Psychon Bull Rev (2016) 23:1681–1712
DOI 10.3758/s13423-016-1049-y

THEORETICAL REVIEW

A unified account of categorical effects in phonetic perception

Yakov Kronrod[1,4] · Emily Coppess[2] · Naomi H. Feldman[3]

Published online: 24 May 2016

Abstract Categorical effects are found across speech sound categories, with the degree of these effects ranging from extremely strong categorical perception in consonants to nearly continuous perception in vowels. We show that both strong and weak categorical effects can be captured by a unified model. We treat speech perception as a statistical inference problem, assuming that listeners use their knowledge of categories as well as the acoustics of the signal to infer the intended productions of the speaker. Simulations show that the model provides close fits to empirical data, unifying past findings of categorical effects in consonants and vowels and capturing differences in the degree of categorical effects through a single parameter.

Keywords Perceptual magnet effect · Categorical perception · Speech perception · Bayesian inference · Rational analysis

✉ Yakov Kronrod
yakovkronrod@gmail.com

1 Department of Psychology, University of Pennsylvania, Philadelphia PA, USA

2 Department of Linguistics, University of Chicago, Chicago IL, USA

3 Department of Linguistics and UMIACS, University of Maryland, College Park MD, USA

4 Institute for Research in Cognitive Science, University of Pennsylvania, 3401 Walnut St. Suite 400A, Philadelphia, PA 19104, USA

Introduction

Assigning categories to perceptual input allows people to sort the world around them into a meaningful and interpretable package. This ability to streamline processing applies to various types of input, both linguistic and non-linguistic in nature (Harnad, 1987). Evidence that these categories affect listeners' treatment of perceptual stimuli has been found in diverse areas such as color perception (Davidoff et al., 1999), facial expressions (Angeli et al., 2008; Calder et al., 1996), familiar faces (Beale & Keil, 1995), artificial categories of objects (Goldstone et al., 2001), speech perception (Liberman et al., 1957; Kuhl, 1991), and even emotions (Hess et al., 2009; Sauter et al., 2011). Two core tendencies are found across these domains: a sharp shift in the identification function between category centers, and higher rates of discrimination for stimuli from different categories than for stimuli from a single category. Nowhere is this more evident than in speech perception, where these perceptual effects are viewed as a core component of our ability to perceive a discrete linguistic system while still allowing for informative variation in the speech signal.

In speech perception, categorical effects are found in a wide range of phonemes. However, different phoneme classes differ in the degree to which the categories influence listeners' behavior. At one end of the spectrum, discrimination of stop consonants is strongly affected by the categories to which they belong. Discrimination is little better than would be expected if listeners used only category labels to distinguish sounds, and between-category differences are extremely pronounced (Liberman et al., 1957; Wood, 1976). At the other end of the spectrum, vowel discrimination

is much more continuous, so much so that some early experiments seemed to suggest that vowels displayed no categorical effects at all (Fry et al., 1962). Since these classic studies, it has become evident that stop consonant perception is not purely categorical, while vowel perception can also exhibit categorical effects (Pisoni & Lazarus, 1974; Pisoni, 1975). In addition, there is evidence that rather than being purely perceptual, the influence of categories on discrimination behavior may arise later in processing (Toscano et al., 2010; but see Lago et al., 2015). Nevertheless, where the categorical effects do occur, the degree to which consonants are affected is much greater than that of vowels. Researchers have proposed a number of qualitative explanations for these differences. For example, the differences have been claimed to stem from the way each type of sound is stored in memory (Pisoni, 1973), to be related to innate auditory discontinuities that could influence stop consonant perception (Pisoni, 1977; Eimas et al., 1971), and to result from different processing mechanisms for steady state and rapidly changing spectral cues (Mirman et al., 2004). However, qualitatively the effects are very similar between consonants and vowels, with a sharp shift in the identification function and a peak in discrimination near the category boundary. These qualitative similarities suggest that these two cases may be interpretable as instantiations of the same phenomenon. That is, perceptual differences among different classes of sounds may be purely quantitative rather than qualitative.

Past models have focused on providing a mechanism by which strong categorical perception may arise for consonants, describing the origin of perceptual warping in vowel perception, or exploring general categorical effects without accounting for differences between stop consonants and vowels, but no model has provided a joint explanation of categorical effects together with an account of the variation in the degree of these effects. In this paper we show that categorical effects in consonant and vowel perception can be captured by a single model. We adapt a Bayesian model proposed by Feldman et al. (2009), which analyzes categorical effects as resulting from the optimal solution to the problem of perceiving the speech sound produced by the speaker. The model predicts that the strength of categorical effects is controlled by a single parameter, representing the degree to which within-category acoustic variability contains information that listeners want to recover. Thus, consonants and vowels may simply differ in how much of their variability is meaningful to listeners. Through simulations, we characterize several classes of sounds along this continuum and show the model can provide a unified framework for both strong and weak categorical effects.

We explore the possibility of a cohesive underlying model purely at the computational level, in the sense of Marr (1982). Marr proposed three possible levels at which one might approach this problem: computation, representation and algorithm, and physical implementation. For understanding speech perception, each level of analysis has a unique contribution. It would be impossible to paint a full picture of speech perception without being able to provide explanations at each level independently and show how these explanations relate to each other. However, it is not necessary to consider all three levels simultaneously, and specifying the model only at the computational level is advantageous in that it allows for the possibility of varying algorithmic and implementational levels of analysis for different sets of sounds, while still retaining the idea that all of these carry out the same basic computation. That is, while the perceptual dimensions that are relevant to perceiving stop consonants and vowels are not likely to have the same neural implementation, we show that the computations performed over those perceptual dimensions serve the same purpose.

In the remainder of the paper, we proceed as follows. First, we review previous findings on categorical effects in stop consonants, vowels, and fricatives, considering whether separate mechanisms are needed to account for the observed effects of categories. Next, we review a Bayesian model of speech perception that was originally proposed to capture categorical effects in vowels, and extend it for evaluating effects for various phonemes. We conduct simulations showing that the model provides a close match to behavioral findings from stop consonants and fricatives as well as vowels, capturing differences in the degree of categorical effects across consonants and vowels by varying a single parameter. We conclude by discussing the significance and implications of these findings.

Categorical effects in speech perception

In order to appreciate the differences and similarities between categorical effects for different phonemes, it is insightful to review the classic findings in these domains. In this section we introduce the methods that are used to study categorical effects in speech perception and review descriptions of categorical effects for the three classes of phonemes that we later consider using our unified model: stop consonants, fricatives, and vowels. This overview centers on two key models that have been put forward for characterizing the effect of categories on stop consonant and vowel discrimination: categorical perception (CP) and the perceptual magnet effect (PME) (Liberman et al., 1957; Kuhl, 1991). Although we do not claim that either of these is an entirely accurate model of perception, they provide a useful historical context for introducing the basic phenomena of interest.

Behavioral measures and perceptual warping

Categorical effects in speech perception are typically studied through behavioral identification and discrimination tasks, which provide data on listeners' ability to classify the sounds (identification) and to differentiate sounds along an acoustic continuum (discrimination). The stimuli that participants hear in each task typically lie along a one-dimensional continuum between two phonemes. For presentation purposes, we consider a continuum between two phonemes, c_1 and c_2, with seven equally spaced stimuli $S_1 ... S_7$. For example, if $c_1 =$ /b/ and $c_2 =$ /p/, stimuli might be created by varying the voice onset time (VOT) of the signal.

The identification task consists of choosing between two competing labels, c_1 and c_2, in a forced choice paradigm. Participants choose one of the two labels for every stimulus heard, even if they are unsure of the proper classification. By examining the frequency with which participants choose each category, we can observe an apparent boundary between the categories and can determine the sharpness of this boundary. The shape of the identification curve provides information about the distribution of sounds in the categories that the listener expects to hear. If the categories are sharply concentrated with little perceptual overlap, then we would expect more absolute identification and a sudden switch between category labels - resulting in a steep curve. Alternatively, if categories are more diffuse, we would see a shallower curve, i.e., a more gradual switch between category labels. An illustrative example in the presence of sharply concentrated categories can be seen in the solid line in Fig. 1.

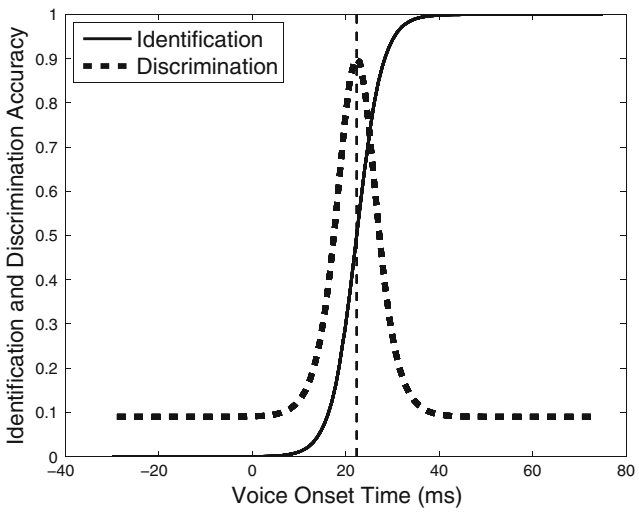

Fig. 1 Hypothetical identification and discrimination functions in the presence of strong category influences

Discrimination can be measured in a number of ways (e.g., AX, ABX, or triad presentation), but all of these methods test the ability of a listener to determine whether sounds are the same or different. Our simulations focus on studies employing the AX discrimination paradigm, where listeners are presented with two stimuli, A and X. Their task is to say whether the X stimulus is identical to the A stimulus or different from it. This task measures how likely discrimination is to occur, and therefore serves as a measure of perceptual distance between the stimuli. By considering all equidistant pairs of stimuli along the continuum we can see how listeners' ability to discriminate sounds changes as we move from the category centers to the category boundary. With no categorical effects, we would expect uniform discrimination along the continuum, due to equal spacing of the stimuli (i.e., listeners should be just as good at telling S_1 and S_3 apart as S_3 and S_5). If perception is biased toward category centers, listeners' ability to differentiate stimuli near category centers should go down while their ability to differentiate stimuli at the category boundary should go up. An illustrative example of discrimination performance in the presence of strong influences of categories can be seen in the dashed line in Fig. 1. The degree of warping in discrimination, corresponding roughly to the height of the discrimination peak near the category boundary, is a key indicator of how strong the category effects are. It will be used in our model to tease apart different sources of variability in the individual sound categories.

Identification and discrimination tasks have been used to investigate sound perception across a variety of phonemes. By considering both of these tasks together, we can determine the characteristics of listeners' perceptual categories as well as the extent to which these categories affect their discrimination of sounds. The remainder of this section reviews categorical effects in perception that have been studied using these paradigms, focusing on the perception of stop consonants, vowels, and fricatives, the three classes of phonemes to whose behavioral data we fit our model.

Phonemes considered in this paper

The three phoneme classes we consider in this paper are stop consonants, vowels, and fricatives. The stop consonants we consider are along the voice onset time continuum between bilabial consonants /p/ and /b/. The primary acoustic cues used in the classification of these consonants are static temporal properties, specifically the amount of time between the release burst and the onset of voicing. The vowels we consider lie along the dimension between /i/ and /e/. These vowels are largely identified by the first and second formant, representing the steady state peaks of resonant energy with a certain bandwidth, a

static spectral cue. Finally we consider the fricatives lying between /ʃ/ and /s/. In terms of cues to identification, these fricatives are identified largely by the static spectral peak locations, though there are many other cues that are also relevant to fricative identification (see (McMurray and Jongman, 2011) for a review). Below we go through more details on the three classes of phonemes more broadly, considering behavioral findings, explanations for categorical effects, and models meant to capture the source of these effects.

