sounds may be acoustically lengthened at the beginning of a phrase, in addition to phrase-finally (Keating et al., 2004; Rakerd et al., 1987), that sounds may be acoustically shorter in function words rather than content words, and that other features of neighboring consonants, for example voicing, may affect the duration of the target vowel (House, 1961; Luce & Charles-Luce, 1985; Umeda, 1975; Van Santen, 1992). In sum, there are a priori reasons to believe that normalization could be helpful for the vowel length contrast.

Testing the efficacy of top-down information on Japanese vowel length

In this section, we test how helpful using contextual information as top-down information can be in categorizing Japanese vowels, by testing to what extent it helps separate short and long vowels in spontaneous speech. We compare various logistic regression models that make use of higher-level contextual factors to the baseline logistic regression that only uses unnormalized duration and formants.

Data

The data we use come from the RIKEN Japanese Mother-Infant Conversational Corpus (R-JMICC) (Mazuka et al., 2006). It is spontaneously produced child-directed speech. Mazuka et al. (2006) collected the data by recording the speech of 22 mothers who visited the lab with their 18- to 24-month-old children. The mothers first played with their child with picture books for about 15 min. They then played with their child with toys for about 15 min. Finally, a female experimenter came into the room and talked to the mother. The mothers' speech in the first two parts, where they interacted only with their child, was labeled as child-directed speech. The mothers' speech in the third part, where they interacted with the experimenter, was labeled as adult-directed speech. The corpus consists of about 14 total hours of speech, and is labeled for both phonetic and prosodic information.

We extracted information about each of the vowels produced by the mothers, but excluded singing, coughing, devoiced vowels, diphthongs, and any segments that the researchers could not transcribe. We also excluded any vowels that were not labeled with prosodic information. This left 92003 total vowels, 30035 of which were in the adult-directed section of the corpus and 61968 of which were in the child-directed section of the corpus. All of the analyses we report were run on the child-directed part of the corpus; however, we also ran these analyses on the adult-directed parts and did not find substantial differences in model performance (see Supplementary Materials).

We extracted both acoustic information and contextual information about each vowel, as described below. The list of the features we extracted is also compiled in Table 1.

Acoustic cues

- **Duration**: We extracted the duration of each vowel in milliseconds.
- **Formants**: We extracted the first three formants, and used these as direct acoustic cues to vowel length. While duration is thought to be the primary acoustic predictor of vowel length in Japanese, previous work has shown that spectral information can improve categorization performance (e.g., Arai et al., 1999; Kinoshita et al., 2002; Lehnert-LeHouillier, 2010). The formants were automatically extracted using Praat (Boersma, 2001) in previous work on this corpus (Antetomaso et al., 2017) and we used the formant values at the midpoint of the vowel.

Contextual factors

In addition to extracting acoustic information, we also extracted contextual information about each vowel that has been shown to be relevant for normalization or top-down information accounts:

- **Vowel quality**: This was a categorical variable that took one of five values (/a/, /e/, /i/, /o/, /u/) and was taken from the coding of what the mother said.
- **Speaker**: This was a categorical variable, with 22 different possible speaker values.
- **Neighboring sounds**: We extracted the identity of the previous sound and the following sound (both

categorical variables), as labeled by the phonetic transcription. This was marked as '#' if the vowel was preceded by silence. Because the vowel length contrast is thought to be learned later than other types of contrasts (Sato et al., 2010), it is reasonable to assume that infants can make use of the other contrasts in their language to learn vowel length.

- **Prosodic position**: We represented prosodic position in three different ways. First, we extracted a categorical variable that ranged from 1 to 4, which indicated whether the word that the vowel occurred in was not phrase-final at all (1), was AP-final (at the end of an accentual phrase) (2), was IP-final (at the end of an intonational phrase) (3), or was utterance-final (4) (BI). Second, we extracted a second categorical variable that ranged from 1 to 4, which indicated whether the word that the vowel was in was not phrase-initial (1), was AP-initial (2), was IP-initial (3), or was utterance-initial (4) (BIstart). Third and finally, we extracted a vector of length 12, which represented the prosodic position of the vowel itself in a bit more detail. Namely, each element of the 12-long vector was a binary categorical variable, with three elements of the 12 elements corresponding to whether the vowel itself was word-initial, word-medial, word-final, three to whether the vowel itself was AP-initial, AP-medial, AP-final, three to whether the vowel itself was IP-initial, IP-medial, IP-final, and three to whether the vowel itself was utterance-initial, utterance-medial, utterance-final. That is, while the first two categorical variables represented the prosodic position of the word the vowel was in, and would, thus, have the same value for every vowel in a given word, the vector represented the prosodic position of the vowel itself.
- **Accented?**: This was a binary variable that took a value of 1 if the vowel was accented and 0 if it was not.
- **Speech rate**: We extracted the duration of the immediately preceding and the immediately following sounds, as proxies for speech rate. If the vowel was immediately preceded (or followed) by silence, we did not use the duration of the silence, but instead used the average duration of the immediately preceding (or following) sound, averaged across all vowels that were not preceded (or followed) by silence.
- **Condition of the vowel**: This was a categorical variable with a value of 'B' if the vowel occurred when mother and child were playing with books and 'T' if it occurred when mother and child were playing with toys. We include this to account for the possibility that the mothers' speech was consistently different (e.g., more or less clear) while playing with books than toys.
- **Part of speech**: This was a categorical variable taken from the annotation in the corpus. In our simulations, we either use full part-of-speech information, or simplified part-of-speech information, which only considers the distinction between function and content words. We vary this because we want our results to be applicable to language acquisition. Infants show evidence of distinguishing function vs. content words using acoustic correlates as early as birth (Shi et al., 1999; Shi & Werker, 2001), so it is relatively likely that they can make use of this knowledge in learning the contrast. However, it is less clear that they could make use of full part-of-speech information for this task, as cross-linguistic evidence suggests that infants have much of this knowledge only after Japanese infants have learned the vowel length contrast (Höhle et al., 2004; Mintz, 2006; Shi & Melançon, 2010). That being said, He and Lidz (2017) show evidence that infants know the distinction between nouns and verbs as early as 12 months, so while infants might not have complete part-of-speech information, they may be able to use more than just the distinction between function and content words for acquiring the vowel length contrast. Testing function vs. content word distinctions in addition to full part-of-speech allows us to determine whether our qualitative results hold true regardless of what infants know.

Methods

We compare the results of four models—divided into three types of models. The baseline model is a logistic regression that learns to predict short/long from only a vowel's absolute duration and formant values (Baseline). The next two models are logistic regressions that learn to predict short/long from contextual factors listed previously and in Table 1, in addition to absolute acoustic cues (Acoustic and Top-Down Information Models). The first of these makes use of all of the contextual factors listed in Table 1, with part-of-speech simplified to just indicate whether the word was a function or content word. The second of these makes use of all of the contextual factors, including detailed part-of-speech, exactly as annotated in the corpus. Finally, we test how much signal just the contextual factors provide, by running a logistic regression model that learns to categorize vowels as short/long using only the contextual factors, without any access to acoustic information (Top-Down Information Model Without Acoustics). Studying the results of this model will allow us to understand how much of the work context does. That is, it will reveal how many vowels can be identified just by the context they occur in, without even turning to acoustic information, or, in other words, how much information is lost when acoustics are removed.

We split the dataset into a training subset (90% of the data) and a test set (the remaining 10% of the data), keeping

the proportions of short and long vowels equal in the two sets. The training and test sets consisted of the same tokens for all of the simulations run in this paper.

Once the logistic regression equations were estimated from the training set, we simply applied each equation to the vowels in the unseen test set to make a prediction about whether that vowel was short or long, as described previously. We compared the models' predictions to the true labels. We report two types of evaluation metrics for each tested model.

First, we report overall categorization accuracy, which is simply the percentage of all of the vowels in the test set that the model categorized correctly, as well as accuracy on just the short vowels and accuracy on just the long vowels. Second, we report the Bayesian Information Criterion (BIC) for each model, computed over the training set. The BIC is a common metric used to select between different models (Schwarz, 1978). The benefit of the BIC is that it balances how well the model works (the likelihood of the model given the data) with how complicated the model is (how many parameters it uses), so it will prefer simpler models, all else being equal. The BIC is calculated as follows and lower values are better:

$$\mathrm{BIC} = -2 * ln(L) + k * ln(n) \tag{5}$$

where L is the likelihood of the model given the data, k is the number of parameters, and n is the number of samples.

We ran each model ten times and averaged performance across these ten runs.

Results

The results from this analysis on child-directed speech are summarized in Table 2.

Baseline model

The baseline model reached an overall accuracy of 91.1%. It correctly categorized 99.1% of short vowels, and 12.2% of long vowels. It had a BIC of 28716. Because 90.9% of vowels in the R-JMICC corpus are short, this model performs comparably to a model that simply categorizes every incoming vowel as short, and has failed to learn anything meaningful about the distinction between short and long vowels.

Acoustic and top-down information model

The following models used contextual factors as direct predictors to category membership, in addition to using absolute duration and formant values. When part-of-speech was simplified to the distinction between function and content words, the model reached an overall accuracy of 95.2%, correctly classifying 98.8% of short vowels and 59.0% of long vowels. The BIC was 15193. When we included full part-of-speech information, the model achieved an overall accuracy of 95.7%, correctly classifying 98.8% of short vowels and 63.9% of long vowels. The BIC was 13106. Including additional part-of-speech information led to performance improvements, but both models substantially outperformed the baseline model. Table 3 analyzes the role of each contextual factor, by showing how well a model with each factor as its only piece of top-down information performs. The most helpful factors include part-of-speech, the previous sound, the following sound, whether the sound is accented, prosodic information (BI and BIstart as described previously), as well as vowel quality.

Top-down information model without acoustics

Even without any acoustic information, only contextual information, the final top model achieved an overall accuracy of 94.5%, correctly classifying 98.6% of short vowels and 54.0% of long vowels. The model BIC was 16301. That is, although there was a slight dip in performance when we removed acoustic information, top-down information models can still perform well even without any acoustic information, suggesting a large role for context.

Table 2 Summary of top-down information results from the R-JMICC dataset

Model	Accuracy	Short accuracy	Long accuracy	BIC
Baseline	91.1	99.1	12.2	28716
Top-down information (with simplified POS)	95.2	98.8	59.0	15193
Top-down information (with POS)	95.7	98.8	63.9	13106
Top-down information (with POS, no acoustics)	94.5	98.6	54.0	16301

Table 3 Results showing the contribution of each contextual factor. This table shows model results when each available contextual factor is included as the only piece of top-down information in a logistic regression model. Factors are ranked from lowest BIC (best) to highest BIC (worst)

Factor	Accuracy	Short accuracy	Long accuracy	BIC
Baseline	91.1	99.9	12.2	28716
Part-of-speech	92.0	99.1	21.0	23165
Previous sound	92.1	99.0	23.0	24231
Following sound	91.7	99.1	18.9	25572
Accented?	91.4	99.1	15.0	26446
BIstart	91.8	99.3	17.5	26870
Quality	91.2	99.0	13.6	27372
BI	91.5	99.1	15.9	27673
Word-medial?	91.0	99.0	11.7	28205
Utterance-final?	91.2	99.0	13.4	28362
Speaker	91.1	99.1	12.4	28463
IP-final?	91.2	99.1	13.3	28500
Word-initial?	91.2	99.1	12.2	28523
Previous sound duration	91.3	99.1	13.1	28540
Word-final?	91.1	99.0	12.2	28563
AP-final?	91.2	99.1	13.3	28605
AP-initial?	91.2	99.1	12.4	28687
Utterance-initial?	91.2	99.1	12.2	28697
AP-medial?	91.2	99.1	12.4	28701
IP-initial?	91.2	99.1	12.2	28703
IP-medial?	91.1	99.1	11.8	28723
Following sound duration	91.2	99.1	12.2	28724
Condition	91.2	99.1	12.2	28726
Utterance-medial?	91.1	99.1	12.2	28727

Discussion

In these analyses, we investigated the hypothesis that infants and adults learn and process the Japanese vowel length contrast by combining bottom-up acoustic cues with top-down expectations about which category is likely to occur in a particular context. To implement this hypothesis, we included contextual factors listed in Table 1 as direct predictors of category membership in the logistic regression model (in addition to absolute acoustic cues), and compared its performance against a model that only makes use of absolute acoustic cues as predictors.

We found that including these additional contextual factors as predictors drastically improved accuracy and lowered BIC scores, suggesting that this method does quite well at separating short vowels from long vowels. Given the relatively small set of factors we used—for example, the only word-level information we used was part-of-speech—it is quite impressive that the model achieved this level of performance, and it suggests that this may be a hypothesis worth pursuing as a way that infants could learn and adults could process the Japanese vowel length contrast.

In fact, although excluding acoustic information did hurt performance, a model relying on contextual information alone still performs very well. Even without *any* acoustic information, this model can correctly identify nearly all short vowels and more than half of all long vowels. This illustrates just how much signal there is in contextual information.

This work shows that top-down information could be very useful in adult speech perception, and also has implications for acquisition. Although these are supervised models that have much more information available to them than infants learning language, and there is still work to be done to show that this is a strategy that could be helpful in acquisition, our analysis does reveal that there is signal to separate short and long vowels that could be exploited in a future unsupervised model.

Testing the efficacy of normalization on Japanese vowel length

In this section, we test whether normalization can help categorize Japanese vowels, by comparing models that use normalized acoustic cues to models that use unnormalized acoustic cues.