Stop consonant effects

Stop consonants provide a prototypical example of a strong effect of categories on listeners' perception of sounds. Liberman et al. (1957) showed that discrimination of stop consonants is only slightly better than would be predicted if listeners only used category labels produced during identification and ignored all acoustic detail. They labeled this observation *categorical perception* (CP). The core tenet of pure CP for stop consonants is that participants do not attend to small differences in the stimuli, rather treating them as coarse categories, ignoring some of the finer detail in the acoustic stream. Under the CP hypothesis, participants assign a category label to each stimulus and then make their discrimination judgment based on a comparison of these labels.

To model participants' identification and discrimination data from a place of articulation continuum ranging from /b/ to /d/ to /g/, Liberman et al. (1957) formulated a probabilistic model which used the probabilities from the identification task to make predictions about how often the listener would be able to discriminate the sounds based only on category assignments. This produces the strongest possible categorical effect. They found that using this formula, they could predict overall discrimination behavior very well. The participants' actual discrimination only outperformed the predictive model slightly. However, the fact that the participants did outperform the model suggests that they were able to use acoustic cues beyond pure category membership.

Since Liberman et al.'s initial findings, there has been extensive research into other phonetic environments with similar categorical effects. Critically for our work here, strongly categorical perception was found by Wood (1976) for the voicing dimension. Strong categorical effects have also been found for /b/-/d/-/g/ by other researchers (Eimas, 1963; Griffith, 1958; Studdert-Kennedy et al., 1963, 1989), as well as for /d/-/t/ (Liberman et al., 1961), /b/-/p/ in intervocalic position (Liberman et al., 1961), and the presence or absence of /p/ in slit vs. split (Bastian et al., 1959; Bastian et al., 1961; Harris et al., 1961). These strong categorical effects can be modulated by contextual effects and task-related factors (Pisoni, 1975; Repp et al., 1979), but are generally viewed as robust.

In contrast with the original formulation of categorical perception, evidence has accumulated showing that perception of stop consonants is not purely categorical. Listeners pay attention to subphonemic detail, as evidenced by various behavioral and neural studies. Studies have shown that goodness ratings of stop consonants vary within categories and are prone to context effects based both on phonetic environment and speech rate (Miller, 1994). Internal structure of consonant categories is further supported by studies of reaction time, with Pisoni and Tash (1974) showing that participants are slower to respond *same* to acoustically different within category pairs of sounds than for pairs of sounds that are acoustically identical. Further, priming studies by Andruski et al. (1994) found priming effects of within-category VOT differences for short inter-stimulus intervals of 50 ms. They showed that stimuli with initial stop consonant VOTs near the category center exhibited a stronger priming effect for semantically-related following stimuli, and that non-central values also elicited longer reaction times. Finally, at the neural level, an fMRI study by Blumstein et al. (2005) showed that there are robust neural correlates to subphonemic VOT differences in stimuli. In related work looking at event-related potentials during word categorization in an auditory oddball task, Toscano et al. (2010) showed that listeners are sensitive to fine acoustic differences in VOT independent of the categorization. Effects were found both at a pre-categorization late perceptual stage 100 ms post stimulus as well as in the post-perceptual categorization stage around 400 ms post stimulus, indicating that fine acoustic detail is carried through the entire perceptual process. Together, these studies strongly suggest that not all members of the category are perceived as truly equal and that identification and discrimination performance is not based on an all-or-none scheme. While this goes against the original CP hypothesis, it takes nothing away from the observation that in discrimination tasks, stop consonants are prone to perceptual warping, and that this warping appears to correlate closely with their classification into categories.

Vowel effects

Vowels exhibit less influence from categories and are perceived more continuously than stop consonants, with listeners exhibiting higher sensitivity to fine acoustic detail. Unlike consonants, vowel discriminability cannot be closely predicted from the identification data, which itself is much more prone to context effects (Eimas, 1963). Relatively continuous perception has been found repeatedly in the

/i/-/I/-/E/ continuum (Fry et al., 1962; Stevens et al., 1963; Stevens et al., 1964), as well as in perception of vowel duration (Bastian and Abramson, 1962) and perception of tones in Thai (Abramson, 1961). Further support for more continuous perception of vowels comes from mimicry experiments (Chistovich, 1960; Kozhevnikov & Chistovich, 1965): When participants were asked to mimic stop consonants and vowels, their ability to reproduce vowels accurately was much greater than for consonants, which tended to be reproduced with prototypical members of the category. It should be noted that these findings are all for steady-state vowels, and do not necessarily pertain to vowels in speech contexts that contain rapidly changing formant structures. Stevens (1966) found that one could obtain nearly categorical perception when looking at vowels between consonants pulled out of a rapidly articulated stream. However, for comparisons in this work, we will focus on findings in the perception of isolated vowels.

Kuhl (1991) took a different approach to investigating the role of categories in vowel perception by examining the relationship between discrimination and goodness ratings. In goodness rating tasks, participants give numerical ratings to indicate how good an example a stimulus is of a specific category. Goodness ratings collected by Kuhl (1991) along a vowel continuum near /i/ confirmed that there was variable within-category structure that people could represent and access; multiple participants shared the center of goodness ratings (i.e., the location where stimuli were rated highest on category fit) and had similar judgments for stimuli expanding radially from the center. These findings suggest that participants have a stable representation of the category for /i/ and that its structure does not represent an all-or-nothing judgment of category membership, but rather a gradient representation.

To determine how this gradient representation relates to the perception and discriminability of individual stimuli, adults, children, and monkeys were asked to discriminate sounds equally spaced around the prototypical and non-prototypical /i/ stimuli. Both adults and children were more likely to perceive stimuli around the prototypical category member as the same sound as compared to sounds around the non-prototype. However, monkeys did not show any effect of prototypicality. This suggested that humans' discrimination abilities depended on linguistically informed representations of category structure and showed that perception of vowels is not entirely veridical, even if the precise nature of the effect differs from that of stop consonants.

These findings led Kuhl to propose the *perceptual magnet effect*. She described this effect as a within-category phenomenon, focusing on the relationship between category goodness ratings and discriminability. The claim was that stimuli that are judged to be better exemplars of a category act as "perceptual magnets", making stimuli around them harder to discriminate. Meanwhile, stimuli judged to be poor exemplars exhibit very little effect on neighboring vowels. As a result, under the perceptual magnet hypothesis, there is a correlation between category goodness judgments and discriminability. Iverson and Kuhl (1995) showed a direct link between goodness ratings and discriminability, proposing this as a central tenet of the perceptual magnet effect. Critically, the effect resembles other categorical effects, but contains additional predictions regarding goodness ratings. If one considers an extreme version of perceptual magnets acting on the surrounding stimuli, we could get something akin to the original formulation of categorical perception. Hence, we can immediately see a possible relationship between this view of categorical effects in vowels and consonants.

The extent to which the perceptual magnet effect generalizes to other sound types is an open question. There were documented replications of the perceptual magnet effect in the /i/ category in German (Diesch et al., 1999) and Swedish (Kuhl et al., 1992; Aaltonen et al., 1997), but also failed replication attempts for American English (Lively & Pisoni, 1997; Sussman & Gekas, 1997) and Australian English (Thyer et al., 2000). Additionally, there was a failure to find evidence of the perceptual magnet effect for certain other vowel categories in English (Thyer et al., 2000). However, it was also found for the Swedish /y/ category (Kuhl et al., 1992; Aaltonen et al., 1997) as well as the lateral and retroflex liquids (/l/,/r/) in American English (Iverson & Kuhl, 1996; Iverson et al., 2003). This suggests that it is a robust effect that extends to at least some types of consonants, even if the precise nature of the stimuli that elicit it is unclear

Fricative effects

Fricatives have spectral properties that make them interesting to consider in relation to work on other phonemes. They are consonants; however, they also share properties with vowels, in that they can largely be identified by their spectral properties. Specifically, sibilant fricatives [s] and [ʃ] are identified by their two primary frication frequencies, analogously to how vowels are identified by their two primary formants (F_1 and F_2). The precise spectral frequency cues are different in that fricatives have higher frequency aperiodic noise and vowels consist primarily of lower frequency periodic energy. However, they are qualitatively similar, in that these frequencies are key to the identification of particular sounds. Because of this similarity, fricatives serve as an interesting case to explore perception behavior that may fall intermediate between vowels and consonants.

Fricatives pattern with vowels in other aspects of perception as well. In their work on classifying the properties of perception of various forms of speech, Liberman et al., (1967) considered a sound's tendency to show restructuring—exhibiting varying acoustic representations as a result of varying context—and how this related to observed categorical effects. Stop consonants were found to exhibit a large amount of restructuring, changing how they appear acoustically even though they have the same underlying phonemic status. This was found in both correlates of place of articulation (Liberman et al., 1967) as well as manner and voicing (Lisker & Abramson, 1964b; Liberman et al., 1954). Steady state vowels, on the other hand, show no such restructuring, when accounting for speaker normalization and speaking rates (Liberman et al., 1967). Liberman et al. consider the noise produced at the point of constriction in both fricatives and stop consonants, and argue that for longer duration of the noise, precisely the kind found in fricatives, the cue does not change with context. This lack of restructuring was shown specifically for the perception of /s/ and /ʃ/ by Harris (1958) and Hughes and Halle (1956). Thus, if categorical effects are related to restructuring, fricatives may pattern with vowels rather than stop consonants in their categorical effects.

Experiments designed to evaluate the effects of categories on the perception of fricatives have led to mixed results. In behavioral experiments with fricatives, Repp (1981) found that participants' behavior was similar to that originally found in experiments with stop consonants, indicating strong effects of categories for fricatives. However, during the course of the same study, some participants exhibited perception that was much more continuous. To accommodate this apparent contradiction in the findings, Repp proposed that participants were using two distinct processing strategies: acoustic and phonetic processing. Phonetic processing refers to a mode of perception where listeners are actively assigning phonetic category classifications, whereas acoustic processing refers to attention to the fine-grained acoustic variability of the signal.

More recently, Lago et al. (2015) investigated categorical effects in fricatives by focusing on the continuum between sibilant fricatives /s/ and /ʃ/. They conducted an identification task, an AX discrimination task, and a goodness judgment task. Their results showed a strong effect of categories on the perception of the stimuli with no strong correlation between discriminability and goodness ratings. Qualitatively, their identification findings showed a sharp change in identification near the category boundary, but a discrimination peak that was markedly shallower than expected. This suggested that fricatives employ a representation that retains more acoustic detail than pure category assignment, making them not as strongly categorical as stop consonants, but also not as continuous as vowel continua.

Models of categorical effects in consonant and vowel perception

Models that have been proposed have largely been split between those focused on strong categorical effects for stop consonants and those focused on weaker effects for vowels. Here we present a brief overview of the models covering a range of approaches. First we consider models focused on categorical perception and generally strong categorical effects. Various models have been put forward to explain the source of these strong categorical effects. Initially, researchers assumed that categorical perception resulted from psychophysical properties of processing speech and argued that it was specific to language processing (Macmillan et al., 1977). Other researchers tended to use more general views of either statistical properties or higher order cognitive processing to explain the effect. Massaro (1987a) used signal detection theory (SDT) to model the identification and discrimination tasks in two stages, sensory and decision operations, leading to a separation of sensitivity and response bias. In their model, categorical behavior can arise from classification behavior even if perception is continuous, with Massaro calling it *categorical partition* instead of *categorical perception*. The separation of perception and decision making processes was further investigated by Treisman et al. (1995), who applied criterion-setting theory (CST) (Treisman and Williams, 1984) to categorical perception. Their work models the sensory system as able to reset the internal criterion for decision making based on most recently available data, much like Bayesian belief updating. Elman (1979) showed that such a model of criterion setting is able to capture the original stop consonant findings (Liberman et al., 1967) even better than their original Haskins model.

Other researchers considered the problem at a different level of analysis, focusing instead on the possible neural implementation of the categorical perception mechanism. Vallabha et al. (2007) proposed a multi-layer connectionist model that operates on Gaussian distributions of speech sounds as the input and produces categorical effects via interactions of three levels of representation: an acoustic input layer, an intermediate perceptual layer, and a categorical classification output layer. The key to their model is the presence of bidirectional connections between the output category level and hidden perceptual layer, whereby the perception influences the classification, but the classification simultaneously biases perception toward category centers.

The setup of the model and use of bidirectional links to create top-down influences is similar to the TRACE model of speech perception proposed by McClelland and Elman (1986), where a feature level, phoneme level, and word level were used to explain various features of speech perception. Other neural network models have also been proposed to show a biologically plausible mechanism by which these categorical perception effect could arise. Damper & Harnad (2000) trained both a Brain-State-In-A-Box (BSB) (following Anderson et al., 1977) and a back-propogation neural network model to show how categorical perception arises through spontaneous generation after training on two endpoint stimuli. They were able to produce typical categorical effects and reproduce the discrepancy between VOT boundaries between different places of articulation found in human participants. Going for even greater biological plausibility, Salminen et al. (2009) exposed a self-organizing neural network to statistical distributions of speech sounds represented by neural activity patterns. Their resulting neural map showed strongly categorical effects from single neurons being maximally activated by prototypical speech sounds, along with the greatest degree of variability in the produced signal at the category boundaries. Kröger et al. (2007) showed that categorical perception arises when using distributions consisting of specific features (bilabial, coronal, dorsal) to train self-organizing maps to learn phonetic categories and discriminate between sounds.

These models suggest that there are many possible processes that underlie strong categorical effects. However, these models are poorly adapted to capture effects going beyond the case of strong categorical perception described above, such as the weaker categorical effects found in vowel perception. For vowels, we focus on models related to explaining the perceptual magnet effect. Several models at different levels of processing have been proposed to explain the source of the perceptual magnet effect. One such theoretical model is the Native Language Magnet Theory (Kuhl, 1993), which proposed that prototypes exert a pull on neighboring sounds. However, this leaves open the question of why prototypes should exert a pull on neighboring speech sounds. An exemplar model was then proposed (Lacerda, 1995) that showed how the perceptual magnet effect could be construed as an emergent property of an exemplar-based model of phonetic memory. In his model, sound perception is guided by a simple similarity metric that operates on collections of exemplars stored in memory, with no need to refer to special prototypes to derive the sorts of effects typical of the perceptual magnet effect. This then leaves open the question of how do we fully account for within-category discrimination. For this we can consider low-level neural network models that attempt to provide a potential explanation of the type of connectionist network that can give rise to these perceptual effects. One such neural network models was proposed by Guenther and Gjaja (1996), where sensory experience guided the development of an auditory perceptual neural map and the vector representing cell firing corresponded to the perceived stimulus. Another model that Vallabha and McClelland (2007) considered modeled learning via distributions of speech sounds and used online mixture estimation and Hebbian learning to derive the effect. Both models showed how the effect might be derived from a biologically plausible mechanism. Finally, Feldman et al. (2009) proposed a Bayesian model in which listeners infer the phonetic detail of a speaker's intended target production through a noisy speech signal. It is this model that serves as the basis for the present work, where we try to show how both strong and weak categorical effects can be accounted for as a unified effect at the computational level.

Common ground in vowel and consonant perception

The existing evidence shows differing degrees of categorical effects across different phonemes. Stop consonant perception is characterized by very sharp identification shifts between two categories and a large peak in discrimination at the center between the categories. Vowel perception elicits more continuous identification functions, shallower peaks in discrimination at the category boundaries, and much greater within-category discrimination. Additionally, goodness ratings for vowels descend in gradient fashion from the center of the category outward (Iverson and Kuhl, 1995), while for stop consonants the goodness ratings vary only slightly within the category (particularly for /p/), while they exhibit a sharp jump in goodness at the category boundary (Miller & Volaitis, 1989). Models proposed for these effects only tend to work for subsets of categories. For stop consonants, the Liberman et al. (1967) model comes close to predicting discrimination based on the identification function, while this prediction fails in vowel perception experiments. Neural network models have been applied to either vowel and liquid perception (Guenther and Gjaja, 1996; Vallabha & McClelland, 2007) or to stop consonant perception (Damper & Harnad, 2000), but the same models have not typically been used to account for perception of both classes. While we now know that no phonemes are perceived purely categorically, the literature has nevertheless continued to treat strongly categorical perception as a separate phenomena from more continuous perception of other sounds, particularly vowels (Table 1).

However, perception of the different sound classes also has much in common qualitatively. Both stop consonants and vowels exhibit greater discriminability at category

Table 1 Stop consonant perception (proposed categorical perception) vs vowel perception (proposed Perceptual Magnet Effect)

Stop consonant perception	Vowel perception
• Poor within-category discrimination	• Graded within-category discrimination
• Greater discrimination across category boundaries	• Decreased discrimination around category centers
• Little within-category gradation	• Graded goodness judgments for category members
	• Correlation of goodness judgments and discriminability

boundaries, with the peak in discrimination being in close correspondence with the boundary found in identification. Stop consonants do exhibit some within-category discriminability, or at least within-category structure, as evidenced by reaction time measures (Pisoni & Tash, 1974; Massaro, 1987b). Lotto et al. (1998) also suggested that Kuhl's (1991) failure in the ability to predict vowel discrimination from identification was due to faulty identification data rather than an inherent difference in perception between consonants and vowels: By retesting identification in paired contexts, they removed the need to appeal to goodness of stimuli to explain reduced discriminability near category centers. They argued based on their analysis that the perceptual magnet effect was nothing more than categorical perception. Furthermore, while Iverson and Kuhl (2000) found that the correlation between discriminability and goodness ratings (a key feature of the perceptual magnet effect model for vowel perception) could be dissociated from the relationship between identification and discriminability (a key feature of categorical perception model for stop consonants), Tomaschek et al. (2011) found that these two relationships co-occur. The fact that fricative perception was not found to be strongly categorical, nor as continuous as vowels, and different studies reached different results depending on the task and measurements involved further suggests that strongly categorical and largely continuous perception are merely two ends of a continuum, and not two separate modes of perception, and that there can be gradient degrees of categorical effects that fall between these two extremes.

The goal of our simulations is to show that a common explanation can account for the behavioral findings in both consonants and vowels. We adapt a model that was originally proposed to account for vowel perception and show that it also provides a close match to empirical data from stop consonants and fricatives. We further show how parametric variation within the model can lead to varying strengths in categorical effects. We argue that while perception of the cues to different sounds is implemented differently at a neural level, strong categorical effects in consonant perception and the largely continuous perception of vowels reflect solutions to the same abstract problem of speech perception at Marr's 1982 computational level. Our analysis appeals to a kind of scientific Occam's Razor to argue that our unified account is the more parsimonious theory; a similar argument was used to substantiate a unified account of cumulative exposure on selective adaptation and phonetic recalibration by Kleinschmidt and Jaeger (2015).

Bayesian model of speech perception

To show that it is possible to interpret categorical effects across speech perception as qualitatively similar processes, we model these effects using a Bayesian model developed by Feldman et al. (2009) that was originally proposed to account for the perceptual magnet effect along the /i/-/e/ continuum. We apply an extension of this model to a broader range of data, encompassing data from stop consonant and

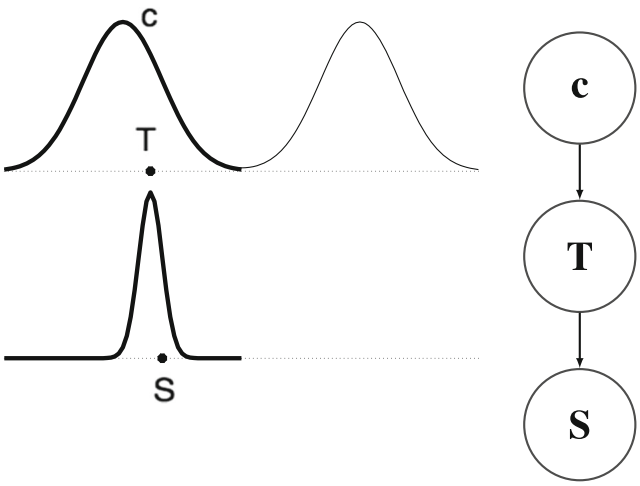

Fig. 2 Bayesian model of production of phonemes used for optimal speech inference. **a** These are the distributions involved in the generative process. c is the underlying phonetic category chosen from the two possible categories, T is the intended target production chosen by the speaker, and S is the perceived speech sound heard by the listener. **b** This is the graphical model of the generative process for speech production under this Bayesian model. T is sampled from a Normal distribution with meaningful category variance σ_c^2 around the category mean μ_c (i.e. $p(T|c) = N(\mu_c, \sigma_c^2)$). S is sampled from a normal distribution with noise variance σ_S^2 around the intended target production T (i.e. $p(S|T) = N(T, \sigma_S^2)$)

fricative perception. In doing so, we show that the categorical effects seen in consonant and vowel perception can be accounted for in a unified fashion.

Generative model

The model lays out the computational problems that listeners are solving as they perform identification and discrimination of sounds. In identification tasks, it assumes that the listener makes a choice among a set of categories while listening to sounds coming from the continuum between these categories. In discrimination tasks, the model assumes that the listener infers a continuous acoustic value that a speaker intended to produce, compensating for noise in the speech signal. For each of these two inference processes, the model formalizes the assumptions that the listener makes about the generative process that produced the sounds in order to determine how these assumptions affect their perception of the sounds. A graphical representation of the model appears in Fig. 2. For presentation purposes, and because of the nature of these particular studies, we restrict our attention to the case of two categories throughout the remainder of this paper. However, the model can in principle be applied to arbitrary numbers of categories, with Eqs. 2 and 9 being used to derive model predictions in the general case.

We begin with the listener's knowledge of the two categories, which we call c_1 and c_2. The next steps concern the process that the listener presumes to have generated the sounds heard. First, the speaker chooses one of the two categories, which we refer to as c. We refer to this category as the *underlying category*. This is not to be confused with the term *underlying category* used in phonology to represent the abstract phonological category that underlies allophonic variation. Instead, we use this term to refer to a phonetic category that has not been corrupted by noise in the speech signal. This category can be represented in our model as a Gaussian distribution around the category mean μ_c with variance σ_c^2. We are agnostic to the particular dimension for the mean and variance, but in practice it can represent any measure including VOT for stop consonants, F_1 and F_2 for vowels, and many others. This mean and variance of the category being used in the generative procedure by the speaker is known by the listener from previous exposure to sounds from this category in the language. For the purposes of our model, we don't concern ourselves with how such categories are learned, but rather the perception that occurs once the categories are already acquired. Hence, the mean represents the center of the category in perceptual space. The variance here is assumed by the listener to be derived from processes that provide useful information about the nature of the sound, indexical variables such as speaker identity, or the identities of upcoming sounds. Because of this, we call the categorical variance of the underlying category 'meaningful'. We consider in more detail in the General Discussion which types of factors might contribute to meaningful variance.

The next step in the generative process is the selection of an intended target production from the normal distribution $N(\mu_c, \sigma_c^2)$. We refer to the intended target production as T. The probability of choosing a specific target production, T, from a phonetic category, c, is $p(T|c) = N(\mu_c, \sigma_c^2)$. Once the intended target production is chosen, it needs to be articulated by the speaker and perceived by the listener. This process introduces additional articulatory, acoustic, and perceptual noise that distorts the signal. We formalize this as an additional Gaussian distribution around the intended target production with mean T and variance σ_S^2. We refer to the actual speech sound that the listener perceives as S and assume that it is sampled from the distribution of speech signal noise, with probability $p(S|T) = N(T, \sigma_S^2)$. We can also consider the overall distribution of possible speech sounds related to the underlying category chosen by the speaker at the beginning of the generative procedure. If we integrate over all possible intended target productions, T, then we can describe the distribution of speech sounds as $S|c = N(\mu_c, \sigma_c^2 + \sigma_S^2)$.