Data

The data are exactly the same from the first analysis, but the contextual factors listed in Table 1 are normalized out of the acoustics (as described below in the Methods section), instead of being included as independent predictors in the logistic regression categorization model. The same training and test sets are used as in the previous analysis, which allows us to directly compare results.

Methods

In testing the efficacy of normalization on spontaneous speech, we implement and test two normalization methods. First, we apply methods from previous work (Cole et al., 2010; McMurray & Jongman, 2011; Nearey, 1990) to the Japanese vowel length contrast, by using linear regression to normalize out systematic variability from vocalic acoustic cues. Second, we implement normalization using a neural network, which has the advantage over past implementations that it can represent more powerful, non-linear normalization functions. Our results can only directly tell us about the two implementations we use, and future work should investigate other ways of normalizing. We return to the question of how these results would generalize to other contrasts in the General Discussion.

Normalization implementation

We use either unnormalized or normalized acoustic cues as predictors of vowel length. Using unnormalized cues simply involves representing the absolute acoustic cues, so this section will focus on how we implement normalization. The basic idea underlying both of the implementations we use is to learn a function that predicts acoustic features (duration and formants) of a vowel from the context that a vowel occurs in (i.e., vowel quality, speaker, prosodic position). Once we learn this function, we can make a prediction about a vowel's duration and formants based on everything we know about where it occurs. We can then use the residuals, or the difference between how long we expect the vowel to be given all of the factors and how long it actually is, to represent a normalized version of this vowel. That is, we have excluded the influence of contextual factors and have recoded the acoustic cues in terms of their difference from expected values. Once we learn this equation from the training set, we recode both the training set and the test set in normalized terms. We use two different methods for representing the function between contextual factors and acoustic cues, linear regression and neural nets, which can learn non-linear functions, which we describe in turn.

Linear regression as normalization Following previous work, we first use linear regression to factor out systematic variability (Cole et al., 2010; McMurray & Jongman, 2011; Nearey, 1990). Linear regression models represent a relationship between a continuous dependent variable and a set of independent variables. In this particular case, we try to estimate an equation that can predict what the acoustic features (duration and formants) of a vowel should be from its context. Each of the factors (e.g., vowel quality, speaker, prosodic position from Table 1) is weighted and combined linearly to yield a prediction. That is, given the factors $x_1, x_2, ..., x_n$, linear regression models take the form:

$$\text{acoustic cue} = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + ... + \beta_n * x_n \quad (6)$$

Learning this function involves learning an intercept (β_0), as well as a weight for each cue ($\beta_1 ... \beta_n$). The data it learns from consist of the information we want to factor out of the acoustic cues, as well as the known acoustic cue values of the vowel, and weights are learned so as to minimize the error in predicting the duration of the vowel.

Neural networks as normalization The linear regression models we use can only learn normalization functions and do not include interactions, even though previous work did. This is because our analyses use a total of 23 contextual factors, so considering all possible interactions would be computationally difficult. To test the possibility that our linear regression without interactions was insufficient to handle spontaneous speech, we also implemented normalization using a neural network. We train a neural network on the training set to predict the duration and formants of a vowel token from its context. Once we have a trained neural network, we use it to predict expected acoustic cues for each vowel, subtract them from the vowel's true acoustic values, and input this into a logistic regression model.

We use a simple feed-forward neural network. We use five-fold cross validation on the training data to tune parameters of the neural network. We manipulate the number of hidden layers, the batch size, the number of nodes

in the hidden layers (either keeping this constant for all of the layers or decreasing the number of nodes progressively deeper into the network), learning rate, number of epochs, and regularization factors. We choose the parameters that minimize average mean squared error on the training set.

Logistic regressions

To test the efficacy of normalization, we compare seven total logistic regression models, which can be grouped into three types of models. The first model is the baseline, which as before uses absolute (unnormalized) duration and formants to predict category membership (short or long). Then, for each type of normalization (linear regression and neural networks), we run three models. First, we regress out all of the contextual factors listed in Table 1 with part-of-speech in simplified form (i.e., function/content word distinctions). Second, we regress out all of the contextual factors listed in Table 1 including full, detailed part-of-speech information. In both of these models, the normalization function is trained completely independently of the subsequent logistic regression. That is, the normalization function is not trained to maximize categorization performance. The third and final model is an oracle model: we choose the subset of contextual factors from Table 1 that maximizes categorization performance, which gives us an estimate of the upper bound on normalization performance. This is useful because it is possible that we are wrongly including some factors in the first three models and underestimating the efficacy of normalization. Running this oracle model allows us to see what the best normalization performance could be.

Results

A summary of the results is presented in Table 4.

Unnormalized model

The baseline model is identical to the baseline model from the previous analysis and uses unnormalized duration and formants as predictors of category membership, without running any linear regression models. As a reminder, this logistic regression model reached an overall accuracy of 91.1%. It correctly classified 99.1% of short vowels and 12.2% of long vowels. Its BIC was 28716.

Linear regression normalization models

When all of the contextual factors with simplified part-of-speech (function vs. content word) were regressed out, the model had an overall categorization accuracy of 91.2%, correctly classifying 99.5% of the short vowels and 8.3% of the long vowels. It had a BIC of 30774. The set of factors used accounted for 26.8% of the variance in duration, 23.0% of the variance in F1, 40.2% of the variance in F2, and 8.1% of the variance in F3.

When all of the contextual factors including full part-of-speech information were regressed out, the model had an overall categorization accuracy of 91.2%, correctly classifying 99.6% of the short vowels and 7.6% of the long vowels. It had a BIC of 30990. The set of factors used accounted for 27.8% of the variance in duration, 23.1% of the variance in F1, 40.3% of the variance in F2, and 8.3% of the variance in F3. Figure 4 plots the normalized durations by vowel length for this model.

Finally, the oracle normalization model included the following five contextual factors: speaker, whether the vowel itself was word-final, whether the vowel itself was AP-final, whether the vowel itself was IP-final, and whether the vowel itself was utterance-final. This oracle model had an overall accuracy of 91.2%, and correctly classified 99.0% of the short vowels and 13.6% of the long vowels. It had an

Table 4 Summary of normalization results on R-JMICC corpus

Model	Accuracy	Short accuracy	Long accuracy	BIC
Unnormalized baseline	91.1	99.1	12.2	28716
Linear regression normalization All factors with simplified part-of-speech	91.2	99.5	8.3	30774
Linear regression normalization All factors with full part-of-speech	91.2	99.6	7.6	30990
Oracle linear regression normalization	91.2	99.0	13.6	28122
Neural network normalization All factors with simplified part-of-speech	91.1	99.8	5.1	32356
Neural network normalization All factors with full part-of-speech	91.1	99.7	5.8	31738
Oracle neural network normalization	91.2	99.0	13.4	28188

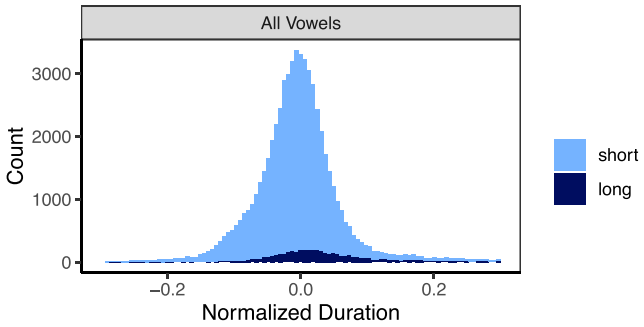


Fig. 4 Distribution of normalized durations (through linear regression). Normalizing does not appear to decrease overlap between short and long vowels

overall BIC of 28122. The set of factors that resulted in the best categorization performance accounted for 11.7% of the variance in duration, 3.6% of the variance in F1, 3.3% of the variance in F2, and 3.8% of the variance in F3.

Neural network normalization models

When all of the contextual factors with simplified part-of-speech (function vs. content word) were normalized out, the model had an overall categorization accuracy of 91.1%, correctly classifying 99.8% of short vowels, and 5.1% of long vowels. The BIC was 32356.

When all of the contextual factors, including fully detailed part-of-speech information, were normalized out, the model reached an overall categorization accuracy of 91.1%, correctly classifying 99.7% of short vowels, and 5.8% of long vowels. The BIC was 31738.

Finally, the oracle model normalized out the following factors from the acoustics: whether the vowel itself was word-final, whether the vowel itself was AP-initial, whether the vowel itself was AP-final, and whether the vowel itself was utterance-final. The oracle model had an overall accuracy of 91.2%, and correctly classified 99.0% of the short vowels and 13.4% of the long vowels. It had an overall BIC of 28188.

Discussion

Previous work has argued that normalization can be helpful in acquisition and processing (Cole et al., 2010; Dillon et al., 2013; McMurray & Jongman, 2011); however, our results on Japanese vowel length did not lend additional support to this hypothesis. We compared the Japanese vowel length categorization performance of a logistic regression model that used unnormalized acoustic cues to the performance of various logistic regression models that used normalized acoustic cues. We considered two different normalization implementations, and three different instantiations of normalized cues for each. The first normalized all available contextual factors, with simplified part-of-speech information (i.e., whether the word containing a vowel was a function or content word). The second normalized all available contextual factors, including detailed part-of-speech information. The third and final normalized out the subset of contextual factors that led to best categorization performance. Crucially, in the first two models, as in past work, normalization was not optimized to give the best categorization. The final, oracle model considered categorization performance in choosing how to normalize, giving it the best possible chance to succeed.

The main finding was that, at its best, normalization resulted in only a modest improvement in accuracy and BIC, regardless of which implementation we used. Although the overall accuracy of all of the models is quite high, just guessing that all of the vowels were short would result in similar results. Normalization never improved accuracy, but improved the BIC from 28716 for the unnormalized version to 28122 for the best normalized version. While this does constitute improvement, it is only modest improvement and a listener would need to learn precisely which factors they should normalize out. Of course, it is possible that results would be better on a larger corpus with more information about the contextual factors. We used previous and following sound duration as a proxy for speech rate, while other measures of speech rate might lead to better performance, and we return to this possibility in the discussion. However, given how prevalent normalization is in the field, the results are surprisingly bad and call into question the efficacy of normalization, at least in this task.

Although it is difficult to directly compare this degree of improvement to the improvement shown in past studies merely on the basis of accuracy, past studies that have implemented and tested normalization reported that normalization resulted in an increase in performance from 28.63% to 54% and 83.3% to 92.9% respectively (Cole et al., 2010; McMurray & Jongman, 2011). In comparison, in this work, the overall accuracy did not change depending on whether cues where unnormalized or normalized, and the long vowel accuracy increased from 12.2% to 13.6%—a much weaker increase in performance than has been observed previously.

It is important to emphasize that there are many ways that normalization could be implemented. Here, we have only tested one that has been proposed and well studied in the literature, as well as a neural network extension of it. It is possible that a different implementation of normalization could yield different results, and future work should test this, in addition to developing additional specific proposals for how normalization could operate. Nonetheless, we have some evidence that this normalization model may not be as helpful as previously thought, and we explore why in the

following sections. Understanding why normalization is not helpful will also let us speculate whether these results will generalize to other normalization implementations.

Do differences between controlled lab speech and spontaneous speech explain discrepancies in results?

Previous results found normalization to be helpful; however, our results were surprising in that they showed that normalization was unhelpful—even when the process was fully supervised. The biggest difference between previous work and our own is that most previous work has explored normalization on controlled and carefully enunciated lab speech, but our work looked at normalization on spontaneously produced speech. To bring these results more in line with each other, we apply the same normalization analyses we used on the R-JMICC Spontaneous Speech corpus to a corpus of read speech that more closely resembles controlled, lab speech. We find that the same linear regression normalization process that was not helpful on spontaneous speech is helpful on read lab speech, suggesting that the discrepancy in results between our work and previous work arises from differences between spontaneous and controlled speech.

Data

The data we use here come from Werker et al. (2007). The data consist of ten mothers teaching their 12-month-old infants a set of 16 nonce CVCV words, while looking at picture books together. This interaction included both a reading task, in which mothers were asked to read sentences containing the nonce words with pictures of the novel object (Werker Read dataset - Fig. 5), as well as a spontaneous speech task, in which mothers were asked to describe a scene that contained the novel object, using the nonce word as much as possible (Werker Spontaneous dataset - Fig. 6). The nonce words were made only using /i/ and /e/ as critical vowels, so the data do not contain any annotated instances of /a/, /o/, or /u/, unlike the R-JMICC corpus. The data were collected in the NTT Communication Science Laboratories in Keihanna, Japan and were labeled by trained phoneticians.

These data were much more similar to datasets that had previously been used to study normalization, though not identical. The experimenter controls the environment in which target sounds occur in, and artificially changes the statistical co-occurrences from that of naturalistic speech. This is especially true for the read portions, but still true for the spontaneous subset, in which researchers still decided what the nonce words were and, therefore, what sounds each target vowel was likely to occur next to. In addition, the productions are relatively well enunciated because the parents are trying to teach their children new words. However, that being said, even the read speech is less constrained than many speech recordings used for research, in which words are often recorded in isolation, or in highly constrained contexts like "Now I will say __."

It is also worth pointing out that though we, and the past researchers, refer to one portion of the Werker data as spontaneous, it is quite different than the spontaneous R-JMICC data, in that nonce words were used, only a subset of vowels are represented, mothers were instructed to teach their infants, and they were provided with highly constrained images to describe.