Given this generative model, we can consider how this model relates to the behavioral tasks described in the phoneme perception sections above. In those tasks, listeners are asked to either identify the category of the sound or to tell if two sounds are the same or different. In the model, the identification task relates to retrieving the underlying category, c. The discrimination task involves recovering the intended target production T for each of the stimuli heard by the listener and then comparing them to see if they are the same or different. The listener is presumed to be recovering both phonetic detail about the target production as well as category choice information when they perceive sounds. By fitting the model to the results of behavioral tasks performed by listeners, we find the optimal setting of parameters that best describes the data. We can then examine how these parameters relate to the degrees of categoricity seen in the perception of different phoneme continua.

Let us consider how the listener might be able to retrieve this information that they need. First, note that the listener does not have access to the intended target production, T, that the speaker meant to say. The listener does have knowledge of the underlying categories, $N(\mu_{c1}, \sigma_{c1}^2)$ and $N(\mu_{c2}, \sigma_{c2}^2)$, noise variance along the relevant perceptual dimension, σ_S^2, and the actual speech stimulus that they perceived, S. This means that the listener will use a combination of actual perceived speech information and knowledge of underlying categories in inferring what the speaker intended. This relationship between the contribution of S and μ_{c1} and μ_{c2} will become important as we

move forward in evaluating the varying effects of categories in identification and discrimination tasks. In terms of our model, the identification task will correspond to finding the probability of a given category given the speech sound. In other words, it entails computing $p(c|S)$. The discrimination task corresponds to finding the probability of a given target production given the same speech stimulus, or computing $p(T|S)$. For both of these inference procedures our model uses Bayes' rule, which we discuss below in relation to each of these tasks.

Bayes' rule

Bayes' rule is derived from a simple identity in probability theory (Bayes, 1763). It allows us to compute a belief in a hypothesis based on observed data, stating that the posterior probability of a hypothesis given some data can be calculated from the probability of the data given the hypothesis multiplied by the prior belief in the hypothesis and then normalized by the overall (marginal) probability of the data given all possible hypotheses,

$$p(h|d) = \frac{p(d|h)p(h)}{\sum_h p(d|h)p(h)} \tag{1}$$

The denominator on the right-hand side of the equation is the marginalized representation of the overall probability $p(d)$. The term of the left-hand side, $p(h|d)$ is called the posterior probability. The term $p(d|h)$ is called the likelihood, since it describes how likely the data are under a certain hypothesis. The final term, $p(h)$ is called the prior probability, since it is the probability of the hypothesis (i.e., belief in the hypothesis) before seeing data. Often these prior probabilities are uninformed and are set using a heuristic, or are uniformly distributed among all possible hypotheses, as is the case in our simulations of identification data below.

The hypotheses and data are different depending on what behavior we are modeling (identification vs. discrimination). In the section below we go through in detail how the inference procedure for the listener is structured, and what parameters we can extract by fitting the model to the listener's behavioral data.

Bayes' rule for identification

First we consider the behavioral task of identification. For a listener, the task of identifying a sound involves picking the correct category label for the sound. In our generative model, this means inferring the category c from the speech sound S (Table 2). Bayes' rule for identification is

$$p(c|S) = \frac{p(S|c)p(c)}{\sum_c p(S|c)p(c)} \tag{2}$$

If we rewrite this equation using the probability distributions for the prior probability and the likelihood given in our generative model, we can see the critical parameters that can be recovered by fitting the model. We make the simplifying assumption that both categories are equally probable before any speech sound is heard, substituting 0.5 for $p(c)$ in the equation. In other words, we are not taking into account different phoneme frequencies. Although vowel frequencies (Gimson, 1980; Wioland, 1972; Fok, 1979; Fry, 1947) and consonant frequencies (Crawford and Wang, 1960; Mines et al., 1978) differ greatly, it is possible that expectations about phoneme frequency are diminished in a laboratory setting where participants merely choose between two particular phonemes. We thus proceed with this simplifying assumption in our simulations.

The resulting expression for identification, using probability distributions from the model, is

$$p(c_1|S) = \frac{N(\mu_{c1}, \sigma_{c1}^2 + \sigma_S^2)(0.5)}{(0.5)N(\mu_{c1}, \sigma_{c1}^2 + \sigma_S^2) + (0.5)N(\mu_{c2}, \sigma_{c2}^2 + \sigma_S^2)} \tag{3}$$

The values that appear in this equation are: μ_{c1}, μ_{c2}, $\sigma_{c1}^2 + \sigma_S^2$, and $\sigma_{c2}^2 + \sigma_S^2$. It is these values that we would be able to recover via a fit of our model to the behavioral data produced by the listener. The Simulations section below shows how this fits into the overall process of extracting parameters via model fitting, and how these parameters guide our understanding of the gradient effects of categoricity.

Equation 3 can be used to derive the equation that we will fit to the data. In the original application of this model to vowel data in Feldman et al. (2009), there was a simplifying assumption that underlying category variances for the two categories c_1 and c_2 were equal. This meant that only one sum of variances would need to be considered, $\sigma_c^2 + \sigma_S^2$. However, this is an inaccurate assumption for

Table 2 Bayes' rule in the identification and discrimination tasks

	Identification	Discrimination
Data	Speech sound, S	Speech sound, S
Hypotheses	Category, c	Target production, T
Likelihood	Distribution of sounds in a category, $p(S\|c)$	Noise process, $p(S\|T)$
Prior	Probability of choosing the category, $p(c)$	Phonetic category structure, $p(T\|c)$

stop consonants, because voiced and voiceless stop consonants have substantial differences in their variances along the voice onset time (VOT) dimension (Lisker and Abramson, 1964a). Because of this, we extend the model to allow for different variances for the two categories. The expression for the probability of the sound having been generated from category 1 given the perceived stimulus is

$$p(c_1|S) = \frac{1}{1+\sqrt{\frac{\sigma_1^2}{\sigma_2^2}} \times exp\frac{(\sigma_2^2-\sigma_1^2)S^2+2(\mu_{c_2}\sigma_1^2-\mu_{c_1}\sigma_2^2)S+(\mu_{c_1}^2\sigma_2^2-\mu_{c_2}^2\sigma_1^2)}{2\sigma_1^2\sigma_2^2}} \tag{4}$$

where $\sigma_1^2 = \sigma_{c_1}^2 + \sigma_S^2$ and $\sigma_2^2 = \sigma_{c_2}^2 + \sigma_S^2$. A full derivation of this identification function is given in Appendix A.

In the simulations below, the optimal fit of this model to behavioral identification data was found by computing an error function between this model and the behavioral data and then running an error minimization routine in MATLAB to find the best-fitting parameters.

The identification portion of this model is compatible with several previous models that focus on identification behavior, including NAPP (Nearey & Hogan, 1986) and HICAT (Smits, 2001), which assume listeners are performing posterior inference on Gaussian categories. However, these previous models do not contain a variable analogous to a target production T, which in our model denotes the continuous phonetic detail that listeners aim to recover in discrimination tasks. It is this discrimination model that will be critical in accounting for differences between consonants and vowels in the strength of categorical effects.

Bayes' rule for discrimination

Next, we consider the behavioral task of discrimination. Previous models of categorical perception have assumed that discrimination tasks primarily involved listeners' inferences of category labels (e.g., (Liberman et al., 1957), (Damper & Harnad, 2000)). Our model instead posits that listeners are primarily focused on recovering continuous phonetic detail in discrimination tasks, and recruit their knowledge of categories only because it helps them solve this inference problem. As we show below, inference of continuous phonetic detail is predicted to give rise to a pattern that has many of the same properties as categorical perception. Specifically, listeners perceiving sounds through a noisy channel are predicted to bias their perception toward peaks in their prior distribution over sounds. Because sounds occur most often near category centers, and less often near category edges, this results in a perceptual bias toward category centers.

For the listener, the task of discrimination involves inferring the most likely value of the target production, T, for a pair of stimuli and then comparing the values to see if the intended target productions were the same or different (Table 2). The further apart the pair of Ts are judged to be, the higher the probability that the listener will decide that the stimuli are different.

Given the values of the likelihood and prior probability distribution from Table 2, we can calculate the posterior probability $p(T|S)$, which the listener needs to infer in order to perform the discrimination task during behavioral trials. Because the target production, T, could have derived from either underlying category, we can express the posterior distribution as a weighted sum over the two categories. The posterior has the form

$$p(T|S) = \sum_c p(T|S,c)p(c|S) \tag{5}$$

The first term is the posterior distribution on target productions, given that a sound came from a specific category c. The weighting term is the probability of the category being the underlying one chosen given the speech sound heard, $p(c|S)$, which was computed above in the identification section (4).

We can compute the posterior for a specific category, $p(T|S,c)$, using the values introduced above. Bayes' rule for discrimination is

$$p(T|S,c) = \frac{p(S|T)p(T|c)}{\int_T p(S|T)p(T|c)} \tag{6}$$

The summation term from Eq. 1 has been replaced with an integral term because, unlike the category variable in the identification task, the target production is a continuous variable. If we rewrite using the probability distributions from the generative model, we can see the critical parameters that can be recovered via model fitting. The expression is

$$p(T|S,c) = \frac{N(T,\sigma_S^2)N(\mu_c,\sigma_c^2)}{\int_T N(T,\sigma_S^2)N(\mu_c,\sigma_c^2)} \tag{7}$$

Plugging this back into Eq. 5 and expanding the summation term yields

$$p(T|S) = \frac{p(c_1|S)N(T,\sigma_S^2)N(\mu_{c1},\sigma_{c1}^2)}{\int_T N(T,\sigma_S^2)N(\mu_{c1},\sigma_{c1}^2)} + \frac{p(c_2|S)N(T,\sigma_S^2)N(\mu_{c2},\sigma_{c2}^2)}{\int_T N(T,\sigma_S^2)N(\mu_{c2},\sigma_{c2}^2)} \tag{8}$$

The values that appear in this equation are: (1) $p(c_1|S)$, (2) $p(c_2|S)$, (3) T, (4) σ_S^2, (5) σ_{c1}^2, and (6) σ_{c2}^2. Of these values, the noise variance, σ_S^2, is the only one that needs to be fit in our simulations of discrimination data. This is because the first two, $p(c_1|S)$ and $p(c_2|S)$, are known from the identification part of the model fitting, T is the value being calculated, and the last two, σ_{c1}^2 and σ_{c2}^2, can be calculated by subtracting σ_S^2 from the two sums of variance terms

inferred in the identification stage above. We also need a ratio constant, which we refer to as K, to relate the model discrimination predictions to the discriminability metrics used in the behavioral experiments. This additional parameter is merely a constant term that stretches the range of the model values to the range of behavioral findings, and does not provide critical information about the structure of the problem. Hence, we have two free parameters in the model that are estimated via fitting the behavioral discrimination data: σ_S^2 and K.

Equation 8 can be used to derive the model equation that we can fit to the discrimination data, calculating the optimal value under the posterior distribution for the target production given the speech sound. Specifically, we compute the mean value of the posterior distribution. The posterior distribution is a mixture of Gaussians obtained from the Gaussian prior, $p(T|c)$, and the Gaussian likelihood, $p(S|T)$, via Bayes' rule. The basic equation can be seen in Eq. 9, which holds for arbitrary numbers of categories. The expanded form for the case of two categories with different category variances is given in Eq. 10. The full derivation can be found in Appendix B.

$$E[T|S] = \sum_c p(c|S)\frac{\sigma_c^2 S + \sigma_S^2 \mu_c}{\sigma_c^2 + \sigma_S^2} \tag{9}$$

$$E[T|S] = p(c_1|S)\frac{\sigma_{c1}^2 S + \sigma_S^2 \mu_{c1}}{\sigma_{c1}^2 + \sigma_S^2} + p(c_2|S)\frac{\sigma_{c2}^2 S + \sigma_S^2 \mu_{c2}}{\sigma_{c2}^2 + \sigma_S^2} \tag{10}$$

The expected value for the target production is a weighted average of contributions from each possible underlying category. The contribution from each category is itself a linear combination of the speech sounds, S, and the underlying category mean, μ_c.

As with the identification model, the optimal fit of this model to the behavioral discrimination data was found by computing an error function between this model and the behavioral discrimination data and then running an error minimization routine in MATLAB to find the best-fitting parameters.

Degrees of warping: a critical ratio

In Eq. 9 (and expanded upon in Eq. 10), the acoustic value of S is multiplied by the meaningful category variance, σ_c^2. This means that a higher meaningful variance term yields a greater contribution from S in the inferred target production. Meanwhile, the acoustic value for the category mean, μ_c, is multiplied by the noise variance term, σ_S^2. This means that a higher noise variance term leads to a greater contribution from μ_c, the category mean, in the inferred target production. As a result, varying the ratio between these two variance terms leads to varying influences from the category mean and the speech sound, effectively controlling categorical effects. It is this ratio of the meaningful to noise variance that we show to correspond with the degree of categorical effects in phoneme perception. When fitting the behavioral identification data, we are able to extract the sum of the two variances. Then, after fitting the behavioral discrimination curves, we obtain the independent contribution of these two parameters. We call this ratio of variances τ and show that the τ values for different phonemes fall on a continuum that corresponds to the degree of categorical effects, giving us a parametric explanation of the differences within the same model. We do not claim that this measure for τ is explicitly represented or associated with different phonemes, but instead use it as a notational convenience that represents the degree to which the listener is attending to the acoustics of the speech signal based on the two variances associated with the phoneme or phonetic category. An appealing aspect of the τ statistic is that it is dimensionless, which allows us to compare phonemes that fall on continua defined over different acoustic dimensions.

We can gain insight into the continuum of τ values by looking at the extremes. As τ approaches zero, either the meaningful variance of a category approaches zero, or the noise variance grows very large, and in either case listeners have to depend entirely on their existing knowledge of categories. In this case, the entire judgment will be determined by the means of the underlying categories. Perception would look extremely categorical, as the listener discards any contribution of fine acoustic detail and instead uses purely the category mean. At the other extreme, as the ratio approaches infinity, either the meaningful category variance grows large or the noise variance goes to zero. In both cases, the contribution of the underlying category means shrinks to nothing and perception is guided purely by the details in the speech stimulus, S. This means that perception would be entirely continuous and veridical to the acoustic signal. Overall, this relationship represents the degree to which perception is biased by the effect of category membership, and can account for the gradient effects of categoricity we observe in various behavioral tasks.

Figure 3 illustrates the degree of warping along a given continuum with the produced acoustic values of the individual stimuli on the top and the perceived values on the bottom. This process provides an appealing visual perspective of the degree of warping for any given value of τ, showing what happens as we move from a situation with no noise (ratio of infinity) to a condition where there is ten times more noise than meaningful variance (ratio of 0.1). Along with the warping of actual to perceived stimuli, each chart is overlaid on top of the categorical variances that

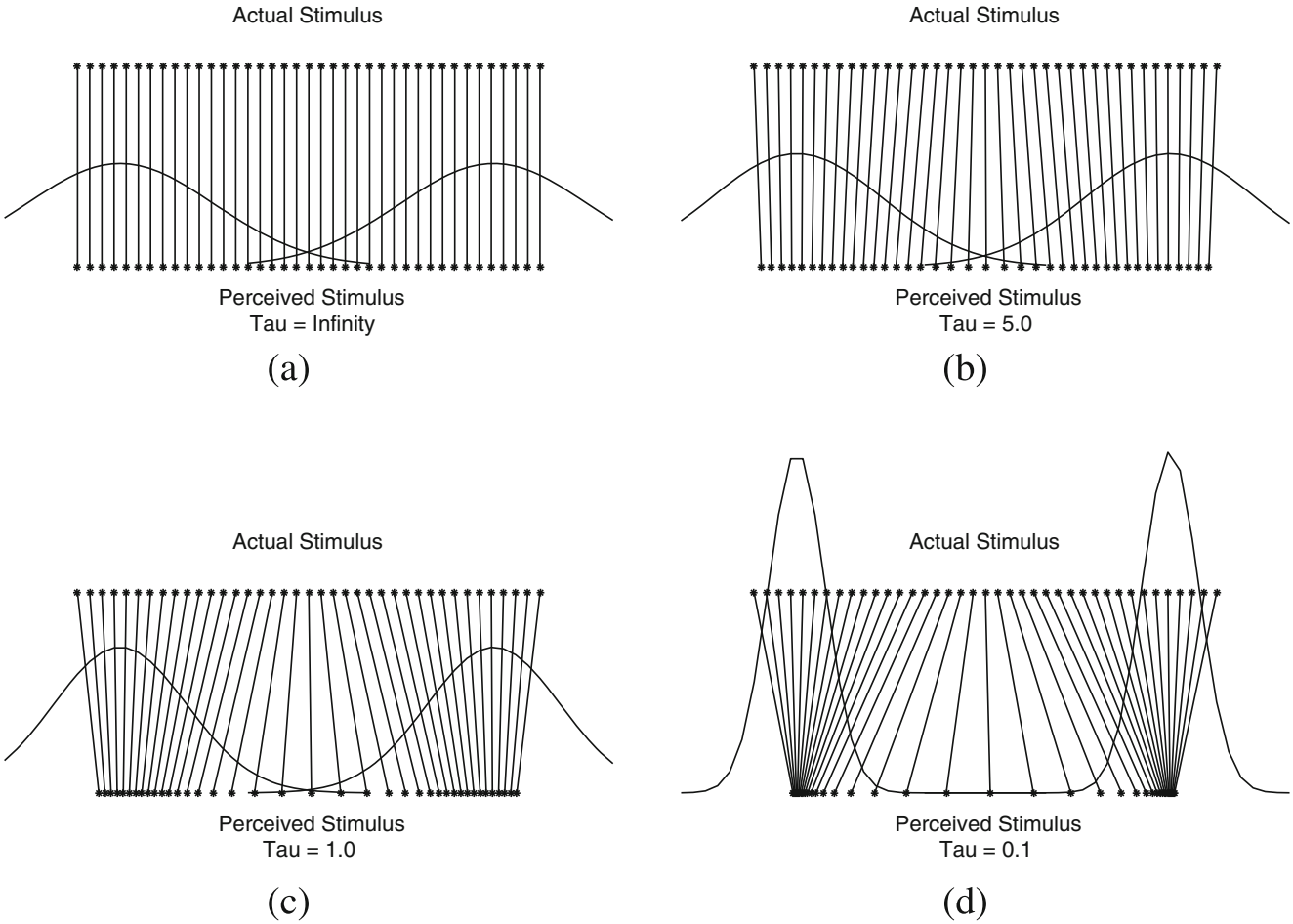

Fig. 3 These simulations illustrate the effect of varying the ratio of meaningful to noise variance. Warping from actual to perceived stimuli is shown in the dispersion of the *vertical bars* toward category centers. Total variance is held constant throughout the simulations, with the amount of variance attributed to underlying category variance shown in the two Gaussian distributions overlaid over the perceptual warping bars. Ratios presented include: **a** infinity, **b** 5.0, **c** 1.0, and **d** 0.1

make up the meaningful variance component of the overall variance (we hold the sum of the meaningful and noise variances constant in these simulations).

In Fig. 3a, with no noise, all stimuli are perceived veridically and there is no warping of the signal. This is the extreme case with an infinite ratio, but is instructive since this is what would happen if we had full faith in the acoustic signal. In terms of our model, this means that the contribution of the perceived speech stimulus, S, completely trumps the mean of the underlying category, μ_c. This graph also serves as a good reference point, with the Gaussian distribution in this case the same as the sum of the two variances in all other graphs (i.e. since there is no noise, the entire variance is due to the meaningful variance). We then see what happens in cases where more and more of the variance is attributed to perceptual noise, thereby shrinking the meaningful:noise variance ratio.

As we move down on the τ continuum, the relative size of the meaningful category variance gets smaller. Holding the sum of the variances constant, greater noise variance means a smaller meaningful variance. Consequently, we see that the Gaussians in the graphs get narrower around the category centers as we reduce the ratio. More importantly, with a decreasing ratio we see greater warping as the individual stimuli get pulled more and more strongly toward the centers of the categories. The contribution of the perceived speech stimulus, S, is decreased compared to the contribution of the mean of the underlying category, μ_c. To motivate why this might happen, we can consider the explanation above in terms of perception in a noisy channel. The meaningful variance going down means that there is a greater contribution from the noise variance. Listeners cannot rely on the information coming in through the noisy channel. Instead, they must rely on their prior knowledge of categories. As a result, the perceived stimuli get pulled into the centers of the categories, of whose structure the listener has prior knowledge. At the last step, with a ratio of 0.1, many of the stimuli are almost completely pulled in by the category center.

Table 3 Steps used to fit parameters in simulations comparing the model to behavioral data from vowels, stop consonants, and fricatives

Step	Description	Derived Parameters
1	Set μ_{c1} on the basis of production data	μ_{c1}
2	Determine μ_{c2}, σ_1^2, and σ_2^2 from identification data using Eq. 4	μ_{c2}, σ_1^2, σ_2^2, giving us the full category structure of the perceived categories
3	Determine the ratio of the meaningful category variances, σ_{c1}^2 and σ_{c2}^2, to noise variance, σ_S^2, by fitting acoustic differences between percepts, $E[T\|S]$, in the model (9) to a distance measure such as d'	σ_S^2 is the only parameter fit by this simulation step, which when subtracted from σ_1^2 and σ_2^2 also gives us σ_{c1}^2 and σ_{c2}^2. After this step, we have the category structure of the underlying categories
4	Compute τ from the meaningful category and noise variances and examine where they fall on the continuum presented above	τ, giving us the degree of warping along our derived continuum

$\sigma_1^2 = \sigma_{c1}^2 + \sigma_S^2$ and $\sigma_2^2 = \sigma_{c2}^2 + \sigma_S^2$

In the remainder of the paper we examine whether the phonemes that we have been considering thus far can be interpreted as falling at different points along this continuum of categoricity. We fit the model to each type of phoneme's identification data and discrimination data to get the relevant ratio of meaningful to noise variance. We then compare these ratios to the warping continuum that we just described. If the phonemes map onto the τ continuum in a way that correlates with the respective behavioral findings, this would indicate that the model captures the range of categorical effects, which had previously been described independently, via parametric variation within a single framework. If all of these effects can be captured by the same model, it may not be necessary to appeal to two independent effects to describe categorical effects for vowels and consonants. The categorical effects would be interpretable as instantiations of the same phenomenon.

Simulations

In this section we describe simulations with vowels, stop consonants, and fricatives. First, we show that the model provides a good fit to the behavioral data with an appropriate setting of the parameters. We evaluate this by examining the fit of the model to both the identification and discrimination curves from behavioral experiments. Second, we show that the derived parameters are precisely the type that yield the proposed single qualitative source of categorical effects proposed above. Particularly, we examine the derived τ values and where they fall on the continuum. All of our simulations follow the procedure that is shown in Table 3 and described in more detail below.

Modeling steps

Setting a category mean

First we set one of the category means. This is necessary because the model is otherwise underspecified; an infinite number of category means provide the same fit to the identification data (i.e., the parameters are not identifiable). This is because fitting the model derives the equivalent of the sum of the category means, so moving one mean up would just move the other one down. Since we need to have a specific set of means in order to get the proper variance ratio in the following steps, we fix one of the means. In our simulations we set the mean based on production data from native speakers. We choose which of the two means to set arbitrarily, since the simulation is symmetric and it does not matter which mean we set. We are able to verify that the mean that is found for the second category is reasonable based on production data, since the mean for the underlying and production categories is the same, with the only difference being the variance. Further, it should be noted that we can't set one of the variances in order to further simplify the calculation since there is no behavioral data that corresponds to the underlying category. Production data serves as a close approximation to the underlying category since it avoids the perceptual noise, but it still includes articulatory noise, leaving us only able to set a category mean and deriving the rest.

Identification fitting

To fit the model's identification predictions to the behavioral data, we use Eq. 4. We compute the mean squared error between this equation and the actual values from behavioral identification data to find the set of parameters σ_1^2, σ_2^2, μ_{c1}, and μ_{c2} that give the optimal fit. In effect, this recovers the sum of the variances that the listener in our generative model assumes, leading to the structure of the Gaussian distributions for $p(S|c) = N(\mu_c, \sigma_c^2+\sigma_S^2)$. This is the structure of the perceived category, as opposed to the underlying category.