Given these data, we extracted information about each of the vowels produced by the mothers, excluding any segments that the researchers could not annotate with certainty. The read speech data consisted of 798 vowels, of which 381 (47.7%) were phonemically short vowels and the remaining 417 (52.3%) were phonemically long vowels. The spontaneous speech data consisted of 1382 vowels,

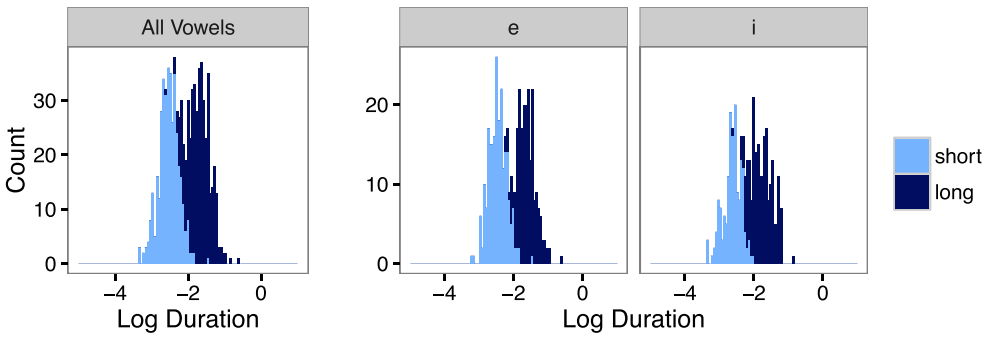


Fig. 5 Distribution of Werker Read IDS vowels (by log-duration). Log-durations will be negative whenever the vowel is less than a second long

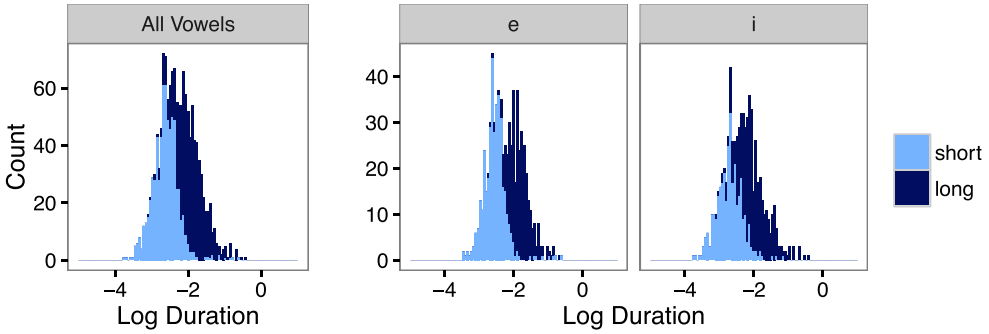


Fig. 6 Distribution of Werker Spontaneous IDS vowels (by log-duration). Log-durations will be negative whenever the vowel is less than a second long

exactly half of which were phonemically short and half of which were phonemically long. Similarly as for the R-JMICC data, the information we extracted was either used as an acoustic predictor or as a contextual factor to be normalized out.

Acoustic cues

As before, we used the duration of the vowel in milliseconds and the first three formants, as direct predictors of vowel length. These acoustic cues were either represented unnormalized as in the corpus, or underwent normalization through linear regression.

Contextual factors

We also extracted all of the contextual information that the original researchers had labeled on this dataset. The set of factors available for the Werker data is largely a subset of what was available for the R-JMICC dataset, with the exception that fundamental frequency (F0) was available for the Werker data while it was not extracted for R-JMICC. In addition, the labeling for prosodic position information was much simpler for the Werker data than for R-JMICC data, as described below. We collected the following pieces of information about each of the extracted vowels.

- **Vowel quality**: This was a categorical variable that took one of two values (/e/ or /i/) and was taken from the coding of what the mother said.
- **Speaker**: This was a categorical variable with one of ten different possible values.
- **Prosodic position**: Prosodic position took one of four values: 'Independent Word,', if the vowel occurred in a free-standing word, or 'Sentence Initial,' 'Sentence Medial,' or 'Sentence Final,' depending on whether the syllable the vowel occurred in was first, last, or in the middle of the sentence. This was controlled in the Werker Read data.
- **Neighboring sounds**: We extracted the previous and following sound. Unlike in the R-JMICC data, these were controlled to always be consonants.
- **Fundamental frequency**: We extracted the F0 at the vowel's midpoint.

Methods

We use linear regression to implement normalization. We did not use neural networks because they require large amounts of data, which we do not have for controlled lab speech, and because they performed worse than linear regression normalization models in our previous analyses. This is a limitation because it does not provide as strong of a test of the normalization hypothesis; however, our results will show that even just using a linear normalization function will suffice for getting better normalization results on the Werker data than the R-JMICC data. The methods were otherwise identical to the normalization methods run on the R-JMICC data set.[2]

Results

The results are summarized in Table 5.

Werker read speech data

Unnormalized model The unnormalized model achieves 91.4% overall accuracy on the Werker Read speech.

[2]Top-down information analyses are not included for the Werker corpus, as they would not be informative above and beyond the analyses performed on the R-JMICC corpus. Top-down information models make use of regularities in where different sounds occur. Any regularities that exist in the Werker corpus arise because researchers created them, and do not reflect true regularities that exist in naturalistic speech. As a result, we only present normalization analyses for the Werker corpus, which are informative above and beyond the analyses performed on the R-JMICC corpus, because they help explain why we see worse normalization performance than in past work.

Table 5 Comparison of normalization results on R-JMICC spontaneous speech corpus and Werker controlled laboratory data. The Werker speech corpus had a read component and a spontaneous component, but even the spontaneous component was relatively controlled by the experimenters, as the experimenters provided nonce words for the parents to teach their children

Data	Model	Accuracy	Short accuracy	Long accuracy	BIC
Werker read	Unnormalized	91.4	89.7	92.9	246
	Normalized (all factors)	86.1	83.9	88.1	399
	Normalized (best factors)	95.1	92.3	97.6	105
Werker spontaneous	Unnormalized	82.9	90.0	75.7	1072
	Normalized (all factors)	78.5	85.9	71.1	1219
	Normalized (best factors)	90.0	92.9	87.1	869
R-JMICC spontaneous	Unnormalized	91.2	99.1	12.2	28716
	Normalized (all factors)	91.2	99.6	7.6	30990
	Normalized (best factors)	91.2	99.0	13.6	28122

Although this is a similar overall accuracy to the R-JMICC spontaneous data, this corpus is much more balanced than the R-JMICC corpus. In the Werker Read speech, about 47.7% of the used vowels are short, compared to 90.9% in the R-JMICC corpus, so a strategy of simply categorizing every vowel as short (or long) will not yield as good results on the Werker Read Speech as on the R-JMICC corpus. The unnormalized model correctly classifies 89.7% of short vowels and 92.9% of long vowels, achieving a BIC of 246.

Normalized models When we normalized out all available factors, the model's overall accuracy is 86.1%, and it correctly classifies 83.9% of the short vowels and 88.1% of the long vowels. Its BIC is 399. That is, normalizing all available factors does not improve performance. When we instead choose the best subset of factors, the model no longer factors out the effect of the following consonant, and shows a boost in performance. It achieves an overall accuracy of 95.1%, a short vowel accuracy of 92.3%, a long vowel accuracy of 97.6%, and a BIC of 105.

Werker spontaneous speech data

Unnormalized model The unnormalized model achieves 82.9% overall accuracy on the Werker Spontaneous speech, and correctly classifies 90% of the short vowels and 75.7% of the long vowels. It achieves a BIC of 1072. Again, in the Werker Spontaneous speech, exactly 50% of the used vowels are short, so the unnormalized model substantially outperforms one that simply guesses that each vowel is short, unlike on the R-JMICC corpus.

Normalized models The model that normalizes out all available contextual factors listed previously and in Table 1 achieves an overall accuracy of 78.5%, correctly classifying 85.9% of short vowels and 71.1% of long vowels. When we allow subsequent categorization results to drive which subset of contextual factors are included in normalization, the model achieves an overall accuracy of 90.0%, and correctly classifies 92.9% of short vowels and 87.1% of long vowels. It achieves a BIC of 869. Similarly to the Werker Read speech, normalizing out all of the factors does not help, but depending on what factors are normalized out, normalization can help—and substantially.

Discussion

In this section, we applied the linear regression analyses that we applied to the R-JMICC spontaneous speech corpus to the Werker corpus. The idea was to test whether we would see similar normalization results as were previously reported, when we used data that more closely resembled that used in previous work. We found that normalization could help on the read speech, as well as the Werker spontaneous speech, even though it did not help when all available contextual factors were factored out.

That is, on the same contrast in the same language, normalization was helpful on carefully controlled lab speech, but was unhelpful on naturalistic, uncontrolled spontaneous speech. This suggests that normalization may be ineffective on spontaneous speech.

Another interesting finding was that the Werker Spontaneous speech patterned similarly to the Werker Read speech, instead of the R-JMICC Spontaneous speech. The overall results were worse on the Werker Spontaneous speech than on the Werker Read speech; however, normalization was helpful on the Werker Spontaneous speech, but not the R-JMICC Spontaneous speech. One reason for this could be that duration seems to be used differently by speakers in the R-JMICC data versus the Werker data, perhaps reflecting the fact that the Werker data is not nearly as natural as the R-JMICC data. In particular, in comparing Figs. 1 to 5 and 6, it seems that the contrast is being produced differently in the two datasets, such that duration is a much better

cue for vowel length in the Werker data than the R-JMICC data. There is less overlap between the short and long vowel categories in the Werker data: there is a duration such that all vowels that are acoustically longer than it are reliably long vowels. In the R-JMICC data, however, this is not the case: some short vowels are as acoustically long as the most acoustically lengthened long vowels.

It is important to emphasize that although both the R-JMICC and Werker Spontaneous speech datasets are referred to as spontaneous speech, they differed quite substantially in nature. In particular, in the Werker Spontaneous speech, mothers were producing nonce words that were created by researchers, were instructed to teach their infants, and were given pictures to describe. In the R-JMICC Spontaneous speech, mothers were given toys and books, but were given very little instruction, so they were free to talk about anything. It is important to keep these types of distinctions in mind when developing and comparing performance across various spontaneously produced speech datasets.

Overall, the simulations we have presented have disentangled normalization and top-down information accounts and evaluated their relative efficacy on relatively naturalistic, spontaneously produced speech. Our results from Japanese vowel length suggest that while top-down information accounts are extremely useful even on spontaneous speech, results that argue for this normalization model only hold for controlled laboratory speech and do not generalize to the type of spontaneous speech that listeners hear. These results force us to scrutinize the role this model and normalization, more broadly, can play in learning and processing, as well as the ways in which the primary cue for a distinction can shift based on the domain of speech.

In the following two sections, we consider what properties of spontaneous speech cause normalization to be ineffective. We provide simulations, followed by a theoretical analysis demonstrating that a listener that makes use of normalization will be impeded if sound categories in their input differ in the types of contexts they are likely to occur in.

Simulating how contextual category imbalances affect normalization performance

We showed that normalization can help reduce category overlap between Japanese short and long vowels when applied to controlled lab speech, but not when applied to spontaneous speech. What are the properties of spontaneous speech that make this normalization implementation ineffective?

In this section, we provide simulations that reveal that one property of spontaneous speech that seems to play an important role is the fact that categories do not occur uniformly across contexts in spontaneous speech, as they do in controlled lab speech. That is, imbalances in where categories occur—precisely the type of signal that is helpful in top-down information accounts—can hurt normalization. We provide an example from the Werker controlled lab speech, in which we take advantage of one contextual factor—the following sound—that is not balanced between short vowels and long vowels. In particular, the consonants /g/, /s/, and /z/ (three of the eight consonants used in the study) each followed either only short vowels or only long vowels. Even within following consonants that occurred both with short and long vowels, there were large imbalances in which vowels occurred with which consonants. These types of imbalances are uncommon in carefully controlled lab speech, where researchers ensure that each vowel occurs in each context—but are extremely common in spontaneous speech, which has phonotactic constraints and phonological alternations. We previously showed that when the effect of the following consonant was one of the contextual factors normalized out, normalization hurt on Werker Read speech, but when it was not normalized out, normalization helped. Here we show that this is because of the large imbalance observed between short and long vowels, by artificially balancing the dataset and showing that normalizing out the following consonant becomes helpful once it is balanced.

Methods and data

The data we use come from the Werker dataset, as described previously. We test the efficacy of normalization (implemented via linear regression) on various subsets of the Werker Read speech data. We limit normalization to one contextual factor—following consonant.

The first dataset is simply the full dataset (Full). As described previously, some of the consonants in the dataset exclusively follow either short vowels or long vowels (i.e., (/g/, /s/, and /z/). To create the second dataset, we remove all vowel tokens that precede one of these consonants and test the efficacy of normalization on this partially balanced dataset. The remaining consonants (/b/, /d/, /k/, and /p/) are still all much more likely to follow one of the vowel categories than the other. For example, /k/ is twice as likely to follow short vowels than long vowels, even though it co-occurs with both. Therefore, to create the third dataset, we randomly remove enough tokens such that each following consonant is preceded by the same number of short and long vowels (Fully Balanced). The Fully Balanced dataset most resembles typical controlled lab speech corpora, as

it completely controls for which vowels occur with which consonants. For each of these three datasets, we test the efficacy of normalizing out the effect of the following consonant, by seeing whether normalized or unnormalized cues result in a better separation between short and long vowels. To ensure that differences in normalization efficacy between datasets are not due to changes in overall proportions of short/long vowels, or due to differences in dataset size, we create two additional control datasets: a Control for the Partially Balanced Data and a Control for the Fully Balanced Data. To create these datasets, we randomly remove the same number of short vowels and long vowels from the full dataset as are removed in the Partially Balanced and Fully Balanced datasets, but remove them randomly and uniformly from all contexts, instead of removing them based on the following consonant. We run normalization using linear regression on each of these five datasets, and test whether normalization is helpful on each of them.

Results

The results are summarized in Table 6. Normalizing for the effect of the following consonant is ineffective on the Full dataset: unnormalized cues result in 91.4% overall accuracy, while normalized cues result in 82.7% overall accuracy. However, normalization was more effective on the Partially Balanced dataset, which removed all vowel tokens that preceded a consonant that only occurred either with long vowels or with short vowels. Unnormalized cues result in 90.1% accuracy, while normalized cues result in 92.6% accuracy. Finally, normalization was even more effective on the Fully Balanced dataset: unnormalized cues resulted in 90.1% accuracy, while normalized cues brought the accuracy up to 93.8% accuracy. That is, each step of removing imbalances in the data resulted in improvements in normalization performance. In fact, when we completely balanced the dataset, normalization was effective. Just reducing the size of the tested dataset or changing the relative proportion of short/long vowels was not enough to explain this effect, as normalization was still ineffective on both control datasets.

Discussion

In this section, we explored why normalization is unhelpful on spontaneous speech. One difference between spontaneous speech and lab speech is that sound categories in spontaneous speech often differ in the contexts they are likely to occur in, while in lab speech, researchers specifically control where sounds occur to make sure that the dataset is fully balanced. We took advantage of one contextual factor within the Werker data for which this was not true, and found that when there were imbalances in a particular context, normalization hurt, but when we artificially balanced the context, normalization was helpful. That is, listeners relying on a normalization strategy when their input contains strong imbalances between categories would be hurt, unless they could somehow learn that they should not normalize for factors that are imbalanced. In this particular case, that would mean learning to normalize for the previous consonant, but not the following consonant.

If category imbalances across contexts were the only factor impeding normalization in our analyses, then we would expect a similar manipulation to make normalization effective on the R-JMICC data. However, in further analyses (not described in detail here), we were unable to show that balancing contextual factors on the spontaneous R-JMICC data made normalization effective. This suggests that although contextual imbalances of this type constitute one key difference between lab speech and spontaneous speech, they are not the only reason that normalization is ineffective on spontaneous speech but not lab speech. Another possibility is that duration is less of a primary cue

Table 6 Results from balancing how often short/long vowels precede different sounds in the Werker Read Speech corpus. Results indicate that the more balanced the corpus, the better normalization performs

Data	Model	Accuracy	Short accuracy	Long accuracy	BIC
Full	Unnormalized	91.4	89.7	92.9	246
	Normalized	82.7	79.5	85.7	525
Partially balanced	Unnormalized	90.1	87.2	92.9	219
	Normalized	92.6	92.3	92.9	241
Fully balanced	Unnormalized	90.1	87.2	92.9	172
	Normalized	93.8	92.3	95.2	112
Control for partially balanced data	Unnormalized	91.4	89.7	92.9	181
	Normalized	81.5	80.8	82.1	377
Control for fully balanced data	Unnormalized	90.1	89.7	90.5	151
	Unnormalized	83.5	79.5	87.1	325

to vowel length in spontaneous speech than lab speech, and this could make normalization ineffective.

That being said, this simulation points to an interesting interaction between normalization and top-down information accounts, because the imbalances that are harmful for normalization are precisely the imbalances that are helpful for top-down information accounts. That is, when there is signal in the input that is helpful for top-down information accounts, normalization suffers. In the following section, we delve into this interaction in more detail.

A mathematical analysis of how properties of naturalistic input affect the efficacy of normalization and top-down information approaches

We have seen that there are two ways that context affects sound production: it affects how likely a particular sound category is to be produced a priori, and, once that is decided, it affects what acoustic realization that sound category is likely to have. As a result, there are also two main ways that listeners might make use of contextual information when processing or learning the sounds of their language. They could either make use of it to normalize the acoustics, or they could make use of it as top-down information that biases their category perception directly. Thus far, we have shown that in the case of Japanese vowel length, top-down information accounts are robust even on naturalistic speech, but that normalization is not effective on naturalistic speech.

The previous simulation suggests that signal in the input that is helpful for top-down information accounts may be harmful for normalization accounts. In this section, we provide a theoretical analysis about how listeners relying on each of these two strategies will fare depending on the kinds of information sources that are present in their input, including what pitfalls they might encounter. We ultimately show that a listener relying on a normalization strategy when their input contains imbalances in categories across contexts may be misled, consistent with our previous simulation, while a listener who relies on a top-down information strategy when their input contains systematic variability resulting from context will not be. Overall, the results in this section suggest that top-down information strategies are much more robust to various types of input than normalization strategies are.

How do contextual category imbalances affect normalization performance?

In this section, we consider how a listener relying on a normalization strategy will fare when their input contains imbalances in category membership—of the type that are helpful in top-down information accounts.

We begin by recapping what inference task we assume the listener is performing, and how exactly we implement normalization. As discussed in Chapter 2, we use a logistic regression categorization model, which involves calculating the relative probability that a particular vowel is long (or short) as follows, where d refers to duration, f_1, f_2, and f_3 refer to formants and all β's refer to learned weights in the logistic regression.

$$P(\text{long}|d^{norm}, f_1^{norm}, f_2^{norm}, f_3^{norm}) = \frac{1}{1 + e^{\beta_0+\beta_1 d^{norm}+\beta_2 f_1^{norm}+\beta_3 f_2^{norm}+\beta_4 f_3^{norm}}} \quad (7)$$

That is, this relies on having normalized duration and formants to categorize a particular vowel as phonemically short or long. There are a number of ways that normalization can be implemented. Our upcoming analyses focus exclusively on the linear regression implementation, which is a simple, but commonly used normalization method.

In order for normalization to be helpful, we would expect normalization to push the means of the short and long vowel categories apart. To study when normalization is or is not helpful, we derive an equation that quantifies how the distance between category means changes as a result of normalization. The mean of short vowels before normalization, $\mu^{\text{unnorm}}_{l=\text{short}}$, is the average of the mean duration of short vowels in each context short vowels occur in, weighted by how many of all the short vowels occur in that context. In the following equation, $N_{l=\text{short},c=j}$ is the number of short (l = short) vowels in context j ($c = j$), $N_{l=\text{short}}$ is the total number of short vowels, and $\mu_{l=\text{short},c=j}$ is the mean duration of short (l = short) vowels in context j ($c = j$).

$$\mu^{\text{unnorm}}_{l=\text{short}} = \sum_j \frac{N_{l=\text{short},c=j}}{n_{l=\text{short}}} \mu^{\text{unnorm}}_{l=\text{short},c=j} \quad (8)$$

An analogous equation holds for the mean of long vowels before normalization, $\mu^{\text{unnorm}}_{l=\text{long}}$. We can then compute a closed form value for the means of the short and long vowel categories after normalization with linear regression—$\mu^{\text{norm}}_{l=\text{short}}$ and $\mu^{\text{norm}}_{l=\text{long}}$, respectively. Each vowel token is normalized by taking the difference between that vowel's acoustic cue and the average acoustic cue of vowels that occur in that vowel's context. Once we obtain closed form values for the mean acoustics of short and long vowels pre- and post-normalization, we can derive the following equation, which shows how the difference between short and long vowel means changes as a result of normalization. This allows us to describe under what conditions category means will move closer together or farther apart as a result of normalization. Of course, the success of categorization

depends not just on the difference in means, but how large this difference is compared to the variance. But in the simplest case, where normalization applies an additive mean shift without changing the variance, it is clear that normalization will hurt performance when the means become closer together. See the Appendix for a full derivation of this equation.

$$\begin{aligned}&\left(\mu^{norm}_{l=\text{long}} - \mu^{norm}_{l=\text{short}}\right) - \left(\mu^{unnorm}_{l=\text{long}} - \mu^{unnorm}_{l=\text{short}}\right)\\ &= \sum_j \left[\frac{N_{l=\text{short},c=j}}{N_{l=\text{short}}} - \frac{N_{l=\text{long},c=j}}{N_{l=\text{long}}}\right]\left[\frac{N_{l=\text{long},c=j}}{N_{c=j}}\mu^{\text{unnorm}}_{l=\text{long},c=j}\right.\\ &\left. + \frac{N_{l=\text{short},c=j}}{N_{c=j}}\mu^{\text{unnorm}}_{l=\text{short},c=j}\right] \qquad (9)\end{aligned}$$

In this equation, $N_{l,c}$ is the number of vowels of length l in context c and $\mu_{l,c}$ is the mean of vowels of length l in context c. The first term in the sum, $\frac{N_{l=\text{short},c=j}}{N_{l=\text{short}}} - \frac{N_{l=\text{long},c=j}}{N_{l=\text{long}}}$, corresponds to the difference between the fraction of all short vowels that occur in the j^{th} context and the fraction of all long vowels that occur in the j^{th} context. The second term in the sum, $\frac{N_{l=\text{long},c=j}}{N_{c=j}}\mu^{\text{unnorm}}_{l=\text{long},c=j} + \frac{N_{l=\text{short},c=j}}{N_{c=j}}\mu^{\text{unnorm}}_{l=\text{short},c=j}$, is a weighted average between the mean of the long vowels in this context weighted by the proportion of vowels in this context that are long and the mean of the short vowels in this context weighted by the proportion of vowels in this context that are short. The product of these two terms is summed over all contexts. When the value of $\left(\mu^{norm}_{l=\text{long}} - \mu^{norm}_{l=\text{short}}\right) - \left(\mu^{unnorm}_{l=\text{long}} - \mu^{unnorm}_{l=\text{short}}\right)$ from Eq. 9 is greater than zero, this means that normalization has pushed the categories apart and when this value is less than zero, this means that normalization has pushed the categories closer together. This equation reveals that if there are imbalances in the contexts in which different categories are likely to occur in, then a listener relying on normalization alone may be misled. To illustrate why, consider a context that is dominated by long vowels (i.e., there are more long vowels than short vowels in this context). For such a contextual factor, we would typically expect the first bracketed term (of two) in Eq. 9 to be negative. This is because it is likely that the proportion of all long vowels that are in this context, $\frac{N_{l=\text{long},c=j}}{N_{l=\text{long}}}$ is greater than the proportion of all short vowels that are in this context, $\frac{N_{l=\text{short},c=j}}{N_{l=\text{short}}}$ (although this need not be the case if, for example, there are many more long vowels than short vowels overall: $N_{l=\text{long}} > N_{l=\text{short}}$). In this long-dominated context, the second bracketed term (of two) would be relatively large for the following reason. Most of the vowels in this context are long (by virtue of it being a long-dominated context), so $\frac{N_{l=\text{long},c=j}}{N_{c=j}}$ will be relatively large and $\frac{N_{l=\text{short},c=j}}{N_{c=j}}$ will be relatively small. The second bracketed term (of two) then consists of putting a higher weight on the long vowel mean than on the short vowel mean, which will push this value more towards the long vowel mean (and thus higher). Taking the product, the value within the sum will be a relatively large negative number for long-dominated contexts. Conversely, in a context that is dominated by short vowels (i.e., there are more short vowels than long vowels in this context), we would typically expect the first term to be positive, and the second term to be relatively small, due to a heavier weighting on the short vowel mean than on the long vowel mean (which will push the value towards shorter durations). Taking the product, the value within the sum will be a relatively small positive number for short-dominated contexts. Overall, then, we would expect the sum over all contexts to be negative, since, as we saw, the negative summands should be relatively large, and the positive summands should be relatively small. This means that imbalances in sound categories across contexts (i.e., large differences in the relative proportion of short and long vowels within particular contexts) can lead to normalization bringing the category means closer together, rather than farther apart.

Another way to think about this is that vowels are normalized relative to the context they occur in, by subtracting the mean of all of the vowels in that vowel's context from that vowel's own acoustic values. An imbalance between short and long vowels in a particular context will cause the mean of that context to be artificially decreased or increased, respectively. All else being equal, in a context that consists of a majority of long vowels, the mean duration will be artificially lengthened, so the normalized cues will be artificially low. A parallel effect will cause the normalized cues for short-dominated contexts to be artificially higher than expected. That is, vowels in contexts that are majority long will be shifted towards shorter durations, and vowels in contexts that are majority short will be shifted towards longer durations, which will push the short and long vowel distributions together. Essentially, the problem is that imbalances in where categories occur make it hard to estimate a proper normalization function.

Consider again the toy example in Fig. 2. In this toy example, there are short vowels and long vowels and only two contexts. Note that the acoustics of the short and long vowels do not change across contexts – the average short vowel and long vowel durations are not shifted. However, there is a large imbalance between phonemically short and phonemically long vowels in particular contexts, such that there are many more long vowels than short vowels in phrase-medial position, and many more short vowels that long vowels in phrase-final position. This will cause the mean duration in phrase-medial context to be much higher

than the mean duration in phrase-final context. A listener relying on normalization would try to normalize and would actually increase the amount of within-category variability present in the speech stream and push the categories together. Overall, differences between sound categories, in terms of the contexts they are likely to occur in, can impede a listener who relies on normalization strategies.