Discrimination fitting

At this point we have the means and the sums of variances, so we need to pull apart the contribution of the meaningful category variance and the articulatory and perceptual noise variance to the overall variance derived in the previous step.

In effect, σ_S^2 is the only free parameter that is inferred in this step. Separating out the noise variance serves a dual purpose. First, it allows us to calculate the ratio that we show to correspond with degree of categorical effects. Second, we get an independent test of the model's ability to fit behavioral data. Having separated the two sources of variance, we can also see the shape of the underlying category distributions that are known to the listener in our generative model and can see if the finding is reasonable by comparing this against production data, which as mentioned above serves as a rough approximation to the underlying category with meaningful variance since it removes the perceptual part of the noise variance. This sanity check is especially relevant in the case of stop consonants, where the distributions of the phonemes at the two ends of the continuum exhibit very different variances that may be reflected in those categories' perceptual signatures.[1]

Calculate the variance ratio τ

As the final step of the simulation process, we compute the value of τ for each phoneme by dividing the meaningful category variance, σ_c^2, by the noise variance, σ_S^2. τ quantifies the strength of the pull toward category centers. We find an independent meaningful category variance for each phoneme, so for each phoneme category we actually find two τ values; τ_{c1} and τ_{c2}. These τ values characterize the degree of perceptual bias toward category centers for either of the phonemes at the ends of the fricative, stop consonant, and vowel continua. In effect, then, they characterize the warping for an idealized continuum if it were to consist of two identical categories at its ends. Looking at the individual warping parameters allows our model to capture varying category structure and varying within-category discrimination, even within a single continuum. This is different from previous investigations of categorical perception, which typically looked across a whole continuum, not at individual phonemes on that continuum. Here, we do not quantify the amount of warping on the stop consonant continuum, but rather for $/b/$ and $/p/$ independently, and likewise for /s/ and $/\int/$ independently. In our simulations, the τ values end up being very close for the phonemes at the two ends of the continuum for all phonetic categories except stop consonants, for which the within category variance for the voiced stop consonant /b/ is much lower than that for /p/.

[1]Another way of fitting model parameters would be to jointly condition parameter inference on the identification and discrimination data. However, we believe that our fitting procedure is more appropriate because it provides an implicit test of the parameters when moving from identification to discrimination fitting. Examining the fit of the model to discrimination data (with parameters derived from identification data) helps to confirm that we are finding an appropriate set of parameters.

Table 4 Formant values for stimuli used in the multidimensional scaling experiment, as reported in Iverson and Kuhl (2000)

Stimulus Number	F_1 (Hz)	F_2 (Hz)
1	197	2489
2	215	2438
3	233	2388
4	251	2339
5	270	2290
6	289	2242
7	308	2195
8	327	2148
9	347	2102
10	367	2057
11	387	2012
12	408	1968
13	429	1925

Vowels

The simulations we consider for vowel perception were conducted by Feldman et al. (2009), who examined the perceptual magnet effect as a case study in categorical effects in cognition. Their model made the simplifying assumption that the two categories that define the continuum endpoints have equal variance, but was otherwise identical to the model described above. We use the parameters of their simulation to compute a ratio of variances along the τ continuum for vowels and use this value as a basis for comparison with stop consonants and fricatives.

Feldman et al. (2009) used their model to simulate data from a paper by Iverson and Kuhl (1995) that used a multidimensional scaling technique to examine discrimination performance by participants on the /i/-/e/ continuum. The formant values for the continuum are reproduced in Table 4 below, as reported by Iverson and Kuhl (2000). Although the parameters here are given in Hertz, the continuum was based on equal sized steps in Mels, a psychoacoustic frequency scale (Stevens et al., 1937). While d' data were available for the same stimuli, the multidimensional scaling data represented the entire range of stimuli and therefore provided a broader range upon which the model could be fit. Identification data were taken from Lotto et al. (1998), who pointed out that there was a discrepancy in how the identification and discrimination data were collected by Iverson and Kuhl (1995). Whereas the stimuli in discrimination trials were presented in pairs, stimuli in the identification trials were presented in isolation. Because of known effects of context on perception, this meant that the category of the same stimulus might be perceived differently in the two experiments. To circumvent this issue, Lotto et al. (1998), repeated the identification experiment from Iverson and Kuhl (1995), but

Table 5 VOT for stimuli from the /b/-/p/ continuum, along with data on listeners' identification and discrimination of these stimuli

Stimulus (S_i) #	VOT (ms)	% Identification as /p/	Hits	False Alarms	d'
S_0	−50	0.00			
S_1	−40	0.00	0.33	0.25	0.24
S_2	−30	0.00	0.30	0.25	0.17
S_3	−20	0.02	0.30	0.30	0.01
S_4	−10	0.01	0.39	0.31	0.20
S_5	0	0.01	0.54	0.36	0.46
S_6	10	0.03	0.73	0.37	0.93
S_7	20	0.30	0.95	0.37	1.98
S_8	30	0.95	0.94	0.34	1.94
S_9	40	1.00	0.64	0.30	0.87
S_{10}	50	1.00	0.47	0.25	0.60
S_{11}	60	1.00	0.42	0.22	0.57
S_{12}	70	1.00			

Identification and discrimination data are obtained from Figures 1 and 5 in Wood et al. (1976)

presented the stimuli for both experiments in pairs. Using the identification data from Lotto et al. (1998) to fit the category means and the sum of the meaningful and noise variances, Feldman et al. showed that the model provided a close fit to the behavioral findings from Iverson and Kuhl (1995).

Feldman et al. (2009) then conducted an experiment to examine how noise affects discrimination judgments and whether the model captures these effects via the noise variance parameter. Listeners made same-different judgments for pairs of stimuli, modeled after those from Iverson and Kuhl (1995) (Table 4), in either quiet or noisy listening conditions. Feldman et al. generated confusion matrices for the stimuli in the experiment and modeled this confusion data using a variant of their model that predicted same-different judgments directly. In effect, they were computing the probability that the distance between the inferred target productions was less than or equal to ϵ given the two perceived speech stimuli, S_1 and S_2. Since the same contrast was presented a total of n times across all participants during the experiment, the overall confusion of the two stimuli is measured by the binomial distribution $B(n, p)$, where p

Table 6 Central Frication Frequencies (in Barks and Hertz, labeled as F_5 and F_6) for stimuli used in the behavioral experiments by Lago et al. (2015)

Stimulus (S_i)	Barks		Hertz		% Ident. of S_i as /s/	$[S_{i-1}, S_{i+1}]$ Disc. d'
	F_5	F_6	F_5	F_6		
S_0	14.5	15.5	2501	2915	0.02	
S_1	15.0	16.0	2698	3152	0.03	0.15
S_2	15.5	16.5	2915	3413	0.00	0.84
S_3	16.0	17.0	3152	3702	0.03	0.63
S_4	16.5	17.5	3413	4025	0.06	1.40
S_5	17.0	18.0	3702	4386	0.13	1.08
S_6	17.5	18.5	4025	4794	0.53	1.50
S_7	18.0	19.0	4386	5258	0.83	1.27
S_8	18.5	19.5	4794	5790	0.99	1.35
S_9	19.0	20.0	5258	6407	0.97	0.91
S_{10}	19.5	20.5	5790	7131	0.99	

Identification study results shown as percent of time stimulus identified as the phoneme /s/. Discrimination results are d' scores for two step discrimination, where the score on line i is the discrimination score for the pair of stimuli on lines i-1 and i+1 (e.g. 0.63 on line S_3 is discrimination score for S_2 and S_4)

is the probability mentioned above and presented below in Eq. 11.

$$p(|T_1 - T_2| \leq \epsilon | S_1, S_2) \tag{11}$$

The role of this parameter in the extended model is similar to that of the observer response criterion in signal detection theory (Green & Swets, 1966), where ϵ determined the size of the judged distance between stimuli necessary to yield a positive response, in their case a response of *different*.

In order to minimize free parameters in the model, they held constant all other parameters. To do this, they used the category means, $\mu_{/i/}$ and $\mu_{/e/}$, as well as the categorical variance, σ_c^2, from the multidimensional scaling simulations. They found that the model was able to find close fits to both conditions and, more importantly, that the noise parameter was independently a good predictor above and beyond the setting for the threshold parameter, ϵ. For our current experiment we are not interested in perception in the presence of additional artificial noise, so we use the noise variance term from the no-noise condition, which was found to be $\sigma_S^2 = 878$ ($\sigma_S = 30$ mels), as the appropriate noise variance to capture discrimination performance for the /i/-/e/ continuum.

The parameters obtained in this discrimination study suggested that the noise parameters derived from fitting the multidimensional scaling analysis were inaccurate, due to skewing introduced via the multi-dimensional scaling procedure. We therefore focus our analysis on the the parameters derived directly from same-different judgments in their discrimination study, and avoid using multidimensional scaling data for the analyses in our simulations below.

The full set of parameters inferred through fitting the model can be found in Table 7 in the row for vowels. With the values for the meaningful and noise variance set to $\sigma_c^2 = 5,873$ and $\sigma_S^2 = 878$, we derive the critical ratio for vowels $\tau_V = \frac{5,873}{878} \approx 6.69$. The corresponding graphical warping picture for this τ value is shown in Fig. 8a together with analogous graphs for other phoneme categories. There is very little warping between the actual and perceived stimuli. There is, however, a small effect of categories, with stimuli closer to the categorical centers pulled together and slightly greater distances at the category boundary as the stimuli are pulled apart. This value and warping picture serve as baselines to consider behavioral data for stop consonants and fricatives.

Stop consonants

Stop consonants have been found to exhibit very strong categorical effects in perception during behavioral experiments. In our model, this would be formalized as a low meaningful to noise variance ratio. As such, in the presence of noise, listeners would rely much more on their knowledge of underlying categories rather than the detail available in the acoustic stream. However, even if the relative contribution of category means and acoustic detail does contribute to consonant perception, factors such as innate phonetic boundaries (Eimas et al., 1971) and auditory discontinuities (Pisoni, 1977) may continue to play a role. If they exert an additional effect on stop consonant perception above and beyond the model we have proposed here, this would prevent our model from fully explaining the behavioral findings. Because of this, a good fit to the behavioral data would be a particularly strong argument in favor of positing a unified account for categorical effects in phoneme perception.

For stop consonants we consider both identification and discrimination data derived from Wood (1976). Their experiments focused on the perception of /b/ and /p/ along a voice onset time (VOT) continuum. Their stimuli were synthetically created along the continuum ranging from −50 to +70 ms VOT. Their identification task was a classic two-alternative forced choice task. For discrimination, they administered both a 10-ms and 20-ms difference AX discrimination task, in which participants heard one stimulus, A, and then had to decide whether the second stimulus, X, was the same or different as the first. For our simulations below, we used data from their 20-ms discrimination condition. Values for both identification and discrimination can be found in Table 5. We use d-prime as the measure of perceptual distance, which we computed from hit and false alarm values reported in Fig. 5 in their paper.[2]

As a first step, we need to set the mean for one of the categories in our simulation. Based on production data in Lisker and Abramson (1964a), we set $\mu_{/p/}$ at 60 ms VOT. Once again, the choice of which mean to set is arbitrary so there is no deeper reason for setting one mean over the other. We then ran the error minimization procedure in Matlab in order to determine the optimal fit of the free parameters involved in the identification simulation: $\mu_{/b/} = -0.3\ ms$, $\sigma_{/b/}^2 + \sigma_S^2 = 96.3$, and $\sigma_{/p/}^2 + \sigma_S^2 = 336.2$ (Table 7).

The fit of the model to identification data can be seen in Fig. 4a. The figure also shows the category structure of the perceived categories, reflecting the two means as well as sums of variances for each category. The value of -0.3 ms that the model inferred as the /b/ category mean is very close to that found in the production data from Lisker and Abramson (1964a) (Fig. 5).