To be clear, this analysis only reveals that normalization will be problematic for factors that can also be useful for top-down information accounts—e.g., neighboring sounds, prosodic position, but not for speech rate or speaker, which only affect how a sound is produced and not which sound is likely to be produced. That being said, normalization will be ineffective whenever it is difficult to estimate a normalization function, and there is reason to believe that this might be the case for factors like speech rate too. In particular, it has been shown that factors like speech rate seem to acoustically lengthen long vowels more than short vowels. In its current form, then, normalization will be incorrect for factors like speech rate, because it only estimates one normalization function for both short vowels and long vowels, instead of using a different function for each category. We return to this point and what it tells us about the efficacy of normalization in the General Discussion.

How does systematic acoustic variability affect top-down information performance?

While a listener relying on normalization when there is signal in the input for top-down information accounts will be misled, the opposite does not hold. A listener relying on a top-down information strategy when there is systematic variability to be normalized in their input will not be misled relative to a listener who simply relies on the acoustics. The model making use of contextual information has access to all the information that the baseline absolute acoustics model does (and more), so it will necessarily perform at least as well. It can always learn to put no weight on contextual factors, and implement exactly the baseline model. Therefore, no matter what the acoustics are like, using top-down information will never mislead a learner more than a listener only relying on acoustic information. More strongly, a listener only relying on contextual information without any access to acoustics cannot be misled by systematic variability in the signal, precisely because it makes no use of acoustic information. Therefore, a listener who makes use of top-down information as a contextual bias in learning and perception will avoid pitfalls that a listener making use of normalization may encounter (as long as they trained on the right distributions).

Discussion

In this section, we provided a mathematical analysis showing that listeners who make use of a normalization strategy may suffer when there are imbalances in category membership of the type that are useful for top-down information accounts. However, the opposite is not true—listeners that make use of top-down information accounts will not be hurt by systematic variability in the signal.

Category imbalances are extremely common in natural language, due to phonotactic constraints, phonological alternations, historical reasons, and more. Our mathematical analysis shows that listeners who rely on normalization strategies may suffer when their input contains these types of category imbalances. Therefore, for factors that affect which sound category is likely to be produced, normalization is not an effective way to deal with context in processing and especially sound category learning for learners who cannot yet separate categories. Instead, a listener would be much better off making use of top-down information, which is immune to systematic variability in acoustics. Overall, in order to make a claim that listeners do use normalization in order to learn and process sounds, it will become important to explain how listeners can overcome the problems presented by contextual imbalances in category membership.

General discussion

In this paper, we tested the efficacy of two ways of using context in helping to disambiguate overlapping categories. We tested top-down information accounts, where listeners make use of context to bias their expectations of what category they will observe, as well as normalization accounts, where listeners use context to help factor out systematic variability. Although well studied, these ideas have been somewhat conflated in past work and have rarely and with limited success been applied to naturalistic spontaneous speech. In this paper, we further explored these two ideas, trying to overcome these issues with past work. We disentangled these two ideas and carefully studied the relative contribution of each of them to the listener's task, applying them to spontaneous speech.

Our simulations showed that a top-down information strategy is effective even on spontaneous speech, but that normalization, at least as it has often been implemented, is not. This result was surprising given that normalization has been found to be effective in the past. We resolved this discrepancy, by showing that normalization was helpful when we ran our same analyses on simplified controlled lab data—of the type generally studied in the normalization

literature—rather than on naturalistic spontaneous speech. We then provided simulations and a mathematical analysis showing that normalization may be ineffective when there are context-specific category imbalances—precisely of the type that are useful for top-down information accounts. This suggests that a learner whose input contains imbalances through phenomena such as phonotactic constraints and phonological alternations, is better off using context to bias their perception in a top-down fashion rather than normalizing it out, at least as we have implemented these strategies here. In what follows, we discuss where this leaves normalization in the literature, how top-down information may be used in acquisition models, how these results generalize to other problems in speech perception and cognition, as well as the importance of testing ideas on spontaneous data in addition to controlled lab data.

The status of normalization

The idea that normalization plays a role in processing and acquisition is a widely held idea, but our results bring up important issues with it that complement other problems discussed in earlier work (Johnson, 1997, 2006; Pierrehumbert, 2002). Although our results can only directly speak to the two implementations that we studied in this paper, the problem appears to arise because of properties of the input that might also hurt other normalization implementations, as we discuss below.

In order to normalize well, it is important to be able to estimate the correct normalization function. Failing to do so can actually increase the amount of variability and overlap between categories, rather than reduce it, as we saw in the toy example in Fig. 2. The results in this paper show that one obstacle to estimating the normalization function well is the fact that different sound categories occur in different contexts with different probabilities. For example, short and long vowels differ in what consonants they are likely to precede, which makes factoring out systematic acoustic variability from the following consonant difficult. This is a problem for any contextual factor that both affects which target sound is likely to occur a priori, and systematically affects a target sound's acoustics (i.e., any contextual factor that would be a good top-down predictor of vowel category). Therefore, this is not a problem for sources of systematic variability like speech rate or gender: short (or long) vowels are no more likely to occur in fast speech than slow speech, or in speech by men than women (and vice versa), so speech rate and speaker gender are unlikely to be informative about whether a particular vowel is short or long.

That being said, there are other problems with normalization that would also affect factors like speech rate and gender. Context can affect how some categories are produced more than others. For example, long vowels might be acoustically lengthened more than short vowels in slow speech. In the implementation adopted here, normalization cannot handle these types of sources of systematic variability because it uses one normalization function across all categories. If different categories are actually differentially affected, then the learned normalization function is guaranteed to be wrong for some tokens, and this may increase variability and overlap rather than reduce it.

Our results directly study linear regression normalization methods; however, our analyses reveal that the properties of naturalistic speech that hurt the studied normalization method would likely also hurt many of the most commonly discussed normalization methods. Essentially, the linear regression normalization method fails because it has an underlying assumption that the only way the mean value of an acoustic cue could change between contexts is if speakers are acoustically altering their productions between those contexts. However, this assumption is not valid in naturalistic speech, where simply having more instances of a particular category in a context can change the mean value of an acoustic cue. Most, if not all, concrete implementations of normalization that have been proposed in the cognitive literature have this same problem. For example, z-scoring, and the normalization method in Dillon et al. (2013) both rely on transforming the acoustics relative to the overall mean acoustic cue in each context. Therefore, while we only directly study one implementation, other implementations could suffer from the same problem.

One possible exception is the idea of relativizing cues. The idea is that, instead of normalizing by learning an explicit normalization function (implemented here as a linear regression or neural network), listeners might rely on an alternative set of acoustic cues that are more invariant than those that are typically measured and described. For example, for Japanese vowel length, listeners might use the ratio of a vowel's duration to the word duration or the duration of a neighboring sound as the primary cue (Hirata, 2004). For other contrasts, researchers have argued that cues like the ratio of first and second formants to third formant values, as well as ratios between stop closure and previous vowel duration could be helpful for perception of vowel qualities and stops, respectively (e.g., Monahan & Idsardi, 2010; Port & Dalby, 1982).

These accounts have support in the literature: MEG experiments have shown that the auditory cortex is sensitive to ratios of these sorts (Monahan & Idsardi, 2010), and analyses have shown that these can be clear cues to category membership (Hirata, 2004; Monahan & Idsardi, 2010), though these analyses have been of controlled lab speech rather than naturalistic speech.

In our own preliminary work and in work by Bion et al. (2013), these relativized cues do not help in the Japanese vowel length contrast, when considering naturalistic speech.

However, it is possible that other untested relativized acoustic cues would help, and future work should be done to systematically study this class of ideas as applied to naturalistic speech.

Given that so many of the already proposed normalization implementations are likely to suffer from the problem exposed in this paper, the question that remains is whether current normalization methods could be altered, or if new normalization methods could be developed, to overcome these issues. The problem we point out is that it is difficult to estimate the correct normalization function. However, it is possible that normalization would be effective with a better estimation of the function. Here, we discuss possible changes that could accomplish that.

With regards to adult speech perception, one possibility would be for listeners to learn different normalization functions for each category type (i.e., one normalization function for short vowels and another for long vowels). This would also allow the process to take into account category imbalances. However, if the listener is equipped with one normalization function for short vowels and one normalization function for long vowels, they will not know which function to use until they have already categorized the sound, so normalization would not, in this case, be helpful during the categorization process (only after). Another possibility would be that listeners build separate normalization functions for separate categories, but average them during categorization, weighting them by the relative proportion of each category type. These ideas have not yet been tested, and it is currently unclear that they would increase the efficacy of normalization on spontaneous speech, but future work should investigate them.

The issues with normalization become even more problematic when considering acquisition. The learner does not yet know the distinction between short and long vowels, and cannot take into account category imbalances. As a result, they will necessarily be applying inaccurate normalization functions, which may actually increase category overlap rather than reduce it. In addition, throughout the paper, we saw that normalization performance depended on the precise set of factors being normalized out. Therefore, a learner would have determine which factors to factor out—and would need to learn that some factors that systematically affect acoustic productions should be factored out, but other factors that similarly affect acoustic productions should not be factored out. These issues complicate the view that normalization is helpful in language acquisition.

Overall, although normalization has received a lot of support in the literature, there is actually little to no current evidence suggesting that this is a strategy that could be helpful for acquisition and processing naturalistic speech. Much of the evidence that has been used to argue for normalization is also consistent with a top-down information strategy, which, unlike normalization, was shown to be effective here, as well as adaptation accounts (Kleinschmidt & Jaeger, 2015). In addition, normalization has mostly been tested on controlled lab speech, rather than the speech that listeners primarily hear and learn from. We showed here that these results from lab speech do not necessarily generalize to naturalistic speech (and did not in the case of Japanese vowel length). This work calls into question the role that normalization could play in acquisition and processing. Certainly it is possible that amending the normalization process helps, but for existing concrete proposals, there is more evidence against normalization than for it. In order to stand by the idea that normalization helps disambiguate overlapping categories, it is critical to find some evidence that normalization—in any form—is actually effective in separating categories when applied to spontaneously produced speech.

That being said, the fact that we show that normalization may not lead to better separation between short and long vowels does not imply that listeners do not normalize. If it is the case that listeners process their input by normalizing acoustics relative to context, then our results indicate that listeners are overcoming even more overlap between short and long vowels than represented in Fig. 1. Our results show that normalization is unlikely to be the solution to the overlapping categories problem.

As discussed before, normalizing is only one way to factor out systematic variability resulting from the context, and other alternatives may be more effective. One particularly promising alternative would be an adaptation strategy, which reduces systematic variability without having to calculate an explicit normalization function. It does so by essentially learning a separate mapping between acoustics and linguistic category for each context observed. This avoids the need to learn a precise normalization function, but can still overcome systematic variability by treating each context separately (see Kleinschmidt & Jaeger, 2015 for a more extensive discussion). Adaptation is promising to pursue because it does not encounter any of the issues that normalization does, and can explain the experimental findings that have been used to argue for normalization. In particular, an adaptation mechanism would also be able to cope with variability due to speech rate or gender, so studies of these factors do not provide independent evidence that normalization must be present. At the same time, the adaptation account is independently supported by experimental evidence regarding other kinds of variability between speakers, such as dialect. Even young infants are capable of learning to adapt to dialect variation given sufficient evidence (van Heugten & Johnson, 2014). Normalization is not a good account of this learning process, since it would require complex phonological alternations such as vowel shifts to be normalized directly in the acoustic

space rather than learned phonologically (Elsner et al., 2013a).

Currently, we run into data sparsity issues when trying to investigate the adaptation idea, as this requires splitting already small datasets by context; however, it is promising to continue pursuing.

Top-down information in acquisition

Our results indicated that using contextual information in a top-down fashion is promising and merits continued study. In the current work, we only included a small set of contextual factors, and adding in additional factors could help separate short and long vowels even further. In particular, with the exception of part-of-speech, our work did not include any word type or word-level information, which has been argued to be helpful in the past (Swingley, 2009; Feldman et al., 2013a, b). This suggests that using contextual information as top-down information or to guide expectations could be extremely helpful in adult speech perception.

All of the models presented here are supervised, meaning that they have information about what the sound categories and relevant contextual factors are. They directly reveal how helpful a top-down information strategy would be for adult speech perception. When applied to acquisition, our results have shown that the prerequisite for a top-down information strategy to be effective in acquisition is met: there is signal in the input that can separate short vowels from long vowels. However, we have not shown how an infant could actually use this contextual information to acquire the vowel length contrast. In order to do this, we would need to propose an unsupervised category learning model. This is the primary challenge facing top-down information accounts, and future work will need to apply these ideas more directly to acquisition, as has been done in the past, to try to gain a more complete account of how the listener solves these overlapping category problems.

A lot of work has already applied top-down information strategies to acquisition. Past research has shown that infants seem to make use of word-level information in acquiring sound categories (Feldman et al., 2013b; Thiessen, 2007) and computational models has shown that word-level information can be helpful (Feldman et al., 2013a). However, as with most speech perception research, these ideas have largely been tested on controlled lab speech, and, in fact, recent work showed that the model from Feldman et al. (2013a) was no longer effective when applied to spontaneous Japanese speech (Antetomaso et al., 2017). As a result, we still do not have a solution for how contextual information could be used in a top-down fashion for learning from spontaneously produced speech. However, most past work has focused exclusively on word-level information, so it is possible that making use of the other aspects of context that we considered in this paper (e.g., prosodic position, neighboring sounds), in addition to word-level information, will result in models that work on naturalistic speech. In what follows, we outline a few possibilities for how top-down information could be useful.