[2]The central measure of discriminability that Wood (1976) reported was $-ln(\eta)$, which is monotonically related to the d-prime parameter. They also included a unit square graph that showed the relationship between hit rates and false alarm rates. By measuring these distances by hand we were able to recover the values needed to compute d-prime scores. We then used these d-prime scores in our simulations to keep the measures of perceptual distance consistent with that used for fricatives.

Table 7 Best fitting model parameters for vowels (Feldman et al., 2009), fricatives and stops consonants

Simulation	Means		Variances			Meaningful:Noise Variance Ratio (τ)
	μ_{c_1}	μ_{c_2}	$\sigma^2_{c_1}$	$\sigma^2_{c_2}$	σ^2_S	
Vowels	$F_1 = 423$ Hz	$F_1 = 224$ Hz	5,873	5,873	878	6.69
(Equal Variance)	$F_2 = 1936$ Hz	$F_2 = 2413$ Hz		(Mels)		
Stop Consonants	−0.3 ms	60 ms	16.3	256.2	80.0	/b/: 0.20 , /p/: 3.20
(Unequal Variance)						
Fricatives	15.99 Barks	19.0 Barks	0.599	0.575	0.310	/∫/: 1.85, /s/: 1.93
(Unequal Variance)						

Fitting the discrimination data we get the following values for the individual variances in our model: $\sigma^2_{/b/} = 16.3$, $\sigma^2_{/p/} = 256.2$, and $\sigma^2_S = 80.0$ The fit of the model to the discrimination data can be seen in Fig. 4b. In addition to providing a good overall fit, the model is able to accurately predict the lower within-category discriminability of voiced stops relative to voiceless stops. This can be seen in Fig. 4b, where the left side of the distribution is substantially lower than the tail on the right.

The underlying categories can be seen, overlaid over the identification data and perceived categories, in Fig. 4a. The distributions for the voiced and voiceless stop consonants were found to be very different, with a very narrow /b/ category and a shallow diffuse /p/ category. While there is no way to independently confirm how well this reflects actual category structure in the brain, we have an approximation in existing production data. As mentioned earlier, production data is closer to the underlying than the perceived categories since it removes perceptual noise from the equation, giving us a measure or the underlying categories with some articulatory noise introduced, but without perceptual noise. Hence, to see if these very different category structures make sense, we consider how well the underlying categories correspond to those in speech production. Figure 5 shows the model categories reflecting the inferred parameters for $\mu_{/b/}$, $\mu_{/p/}$, $\sigma^2_{/b/}$, and $\sigma^2_{/p/}$, together with the the distributions found in production studies by Lisker and Abramson (1964a). The distributions in the model are almost perfectly aligned with the distributions in the data. This suggests that the model was able to accurately infer the categories that lead to the behavioral data.

Moving on to an analysis of the degree of categorical effects, based on the parameters found for stop consonants, we obtain the following critical τ ratios:

- $\tau_{/p/} = 3.20$
- $\tau_{/b/} = 0.20$

As would be expected based on previous findings, the critical variance ratio for stop consonants is found to be substantially lower than that of vowels for both the voiced and voiceless categories. In terms of the model, this means that stop consonants have less meaningful within-category variance and more noise in the signal. This leads the listener to rely to a greater degree to underlying knowledge of categories, leading to greater pull toward category centers, and in turn to what we know as categorical perception. This perceptual bias is especially pronounced for the voiced stop consonant, /b/, due to the low variance in the underlying category.

The corresponding warping views for these values can be seen in context of other phonetic categories in Fig. 8b for parameters corresponding to the /p/ category and Fig. 8c for parameters corresponding to the /b/ category. Although the figures display two categories at either end of the continuum with equal variances, these are not meant to correspond to actual phonetic categories. Instead, they illustrate the way in which the τ ratio for each individual phoneme would warp perceptual space along a hypothetical symmetric continuum. These figures are meant as illustrations of strength of the particular categorical effects, rather than a veridical representation of the warping along any particular continuum. We can see from these figures just how pronounced this pull toward category centers is, especially compared to that of vowels.

Fricatives

Behavioral findings in research on categorical effects in fricative perception have varied, with some studies finding effects that are close to the strong categorical effects of stop consonants, while others find more continuous perception closer to that of vowels. This suggests that task factors may affect the processing mode or the attribution of variance to noise or meaningful variance. On the one hand, this may mean that it is impossible to posit a single ratio for any phoneme in the fricative class. On the other hand, we might expect to find that the relationship between meaningful category variance and noise variance is somewhere in between that of stop consonants and vowels. We select data from a task and a measure that are similar to those

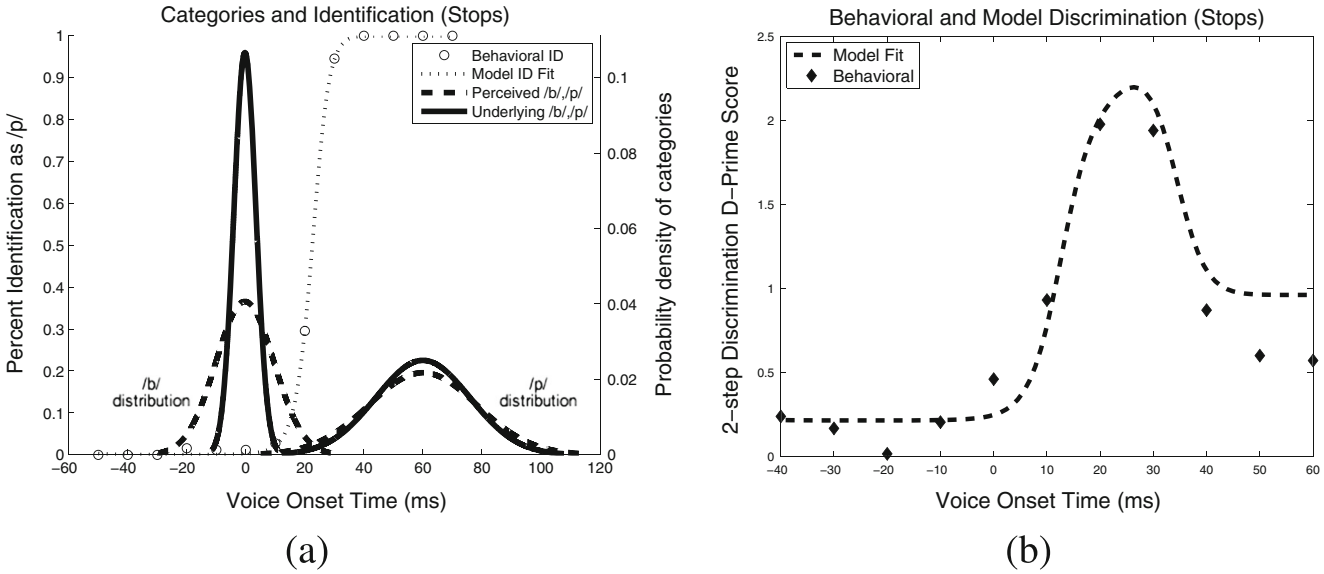

Fig. 4 Model fit to identification and discrimination data for stop consonants: **a** identification function, perceived categories, and underlying categories for stop consonant simulations along the /b/-/p/ continuum, **b** behavioral d-prime scores for discrimination along with optimal model fit to the data

from stop consonant studies, and ask how the derived ratios line up with those found in the previous simulations. We have already seen that the model can both fit the behavioral data and provide meaningful parameters for stop consonants and vowels. If we find that it can do the same for fricatives, then this would further confirm that the model can reliably explain behavioral data for a range of phonemes. This would, in turn, provide additional evidence that we do not need to posit qualitative differences across phoneme classes in the computations involved in categorical effects in phoneme perception.

Our simulations used behavioral data from Lago et al. (2015) for fricative identification and discrimination. Lago et al. used stimuli along a continuum ranging between the the sibilant fricatives /s/ and /∫/. Both phonemes are primarily identified by their two central frication frequencies. Shifting these from one to the other creates a fricative continuum that is analogous to a voice onset time continuum in stop consonants or a formant continuum in vowels.

The /s/-/∫/ continuum was created by varying the two central frication frequencies of the noise portion of a consonant-vowel syllable with a standard /a/ vowel. This continuum relates to the shift in place of articulation moving from alveolar to post-alveolar. The stimuli were artificially created based on human productions, by employing a procedure similar to McQueen et al. (2009). Utilizing the Klattworks interface (McMurray, 2009), frication, aspiration, voicing, formant bandwidths, and amplitues were set by hand and smooth logistics were applied to create the

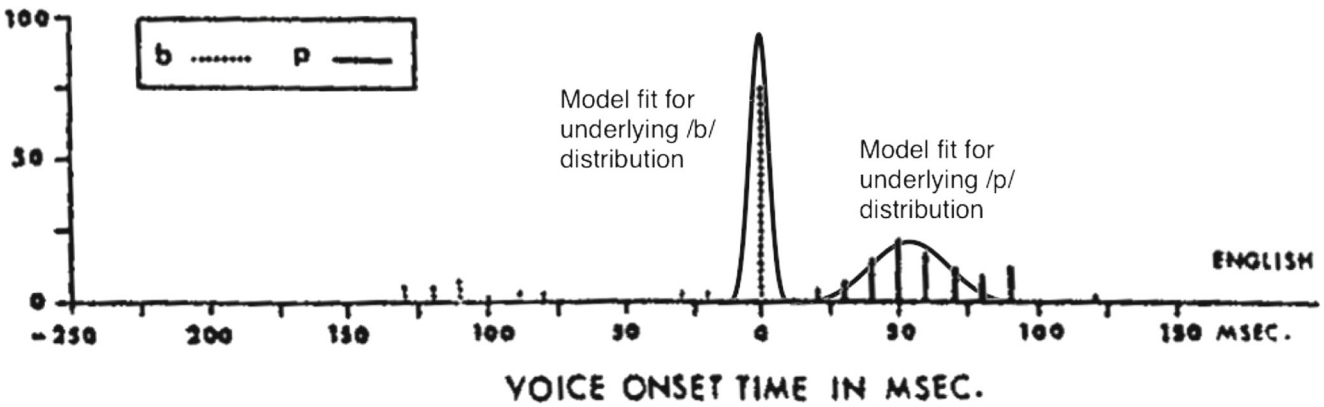

Fig. 5 Production data for the /b/-/p/ continuum (reprinted from Lisker and Abramson (1964a)) overlaid with underlying categories found by the model

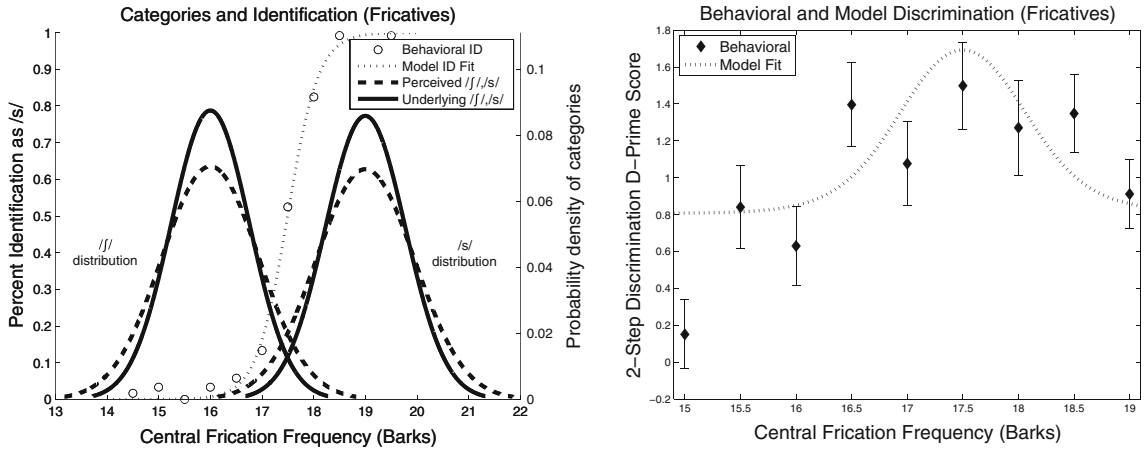

Fig. 6 Model fit to identification and discrimination data for fricatives: **a** identification function, perceived categories, and underlying categories for fricative simulations along the /ʃ/-/s/ continuum, **b** behavioral d-prime scores with standard error intervals representing inter-subject variance along with optimal model fit to the data

initial stimuli. The results were then converted to formats readable by Praat (Boersma and Weenink, 2005), shifting the relevant frication frequencies along the way. The values for the continuum are presented in Table 6 in the first three columns, with values reported in Barks, a psychologically plausible acoustic scale (Zwicker, 1961).