An adaptation strategy, which builds a separate mapping from acoustics to categories for each context encountered, could again be helpful (Kleinschmidt & Jaeger, 2015). In doing so, it has access to information about which categories are more/less likely to occur within a particular context. Therefore, it is possible that within particular contexts, the short and long vowel categories are more separated than they appear overall. For example, short vowels and long vowels might be well separated when they occur in phrase-final position and preceded by a particular consonant. If the distribution is bimodal along the duration dimension in a particular context like this one, then the learner could learn that there are two categories along the duration dimension via a process of distributional learning (Maye et al., 2002), and then generalize this to other contexts where the distinction is less clear.

In particular, our results have shown that there seems to be something about carefully enunciated Japanese speech that reliably elicits different durations for short versus long vowels. While most of the input infants hear is highly-variable and spontaneous, infants do sometimes get some exposure to clearer instances of speech. Infants may hear their parents read them books. In addition, parents often use repetitions when speaking, which could help children by providing them points of comparison. Finally, Japanese child-directed speech involves a lot of mimetics, or onomatopoeic words, which have been shown to help in lexical acquisition (e.g., Imai & Kita, 2014), and might differ from other subsets of IDS in terms of the relative proportion of short and long vowels, or in terms of how well the length distinction is enunciated. Therefore, it is, in principle, possible that children learn which words have which vowels precisely by listening to subsets of their data in which their parents speak carefully.

This type of adaptation strategy, in which children learn about the distinction in a particular context and generalize it, is a particularly promising strategy, as it provides both a way to take advantage of top-down expectations of category membership, as well as a way to remove systematic variability. By building a separate mapping from signal to category for each context, it has access to information about top-down information about which categories are more or less likely to occur in a particular context, and, therefore, can account for results that have been used to argue for top-down information accounts. At

the same time, experimental results that have been used to argue for normalization functions can also be explained by adaptation strategies (Kleinschmidt & Jaeger, 2015), as these results show that listeners account for the fact that sounds are produced differently in different contexts, but cannot dissociate whether listeners do so via an explicit normalization function, or by building a separate model for each context. As a result, adaptation accounts can explain the experimental findings that have been used to argue in favor of both top-down information accounts and normalization accounts in a unified way. However, again, infant-directed speech corpora are generally quite small, and it is difficult to test adaptation accounts without running into data sparsity issues.

The above possibility relies on the learner considering the distribution of vowels across many different contexts, and observing a bimodal (or substantially less overlapping) distribution within one of those contexts. Another possibility is that the overall distribution remains unimodal across contexts, but that the shape of the distribution, nonetheless, changes, and this could reveal the presence of multiple categories. What our results have revealed is that there are radically different proportions of short and long vowels across different contexts. As a result, the overall distribution of vowels (along the duration dimension) will change depending on the relative proportion of short and long vowels within it. For example, a context in which almost all vowels are long will be more heavily skewed to the right than one in which all of the vowels are short. It is possible that these types of distribution shape changes only occur along acoustic dimensions that are contrastive for a language (e.g., for duration in Japanese, but not for French which does not use duration contrastively). If this is true, then a learner might be able to detect that a particular dimension is contrastive in their language by observing these types of changes in distribution shape across contexts. A learner could apply this across contexts, or alternatively, within frequent words—to see whether the vowel distributions within particular word frames differ (i.e., depending on what proportion of the time the word frames involve short vowels or long vowels).

There are other ways that a learner could take advantage of contextual information in a top-down fashion, and future work will implement these strategies computationally and test how effective they are at learning the contrast between short vowels and long vowels.

How generalizable are these results?

We focus on the Japanese vowel length contrast as a test case, but to what extent do our results generalize to other similar overlapping category problems in speech perception and more generally? For a number of reasons, the vowel length contrast is unique and it is possible that these properties explain the results we observe. For example, vowel length is acquired relatively late (Sato et al., 2010; Mugitani et al., 2009), and it could be that earlier learned contrasts rely more on normalization.

First, it is relatively likely that our results showing the efficacy of top-down information accounts would generalize to other tasks in speech perception, phonetic learning, and cognition. In all of these areas, there is already ample evidence that top-down information is useful (though mostly from controlled lab data). Our results suggest that this usefulness will generalize to more realistic data, because there are systematic regularities in which contexts sounds (or objects) of all types occur in; however, this hypothesized generality will need to be demonstrated in future work.

The picture is a bit more complicated for our results on normalization. Our analysis reveals that normalization is ineffective when it is difficult to estimate the normalization function. As we have shown here, it will be difficult to estimate normalization functions for contextual factors that would helpful for top-down information accounts (i.e., when there are regularities in which categories occur in which contexts). In addition, as we have discussed in the previous section, it will also be difficult to estimate one good normalization function for contextual factors that affect the productions of different categories differently (e.g., a contextual factor that acoustically lengthens long vowels more than they acoustically lengthen short vowels). To the extent that people are dealing with contextual factors that do not fall into one of these classes, normalization could very well help for the tasks of speech perception and phonetic learning. In particular, it is possible that in the case of Japanese vowel length, there is sufficient signal via top-down information to distinguish most short/long minimal pairs without attending to the acoustic duration at all, so that in conversational speech, the durational contrast is mostly neutralized. It may be the case that normalization is ineffective for contrasts with low functional load (like Japanese vowel length), but more effective for contrasts with high functional load, where speakers must produce a perceptible contrast in order to be understood. We, nonetheless, speculate that the ineffectiveness of normalization will generalize to many other contrasts, as naturalistic speech is full of top-down information, which helps predict which sound will be uttered, even without hearing the acoustics of the sound, but hurts normalization. However, further work will need to be done to study the extent to which these findings generalize to other contrasts within the domains of speech perception and phonetic learning.

Dissociating top-down information from normalization accounts

The two ideas we study here—normalization and top-down information accounts—have often been conflated in the literature. Part of the reason why this might be the case is that they are difficult to dissociate experimentally. Many of the studies that have been used to argue for one or the other are actually compatible with both alternatives, because they show that contextual information is used, but cannot pinpoint exactly how. In addition, these two ideas have largely been treated in separate literatures, such that where computational models have been proposed, they have never been directly compared.

In this work, by implementing these two ideas separately, and testing their relative efficacy on the same task, we are able to dissociate these two ideas. On the same task, top-down information accounts performed well, but normalization accounts performed poorly, showing that these two accounts are theoretically very different, even if it has been difficult to separate them empirically.

It is very well known that contextual information is used in speech perception, but, as this paper highlights, there are many ways that contextual information is used, and it will be important to get a better understanding of how exactly listeners use context. Towards this end, future work should devise ways to test how listeners use context in speech perception and acquisition, in a way that can differentiate between different accounts.

There has been some work that has succeeded in dissociating these two accounts. Much of this work comes from testing contextual factors that could be helpful under one account, but not for the other, and showing that listeners use them. For example, the Ganong effect, in which participants preferentially categorize sounds so as to form words (Ganong, 1980), and the phonemic restoration effect, in which participants report hearing a sound that is not physically present in the speech (Warren, 1970), show that top-down information is used, and are incompatible with normalization accounts. On the other hand, experiments showing that speakers change their perception based on speech rate or speaker (e.g., Nearey, 1978; Fujisaki et al., 1975), are incompatible with top-down information accounts, because e.g., a speaker is unlikely to produce more /s/ phones just because they are speaking quickly. These studies show us that listeners are factoring out systematic variability in one way or another, though it is also unclear how to dissociate using both normalization and top-down information from an adaptation account. Finally, there have also been experiments which have directly compared two different ways of using context (Toscano & McMurray, 2012). In so doing, the researchers showed that an effect that is typically taken as evidence for normalization can also be explained by other ways of using context. Studies that can put these two theories in conflict can be particularly helpful, although because the strategies are not mutually exclusive, it is possible for listeners to use both.

Future work should build off of these cases to help us gain a more nuanced view of how listeners rely on context in speech perception.

Controlled lab speech vs. naturalistic speech

The results of this paper reiterate once again that there is a crucial distinction between controlled laboratory speech and spontaneously produced naturalistic speech. Essentially all of our understanding of speech perception comes from work on carefully controlled and carefully enunciated laboratory speech, but almost all of our experience as listeners comes from messy, variable spontaneous speech. These two types of speech differ quite substantially from each other in nature, both in how the speech is produced, as well as the content of the speech. Indeed, where tested, many of the ideas developed on controlled lab speech have been shown to be ineffective on spontaneous speech. Previous work has shown that top-down information accounts developed and tested on carefully controlled or synthesized speech do not generalize to spontaneously produced lab speech (Antetomaso et al., 2017). The current work shows that normalization is helpful on lab speech, but ineffective on spontaneously produced speech.

There is obviously a great deal of value that comes from working on speech where various factors are controlled for and isolated. In addition, listeners can process synthesized and controlled lab speech effortlessly, so our theories must be able to handle clear, enunciated speech, in addition to more naturalistic daily speech. However, what is critical is for ideas generated and tested on this lab speech to then be applied to spontaneous speech, to make sure that researchers are working on the same problem that listeners are solving.

There has certainly been some research starting to look at spontaneous speech, especially with the development of hand-annotated child-directed speech corpora such as from (Mazuka et al., 2006). As we have discussed, Antetomaso et al. (2017) applied the model from Feldman et al. (2013a) to spontaneous Japanese speech, showing that the model's success did not generalize to spontaneous speech. Other work has also investigated spontaneous speech corpora both in the case of the overlapping categories problem (Narayan et al., 2017; Swingley & Alarcon, 2018) and more widely (Guevara-Rukoz et al., 2018; Ludusan et al., 2016; Ludusan et al., 2017; Martin et al., 2016). However, it is still not prevalent, and our work aligns with previous work in revealing that studying spontaneous speech is critical for ensuring our ideas apply to naturalistic listening situations.

Conclusions

In this paper, we compared the relative efficacy of two ways of using context to help in phonetic learning. The first involved making use of contextual information as top-down information to guide expectations about what category was likely to be heard. The second involved factoring out systematic acoustic variability that resulted from the context a sound was produced in. These ideas have been conflated and almost entirely studied on controlled lab speech, not naturalistic speech. In this work, we showed that, for the case of the Japanese vowel length distinction, a top-down information strategy is effective even on spontaneous speech, but, contrary to previous findings, normalization is not. We resolved this discrepancy in findings, by demonstrating that the same normalization procedure is helpful on lab speech—the focus of most previous studies—but ineffective on spontaneous speech—the focus of our study. We then provided simulations and a mathematical analysis showing that normalization may be ineffective when there are context-specific category imbalances—precisely of the type that are useful for top-down information accounts. These results suggest the need to reevaluate the role that normalization can play in acquisition and processing. In addition, they reveal the importance of applying ideas tested on well enunciated lab speech to highly variable spontaneous speech which is present in most listening situations.

Acknowledgements This work was supported by National Science Foundation grants #IIS-1421695, #IIS-1422987, #DGE-1449815, and NSF/JSPS EAPSI grant #1713974. We thank Laurel Fais, Savannah Nijeboer, Ryoko Mugitani, and Janet Werker for providing the Werker data used in this work. We also thank Adam Albright, Stephanie Antetomaso, Robert Daland, Edward Flemming, Bill Idsardi, Chiyuki Ito, Kyoji Iwamoto, Jeff Lidz, Bob McMurray, Thomas Schatz, Kristine Yu, the RIKEN Lab for Language Development, the MIT Computational Psycholinguistics Lab, the MIT Phonology Circle, the MIT CompLang group, NECPhon 11, Brown University LingLangLunch, and the UMD ProbMod group for their help and feedback on this work.

Open Practices Statement This work made use of two previously collected corpora, and the researchers who originally collected them control their distribution. The R-JMICC corpus, previously collected by one of the authors of this paper (Reiko Mazuka), is not publicly available due to IRB requirements. The code used in this work is available on the first author's webpage.

Appendix

Notation

$N_{l,c}$ - The number of vowels of length l (phonemically short or long) in context c

$\mu_{l,c}$ - The mean duration of vowels of length l (phonemically short or long) in context c

v_i - The unnormalized duration of vowel token i

Derivation

The following derives Eq. 9, which characterizes how the means of two categories will move relative to one another as a result of normalization when top-down expectations are present. First, we write out what the unnormalized average duration of long vowels and short vowels is, beginning with long vowels. The average duration of long vowels is simply the sum of every long vowel's duration, divided by the total number of long vowels:

$$\mu^{\text{unnorm}}_{l=\text{long}} = \frac{1}{N_{l=\text{long}}} \sum_{i:\, l_i=\text{long}} v_i \tag{10}$$

Summing every long vowel's duration is equivalent to summing every long vowel's duration in every context, and then adding up the sums from each context, which can be written as

$$\mu^{\text{unnorm}}_{l=\text{long}} = \frac{1}{N_{l=\text{long}}} \sum_{j} \sum_{i:\, l_i=\text{long}, c_i=j} v_i \tag{11}$$

We then multiply in $\frac{N_{l=\text{long},c=j}}{N_{l=\text{long},c=j}}$ to obtain:

$$\mu^{\text{unnorm}}_{l=\text{long}} = \frac{1}{N_{l=\text{long}}} \sum_{j} \frac{N_{l=\text{long},c=j}}{N_{l=\text{long},c=j}} \sum_{i:\, l_i=\text{long}, c_i=j} v_i \tag{12}$$

The value $\frac{1}{N_{l=\text{long},c=j}} \sum_{i:\, l_i=\text{long}, c_i=j} v_i$ is simply $\mu^{\text{unnorm}}_{l=\text{long},c=j}$. This is because we are summing the durations of all long vowels in context j and then dividing that by the total number of long vowels in context j, which is equivalent to the mean duration of long vowels in context j. This gives us the following equation for the unnormalized average duration of long vowels:

$$\mu^{\text{unnorm}}_{l=\text{long}} = \sum_{j} \frac{N_{l=\text{long},c=j}}{N_{l=\text{long}}} \mu^{\text{unnorm}}_{l=\text{long},c=j} \tag{13}$$

Similarly, the unnormalized average duration of short vowels is

$$\mu^{\text{unnorm}}_{l=\text{short}} = \sum_{j} \frac{N_{l=\text{short},c=j}}{N_{l=\text{short}}} \mu^{\text{unnorm}}_{l=\text{short},c=j} \tag{14}$$

Next, we compute the normalized average duration of long vowels and short vowels, starting with long vowels. To normalize a particular vowel's duration based on the context it occurs in, we take that vowel's unnormalized duration, v_i, and subtract from it the average duration of vowels in that context. The average duration of vowels in that context can be calculated by taking the sum of all short vowel durations in that context, adding that to the sum of all long vowel durations in that context, and then dividing

the total sum (of short vowel and long vowel durations) by the total number of vowels in that context. This difference, $\left(v_i - \frac{1}{N_{c=c_i}}\left(\sum_{k:\, l_k=\text{long}, c_k=c_i} v_k + \sum_{k:\, l_k=\text{short}, c_k=c_i} v_k\right)\right)$, is the normalized duration value for one particular long vowel, i. We can then take the sum of this value from each long vowel token, and divide by the total number of long vowels to obtain the average long vowel normalized duration:

$$\mu^{\text{norm}}_{l=\text{long}} = \frac{1}{N_{l=\text{long}}} \sum_{i:\, l_i=\text{long}} \left(v_i - \frac{1}{N_{c=c_i}} \left(\sum_{k:\, l_k=\text{long}, c_k=c_i} v_k + \sum_{k:\, l_k=\text{short}, c_k=c_i} v_k \right) \right) \tag{15}$$

Multiplying everything out yields

$$\mu^{\text{norm}}_{l=\text{long}} = \frac{1}{N_{l=\text{long}}} \sum_{i:\, l_i=\text{long}} v_i - \frac{1}{N_{l=\text{long}}} \sum_{i:\, l_i=\text{long}} \frac{1}{N_{c=c_i}} \sum_{k:\, l_k=\text{long}, c_k=c_i} v_k - \frac{1}{N_{l=\text{long}}} \sum_{i:\, l_i=\text{long}} \frac{1}{N_{c=c_i}} \sum_{k:\, l_k=\text{short}, c_k=c_i} v_k \tag{16}$$

The first term in Eq. 16 is summing the unnormalized durations of all long vowels and dividing by the total number of long vowels there are, so this first term is equivalent to the mean unnormalized duration of long vowels. Therefore, we can rewrite Eq. 16 as:

$$\mu^{\text{norm}}_{l=\text{long}} = \mu^{\text{unnorm}}_{l=\text{long}} - \frac{1}{N_{l=\text{long}}} \sum_{i:\, l_i=\text{long}} \frac{1}{N_{c=c_i}} \sum_{k:\, l_k=\text{long}, c_k=c_i} v_k - \frac{1}{N_{l=\text{long}}} \sum_{i:\, l_i=\text{long}} \frac{1}{N_{c=c_i}} \sum_{k:\, l_k=\text{short}, c_k=c_i} v_k \tag{17}$$

From here, we will rewrite both the second and third terms of Eq. 17, and we make an aside here to show how. Consider first the second term of Eq. 17. In it, we are summing over all long vowels. Just as in the transition from Eqs. 10 to 11, we can rewrite this as summing over all long vowels in a particular context, and then summing over these contexts, which yields (18). To get from Eqs. 18 to 19, notice that the term, $\frac{1}{N_{c=j}} \sum_{k:\, l_k=\text{long}, c_k=j} v_k$ in the inner sum will be the same for every long vowel in context j, so this term will be repeated exactly $N_{l=\text{long}, c=j}$ times.

$$\frac{1}{N_{l=\text{long}}} \sum_{i:\, l_i=\text{long}} \frac{1}{N_{c=c_i}} \sum_{k:\, l_k=\text{long}, c_k=c_i} v_k = \frac{1}{N_{l=\text{long}}} \sum_{j} \sum_{i:\, l_i=\text{long}, c_i=j} \frac{1}{N_{c=j}} \sum_{k:\, l_k=\text{long}, c_k=j} v_k \tag{18}$$

$$= \frac{1}{N_{l=\text{long}}} \sum_{j} N_{l=\text{long}, c=j} \frac{1}{N_{c=j}} \sum_{k:\, l_k=\text{long}, c_k=j} v_k \tag{19}$$

Using the same logic for the third term, we can, therefore, rewrite Eq. 17 as follows:

$$\mu^{\text{norm}}_{l=\text{long}} = \mu^{\text{unnorm}}_{l=\text{long}} - \sum_{j} \frac{1}{N_{c=j}} \frac{N_{l=\text{long}, c=j}}{N_{l=\text{long}}} \sum_{i:\, l_i=\text{long}, c_i=j} v_i - \sum_{j} \frac{1}{N_{c=j}} \frac{N_{l=\text{long}, c=j}}{N_{l=\text{long}}} \sum_{i:\, l_i=\text{short}, c_i=j} v_i \tag{20}$$

As before, the mean duration of long vowels in a particular context is equivalent to the sum over all long vowel durations in that context, divided by the total number of long vowels in that context. Writing this out notationally will help us rewrite Eq. 20:

$$\mu^{\text{unnorm}}_{l=\text{long}, c=j} = \frac{1}{N_{l=\text{long}, c=j}} \sum_{i:\, l_i=\text{long}, c_i=j} v_i \tag{21}$$

Using Eq. 21, we can rewrite Eq. 20 as

$$\mu^{\text{norm}}_{l=\text{long}} = \mu^{\text{unnorm}}_{l=\text{long}} - \sum_{j} \frac{N_{l=\text{long}, c=j}}{N_{c=j}} \frac{N_{l=\text{long}, c=j}}{N_{l=\text{long}}} \mu^{\text{unnorm}}_{l=\text{long}, c=j} - \sum_{j} \frac{N_{l=\text{long}, c=j}}{N_{c=j}} \frac{N_{l=\text{short}, c=j}}{N_{l=\text{long}}} \mu^{\text{unnorm}}_{l=\text{short}, c=j} \tag{22}$$

Factoring out $\frac{N_{l=\text{long}, c=j}}{N_{c=j} N_{l=\text{long}}}$ gives us the following equation for the mean normalized duration of long vowels,

$$\mu^{\text{norm}}_{l=\text{long}} = \mu^{\text{unnorm}}_{l=\text{long}} - \sum_{j} \left[\frac{N_{l=\text{long}, c=j}}{N_{c=j} N_{l=\text{long}}} \left(N_{l=\text{long}, c=j} \mu^{\text{unnorm}}_{l=\text{long}, c=j} + N_{l=\text{short}, c=j} \mu^{\text{unnorm}}_{l=\text{short}, c=j} \right) \right] \tag{23}$$

Similarly, the mean normalized duration of short vowels is

$$\mu^{\text{norm}}_{l=\text{short}} = \mu^{\text{unnorm}}_{l=\text{short}} - \sum_{j} \left[\frac{N_{l=\text{short}, c=j}}{N_{c=j} N_{l=\text{short}}} \left(N_{l=\text{short}, c=j} \mu^{\text{unnorm}}_{l=\text{short}, c=j} + N_{l=\text{long}, c=j} \mu^{\text{unnorm}}_{l=\text{long}, c=j} \right) \right] \tag{24}$$

Up until this point, we have calculated the mean unnormalized duration of long vowels and short vowels, as well as the mean normalized duration of long vowels and short vowels. We can subtract the average unnormalized short vowel duration from the average unnormalized long vowel duration to obtain a measure of how far apart the two vowel categories are before normalization. Similarly, we can subtract the average normalized short vowel duration from the average normalized long vowel duration to obtain a measure of how far apart the two vowel categories are after normalization. To compare whether the means of the two categories move closer together or farther apart after normalization, we can calculate the value

of $\left(\mu^{\text{norm}}_{l=\text{long}} - \mu^{\text{norm}}_{l=\text{short}}\right) - \left(\mu^{\text{unnorm}}_{l=\text{long}} - \mu^{\text{unnorm}}_{l=\text{short}}\right)$, by simply plugging in the relevant terms from above.

$$
\begin{aligned}
&\left(\mu^{\text{norm}}_{l=\text{long}} - \mu^{\text{norm}}_{l=\text{short}}\right) - \left(\mu^{\text{unnorm}}_{l=\text{long}} - \mu^{\text{unnorm}}_{l=\text{short}}\right) \\
&= \left(\mu^{\text{unnorm}}_{l=\text{long}} - \mu^{\text{unnorm}}_{l=\text{short}}\right) - \left(\mu^{\text{unnorm}}_{l=\text{long}} - \mu^{\text{unnorm}}_{l=\text{short}}\right) \qquad (25) \\
&\quad - \sum_j \frac{N_{l=\text{long},c=j}}{N_{c=j} N_{l=\text{long}}} \left(N_{l=\text{long},c=j} \mu^{\text{unnorm}}_{l=\text{long},c=j} \right. \\
&\quad \left. + N_{l=\text{short},c=j} \mu^{\text{unnorm}}_{l=\text{short},c=j}\right) \\
&\quad + \sum_j \frac{N_{l=\text{short},c=j}}{N_{c=j} N_{l=\text{short}}} \left(N_{l=\text{long},c=j} \mu^{\text{unnorm}}_{l=\text{long},c=j} \right. \\
&\quad \left. + N_{l=\text{short},c=j} \mu^{\text{unnorm}}_{l=\text{short},c=j}\right) \\
&= \sum_j \left[\left(\frac{N_{l=\text{short},c=j} \mu^{\text{unnorm}}_{l=\text{short},c=j} + N_{l=\text{long},c=j} \mu^{\text{unnorm}}_{l=\text{long},c=j}}{N_{c=j}} \right) \right. \\
&\quad \left. \times \left(\frac{N_{l=\text{short},c=j}}{N_{l=\text{short}}} - \frac{N_{l=\text{long},c=j}}{N_{l=\text{long}}} \right) \right] \qquad (26)
\end{aligned}
$$

This gives us Eq. 9 from the main text:

$$
\begin{aligned}
&\left(\mu^{norm}_{l=\text{long}} - \mu^{norm}_{l=\text{short}}\right) - \left(\mu^{unnorm}_{l=\text{long}} - \mu^{unnorm}_{l=\text{short}}\right) \\
&\quad = \sum_j \left[\frac{N_{l=\text{short},c=j}}{N_{l=\text{short}}} - \frac{N_{l=\text{long},c=j}}{N_{l=\text{long}}} \right] \left[\frac{N_{l=\text{long},c=j}}{N_{c=j}} \mu_{l=\text{long},c=j} \right. \\
&\left. + \frac{N_{l=\text{short},c=j}}{N_{c=j}} \mu_{l=\text{short},c=j} \right] \qquad (27)
\end{aligned}
$$

We can then study whether this value is positive or negative. This value will be positive when the difference between the normalized means is greater than the difference between the unnormalized means (i.e., when normalization is effective and reduces the overlap between categories). Likewise, this value will be negative when normalization is ineffective and actually increases the overlap between categories.

As stated in the main text, this equation reveals that when different categories differ in the contexts that they are likely to occur in, then normalization may actually increase the amount of overlap between different categories.

References

Adelson, E. H. (1993). Perceptual organization and the judgment of brightness. *Science*, *262*(5142), 2042–2044.

Ainsworth, W. (1973). Durational cues in the perception of certain consonants. *Proceedings of the British Acoustical Society*, *2*, 1–4.

Ainsworth, W. (1974). The influence of precursive sequences on the perception of synthesized vowels. *Language and Speech*, *17*(2), 103–109.

Allen, J. S., Miller, J. L., & DeSteno, D. (2003). Individual talker differences in voice-onset-time. *The Journal of the Acoustical Society of America*, *113*(1), 544–552.

Antetomaso, S., Miyazawa, K., Feldman, N., Elsner, M., Hitczenko, K., & Mazuka, R. (2017). Modeling phonetic category learning from natural acoustic data. In *BUCLD 41: Proceedings of the 41st Annual Boston University Conference on Language Development*.

Apfelbaum, K. S., & McMurray, B. (2015). Relative cue encoding in the context of sophisticated models of categorization: Separating information from categorization. *Psychonomic Bulletin & Review*, *22*(4), 916–943.

Arai, T., Behne, D., Czigler, P., & Sullivan, K. (1999). Perceptual cues to vowel quantity: Evidence from Swedish and Japanese. In *Proceedings of the Swedish Phonetics Conference (FONETIK)*, (Vol. 81, pp. 8-11).

Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience*, *5*(8), 617.

Bar, M., & Ullman, S. (1996). Spatial context in recognition. *Perception*, *25*(3), 343–352.

Bion, R. A., Miyazawa, K., Kikuchi, H., & Mazuka, R. (2013). Learning phonemic vowel length from naturalistic recordings of Japanese infant-directed speech. *PLOS ONE*, *8*(2), e51594.

Boersma, P. (2001). Praat: A system for doing phonetics by computer. *Glot International*, *5*(9/10), 341–345.

Boucher, V. J. (2002). Timing relations in speech and the identification of voice-onset times: A stable perceptual boundary for voicing categories across speaking rates. *Perception & Psychophysics*, *64*(1), 121–130.

Brown, R. W., & Hildum, D. C. (1956). Expectancy and the perception of syllables. *Language*, *32*(3), 411–419.

Chen, H., Yamane, N., Rattasone, N. X., Demuth, K., & Mazuka, R. (2016). Japanese infants are aware of phonemic vowel length in novel words at 18 months. In *BUCLD 40: Proceedings of the 40th Annual Boston University Conference on Language Development*.

Cole, J., Linebaugh, G., Munson, C., & McMurray, B. (2010). Unmasking the acoustic effects of vowel-to-vowel coarticulation: A statistical modeling approach. *Journal of Phonetics*, *38*(2), 167–184.

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), 101–112.

Dilley, L. C., & Pitt, M. A. (2010). Altering context speech rate can cause words to appear or disappear. *Psychological Science*, *21*(11), 1664–1670.

Dillon, B., Dunbar, E., & Idsardi, W. (2013). A single-stage approach to learning phonological categories: Insights from Inuktitut. *Cognitive Science*, *37*(2), 344–377.

Elsner, M., Goldwater, S., Feldman, N., & Wood, F. (2013a). A joint learning model of word segmentation, lexical acquisition, and phonetic variability. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 42–54).

Feldman, N. H., Griffiths, T. L., Goldwater, S., & Morgan, J. L. (2013a). A role for the developing lexicon in phonetic category acquisition. *Psychological Review*, *120*(4), 751.

Feldman, N. H., Myers, E. B., White, K. S., Griffiths, T. L., & Morgan, J. L. (2013b). Word-level information influences phonetic learning in adults and infants. *Cognition*, *127*(3), 427–438.

Fujisaki, H., & Kunisaki, O. (1978). Analysis, recognition, and perception of voiceless fricative consonants in Japanese. *IEEE Transactions on Acoustics on Speech, and Signal Processing*, *26*(1), 21–27.

Fujisaki, H., Nakamura, K., & Imoto, T. (1975). Auditory perception of duration of speech and non-speech stimuli. *Auditory Analysis and Perception of Speech*, 197–219.

Fukui, S. (1978). Perception for the Japanese stop consonants with reduced and extended durations. *Onsei Gakkai Kaihou*, *59*, 9–12.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, *6*(1), 110.

Guevara-Rukoz, A., Cristia, A., Ludusan, B., Thiollière, R., Martin, A., & Mazuka, R. (2018). Are words easier to learn from infant-than adult-directed speech? A quantitative corpus-based investigation. *Cognitive Science*, *42*(5), 1586–1617.

Han, M. S. (1994). Acoustic manifestations of mora timing in Japanese. *The Journal of the Acoustical Society of America*, *96*(1), 73–82.

He, A. X., & Lidz, J. (2017). Verb learning in 14-and 18-month-old English-learning infants. *Language Learning and Development*, *13*(3), 335–356.

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.

Hillenbrand, J., Clark, M. J., & Nearey, T. M. (2001). Effects of consonant environment on vowel formant patterns. *The Journal of the Acoustical Society of America*, *109*(2), 748–763.

Hirata, Y. (2004). Effects of speaking rate on the vowel length distinction in Japanese. *Journal of Phonetics*, *32*(4), 565–589.

Hisagi, M., Shafer, V. L., Strange, W., & Sussman, E. S. (2010). Perception of a Japanese vowel length contrast by Japanese and American English listeners: Behavioral and electrophysiological measures. *Brain Research*, *1360*, 89–105.

Höhle, B., Weissenborn, J., Kiefer, D., Schulz, A., & Schmitz, M. (2004). Functional elements in infants' speech processing: The role of determiners in the syntactic categorization of lexical elements. *Infancy*, *5*(3), 341–353.

House, A. S. (1961). On vowel duration in English. *The Journal of the Acoustical Society of America*, *33*(9), 1174–1178.

Imai, M., & Kita, S. (2014). The sound symbolism bootstrapping hypothesis for language acquisition and language evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1651), 20130298.

Isei-Jaakkola, T. (2004). Lexical quantity in Japanese and Finnish. Unpublished doctoral dissertation.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. *Talker Variability in Speech Processing*, 145–165.

Johnson, K. (2006). Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of Phonetics*, *34*(4), 485–499.

Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives. *The Journal of the Acoustical Society of America*, *108*(3), 1252–1263.

Kawahara, S. (2006). A faithfulness ranking projected from a perceptibility scale: The case of [+ voice] in Japanese. *Language*, 536–574.

Keating, P., Cho, T., Fougeron, C., & Hsu, C. S. (2004). Domain-initial articulatory strengthening in four languages. *Phonetic Interpretation: Papers in Laboratory Phonology VI*, 143–161.

Kinoshita, K., Behne, D. M., & Arai, T. (2002). Duration and F0 as perceptual cues to Japanese vowel quantity. In *Seventh international conference on spoken language processing*.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*(2), 148.

Kubozono, H. (2002). Temporal neutralization in Japanese. In *Laboratory Phonology 7*, (pp. 171–2002). Cambridge: Cambridge University Press.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, *255*, 606–608.

Lehnert-LeHouillier, H. (2010). A cross-linguistic investigation of cues to vowel length perception. *Journal of Phonetics*, *38*(3), 72–482.

Luce, P. A., & Charles-Luce, J. (1985). Contextual effects on vowel duration, closure duration, and the consonant/vowel ratio in speech production. *The Journal of the Acoustical Society of America*, *78*(6), 1949–1957.

Ludusan, B., Cristia, A., Martin, A., Mazuka, R., & Dupoux, E. (2016). Learnability of prosodic boundaries: Is infant-directed speech easier? *The Journal of the Acoustical Society of America*, *140*(2), 1239–1250.

Ludusan, B., Mazuka, R., Bernard, M., Cristia, A., & Dupoux, E. (2017). The role of prosody and speech register in word segmentation: A computational modelling perspective. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (volume 2: Short papers)*, (Vol. 2, pp. 178–183).

Mann, V. A., & Repp, B. H. (1980). Influence of vocalic context on perception of the [ʃ]-[s] distinction. *Attention, Perception, & Psychophysics*, *28*(3), 213–228.

Martin, A., Igarashi, Y., Jincho, N., & Mazuka, R. (2016). Utterances in infant-directed speech are shorter, not slower. *Cognition*, *156*, 52–59.

Massaro, D. W., & Cohen, M. M. (1983). Phonological context in speech perception. *Attention, Perception, & Psychophysics*, *34*(4), 338–348.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3), B101–B111.

Mazuka, R., Igarashi, Y., & Nishikawa, K. (2006). Input for learning Japanese: RIKEN Japanese mother–infant conversation corpus. The technical report of the Proceedings of the Institute of Electronics. *Information and Communication Engineers*, *106*(165), 11–15.

McMurray, B., & 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.

Miller, J. L. (1981). Effects of speaking rate on segmental distinctions. *Perspectives on the Study of Speech*, 39–74.

Miller, J. L., Grosjean, F., & Lomanto, C. (1984). Articulation rate and its variability in spontaneous speech: A reanalysis and some implications. *Phonetica*, *41*(4), 215–225.

Miller, J. L., & Liberman, A. M. (1979). Some effects of later-occurring information on the perception of stop consonant and semivowel. *Perception & Psychophysics*, *25*(6), 457–465.

Miller, J. L., O'Rourke, T. B., & Volaitis, L. E. (1997). Internal structure of phonetic categories: Effects of speaking rate. *Phonetica*, *54*(3-4), 121–137.

Minifie, F., Kuhl, P., & Stecher, E. (1977). Categorical perception of /b/ and /w/ during changes in rate of utterance. *The Journal of the Acoustical Society of America*, *62*(S1), S79–S79.

Mintz, T. H. (2006). Finding the verbs: Distributional cues to categories available to young learners. *Action Meets Word: How Children Learn Verbs*, 31–63.

Monahan, P. J., & Idsardi, W. J. (2010). Auditory sensitivity to formant ratios: Toward an account of vowel normalisation. *Language and Cognitive Processes*, *25*(6), 808–839.

Moreton, E., & Amano, S. (1999). Phonotactics in the perception of Japanese vowel length: Evidence for long-distance dependencies. In *EUROSPEECH*.

Mugitani, R., Pons, F., Fais, L., Dietrich, C., Werker, J. F., & Amano, S. (2009). Perception of vowel length by Japanese- and English-learning infants. *Developmental Psychology*, *45*(1), 236.

Narayan, C. (2008). The acoustic–perceptual salience of nasal place contrasts. *Journal of Phonetics*, *36*(1), 191–217.

Narayan, C. (2013). Developmental perspectives on phonological typology and sound change. *Origins of Sound Change: Approaches to Phonologization*, 128–146.

Narayan, C., Peters, A., & Woldenga-Racine, V. (2017). Fragile phonetic contrasts in longitudinal infant-directed speech: Implications for infant speech perception. In *BUCLD 42: Proceedings of the 41st Annual Boston University Conference on Language Development*.

Nearey, T. (1978). Vowel space normalization in synthetic stimuli. *The Journal of the Acoustical Society of America*, *63*, 1.

Nearey, T. (1990). The segment as a unit of speech perception. *Journal of Phonetics*.

Newman, R. S., Clouse, S. A., & Burnham, J. L. (2001). The perceptual consequences of within-talker variability in fricative production. *The Journal of the Acoustical Society of America*, *109*(3), 1181–1196.

Newman, R. S., & Sawusch, J. R. (1996). Perceptual normalization for speaking rate: Effects of temporal distance. *Attention, Perception, & Psychophysics*, *58*(4), 540–560.

Pickett, J., & Decker, L. R. (1960). Time factors in perception of a double consonant. *Language and Speech*, *3*(1), 11–17.

Pierrehumbert, J. (2002). Word-specific phonetics. *Laboratory Phonology*, 7.

Port, R. F., & Dalby, J. (1982). Consonant/vowel ratio as a cue for voicing in English. *Attention, Perception, & Psychophysics*, *32*(2), 141–152.

Rakerd, B., Sennett, W., & Fowler, C. A. (1987). Domain-final lengthening and foot-level shortening in spoken English. *Phonetica*, *44*(3), 147–155.

Richter, C., Feldman, N. H., Salgado, H., & Jansen, A. (2017). Evaluating low-level speech features against human perceptual data. In *Transactions of the Association for Computational Linguistics*.

Sato, Y., Sogabe, Y., & Mazuka, R. (2010). Discrimination of phonemic vowel length by Japanese infants. *Developmental Psychology*, *46*(1), 106.

Sawusch, J. R., & Newman, R. S. (2000). Perceptual normalization for speaking rate II: Effects of signal discontinuities. *Attention, Perception, & Psychophysics*, *62*(2), 285–300.

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

Shi, R., & Melançon, A. (2010). Syntactic categorization in French-learning infants. *Infancy*, *15*(5), 517–533.

Shi, R., & Werker, J. F. (2001). Six-month-old infants' preference for lexical words. *Psychological Science*, *12*(1), 70–75.

Shi, R., Werker, J. F., & Morgan, J. L. (1999). Newborn infants' sensitivity to perceptual cues to lexical and grammatical words. *Cognition*, *72*(2), B11–B21.

Strand, E. A., & Johnson, K. (1996). Gradient and visual speaker normalization in the perception of fricatives. *Konvens*, 14–26.

Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, *7*(5), 1074.

Swingley, D. (2009). Contributions of infant word learning to language development. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *364*(1536), 3617–3632.

Swingley, D., & Alarcon, C. (2018). Lexical learning may contribute to phonetic learning in infants: A corpus analysis of maternal Spanish. *Cognitive Science*.

Thiessen, E. D. (2007). The effect of distributional information on children's use of phonemic contrasts. *Journal of Memory and Language*, *56*(1), 16–34.

Todorović, D. (2010). Context effects in visual perception and their explanations. *Review of Psychology*, *17*(1), 17–32.

Toscano, J. C., & McMurray, B. (2012). Cue-integration and context effects in speech: Evidence against speaking-rate normalization. *Attention, Perception, & Psychophysics*, *74*(6), 1284–1301.

Umeda, N. (1975). Vowel duration in American English. *The Journal of the Acoustical Society of America*, *58*(2), 434–445.

Vance, T. J. (1987). *An introduction to Japanese phonology*. SUNY Press.

van Heugten, M., & Johnson, E. K. (2014). Learning to contend with accents in infancy: Benefits of brief speaker exposure. *Journal of Experimental Psychology: General*, *143*(1), 340.

Van Santen, J. P. (1992). Contextual effects on vowel duration. *Speech Communication*, *11*(6), 513–546.

Verbrugge, R. R., Strange, W., Shankweiler, D. P., & Edman, T. R. (1976). What information enables a listener to map a talker's vowel space? *The Journal of the Acoustical Society of America*, *60*(1), 198–212.

Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science*, *167*(3917), 392–393.

Wayland, S. C., Miller, J. L., & Volaitis, L. E. (1992). The influence of sentence articulation rate on the internal structure of phonetic categories. *The Journal of the Acoustical Society of America*, *92*(4), 2465–2465.

Wayland, S. C., Miller, J. L., & Volaitis, L. E. (1994). The influence of sentential speaking rate on the internal structure of phonetic categories. *The Journal of the Acoustical Society of America*, *95*(5), 2694–2701.

Werker, J. F., Pons, F., Dietrich, C., Kajikawa, S., Fais, L., & Amano, S. (2007). Infant-directed speech supports phonetic category learning in English and Japanese. *Cognition*, *103*(1), 147–162.