Using these stimuli, Lago et al. (2015) conducted an identification task and an AX discrimination task. The results of these tasks are shown in Table 6 in the final two columns. The results showed strong categorical effects on the perception of the stimuli, though with perhaps a shallower discrimination peak that one might expect, suggesting representation that incorporates some more of the acoustic signal beyond pure category assignment.

As before, we need to set one of the two category means before fitting the behavioral data. Based on productions by an adult male participant, we set the value of $\mu_{/s/}$ to be 19.0 Barks. We then proceeded to fit the first set of parameters via an error minimization procedure in Matlab. The optimal model fit produces the following parameters: $\mu_{/\int/} = 15.99$ Barks, $\sigma^2_{/\int/} + \sigma^2_S = 0.909$, and $\sigma^2_{/s/} + \sigma^2_S = 0.885$ (Table 7). The fit of the model to the identification data can be seen in the identification curve overlaid over the behavioral data points in Fig. 6a. In the same figure we also see the perceived categories, representing the parameters for the means of the categories as well as the sums of meaningful and noise variance.

Fitting the discrimination data we obtain the following value for the individual variances in our model: $\sigma^2_{/\int/} = 0.599$, $\sigma^2_{/s/} = 0.575$, and $\sigma^2_S = 0.310$ (Table 7). The fit of the model to the discrimination data can be seen in Fig. 6b. Although the fit here is not as close as that of the stop consonants, the peak in discrimination in both the underlying data and the model fit occurs in the same location, and that location corresponds to the inflection point of the identification curve. While our model does not fit every mean of the subjects' d-prime data, it fits the data within the margin of error for almost all data points. It is possible that fitting a separate model to each participant would give us insight into individual variability and the source of this disparity, but the individual data would be sparse, and might lead to unreliable model fits.

Based on the parameters found for fricatives, we obtain the following critical τ ratios:

- $\tau_{/\int/} = 1.85$
- $\tau_{/s/} = 1.93$

As reviewed above, previous studies have been mixed as to the degree of categorical effects for fricatives. Therefore, we might expect values for the ratio somewhere between the extremes for categorical perception and continuous perception. We find this to a limited extent (Fig. 7). The fricatives are much lower on the continuum than are vowels, and are well above the voiced stop consonants, though they are close to the voiceless stop consonant /p/. Based on the model, this means that the listener is depending on the underlying categories more than for vowels, but depending on the finer acoustic detail more than for voiced stop consonants. However, given that the fricatives are lower on the continuum than the voiceless stops and we have only considered one case of stop consonants and fricatives, more research looking at other phonemes is warranted before we say exactly what the relationship is among the phonemes within the cluster of consonants on the continuum. The corresponding

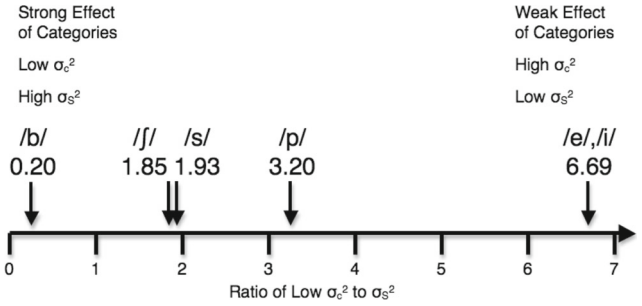

Fig. 7 Fitted τ values for vowels, stop consonants, and fricatives

warping view can be seen in Fig. 8d in the context of other phonetic categories. Given that the two sibilant fricatives yielded nearly identical τ values, only /s/ is pictured as the representative case.

Simulation summary

The simulations in this section have shown that the model provides close empirical fits to behavioral data and also derives ratio values consistent with the continuum of categorical effect sizes. Figures 4 through 6 show that the model is able to provide reasonable fits to identification and discrimination data from multiple experiments: Not only are the means and variances reasonable compared to production data, but further details such as differences in discriminability between voiceless and voiced stop consonants are captured as well. To see whether the model captures categorical effects with parametric variation within a single model, we considered the critical ratio of meaningful to noise variance for each phoneme. Based on previous research, we expected vowels to show a very high ratio, reflecting weak effects of categories with high meaningful variance and low noise variance. Conversely, we expected stop consonants to show a low ratio, reflecting strong effects of categories with low meaningful variance and high noise variance. Finally, we expected fricatives to be somewhere in between, due to mixed previous findings in perception experiments. Findings largely conformed to these predictions, with the single exception being the voiceless stop consonant ratio being higher along the continuum than either of the sibilant fricatives. Figure 7 provides a succinct visual aid to the relevant continuum of ratio values, while Fig. 8 illustrates the relevant warping for these same ratio values.

General discussion

Effects of categories are observed in the perception of various phonemes, but the effects are of different magnitudes. Past research has labeled the case of strong categorical effects of stop consonants as categorical perception and the case of weak categorical effects observed in vowel perception as the perceptual magnet effect. More recent research has demonstrated qualitative similarities across consonant and vowel perception by showing that there is a continuous aspect to consonant perception (Andruski et al., 1994; McMurray et al., 2002; Miller, 1997) and a categorical aspect to vowel perception (Gerrits and Schouten, 2004; Stevens, 1966), but the literature has not provided strong evidence that the differences between them can be accounted for within a unified framework. Our work fills this gap by showing that the same principled mathematical model provides a close match to both identification and discrimination data from both consonants and vowels. We extended a Bayesian model that treats the task of speech perception as a case of optimal statistical inference and showed that it was able to capture the range of categorical effects covering stop consonants, fricatives, and vowels. Not only does our model fit the behavioral identification and discrimination data from the range of phonemes, but it derives the varying degrees of categoricity by varying a single parameter. This statistical parameter, τ, is the ratio of the meaningful category variance (the informative variation of possible target productions) and the interfering noise variance (uninformative variation that masks potentially informative cues in the acoustic stream). An increase in τ corresponds to a greater contribution from meaningful variance, thereby leading the listener to pay more attention to fine acoustic detail, and in turn to more continuous perception. As τ decreases, more of the variance is interfering noise, leading the listener to employ their knowledge of underlying categories more. Correspondingly, we found that strongly categorical stop consonants fall lowest on the continuum (0.20) and much more continuously perceived vowels fall highest on the continuum (6.69), with voiceless stop consonants and fricatives in between. Our results support the idea that this computational model captures the problem being solved by the listener in identification and discrimination of phonemes independent of the specific phoneme class.

Importantly, this is not just a classification model, but a model of continuous discrimination. Many previous models have focused primarily on phoneme classification in identification tasks. In these models, continuous perception is a by-product of the computations necessary for identification. For example, the HICAT model (Smits, 2001) focuses on classification of successive phonemes whose acoustics influence each other, but the only continuous values that are defined in the model are goodness values corresponding to likelihood terms $p(S|c)$ for each category. The NAPP model of Nearey and Hogan (1986) is also focused on identification tasks. Feldman et al. (2009) were the first to introduce the idea that discrimination tasks involve inferring a continuous value T, which is a denoised version of the stimulus

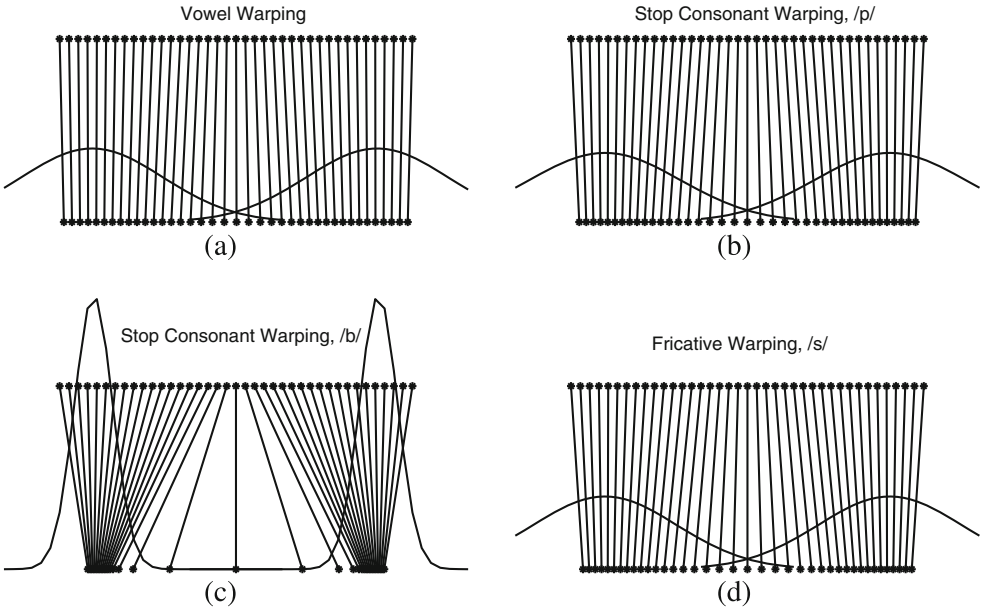

Fig. 8 Warping of perceptual space for phonemes belonging to four classes: **a** vowels, **b** voiceless stop consonants, **c** voiced stop consonants, **d** sibilant fricatives. These figures represent idealized effects that would be predicted along a hypothetical continuum that had two phonemes with identical variances. They are meant as illustrations of strength of the particular categorical effects, rather than a veridical representation of the warping along any particular continuum

S. This focus on recovering continuous detail from the speech signal provides an advantage over previous models that have treated only the classification problem, and is consistent with evidence that listeners perceive acoustic detail (Andruski et al., 1994; Blumstein et al., 2005; Joanisse et al., 2007; Pisoni and Tash, 1974; Toscano et al., 2010). The model explains how the perceptual process leads to effects we see in both identification and discrimination, not just categorical effects in identification.

Ockham's razor and levels of analysis

There is a principle, or heuristic, often employed in scientific arguments called Ockham's Razor that states that all things being equal, the simplest explanation is the right one. In general, models that make fewer assumptions and limit the number of entities involved in the explanation tend to be preferred over more complex models (see Thorburn (1915), Thorburn (1918) for a thorough history and various formulations of this argument). In the spirit of Ockham's Razor we can consider whether it is more likely that there are separate underlying phenomena involved in vowel and consonant perception, or for different phonemes more generally, or that the results are derived from a single source. Our model's close fit to the behavioral results that were the motivating force for proposals such as categorical perception for strong effects of categories for stop consonants and the perceptual magnet effect for mostly continuous vowels makes a unified explanation more parsimonious, without loss of explanatory power. We must consider, however, at what level of analysis this unified explanation holds.

Our model is meant to capture speech perception at the computational level (Marr, 1982). It makes no direct algorithmic or implementation claims, even if the computations involved may suggest certain process-level accounts. This is an important distinction, because stating that a computation involving certain parameters is being solved is not the same as specifying that these parameters are either explicitly represented in memory or derived somehow on the fly from acoustic signals. However, for such a computational approach to guide perception, these parameters do need to be compatible with the underlying algorithm and implementation. As such, our findings do have implications for models at other levels of analysis. In our model, probabilities of sounds in a listener's underlying categories are employed even when there is no explicit identification task being performed.

There are models that have been proposed utilizing exemplars, prototypes, and neural representations that are all compatible with our computational-level account. One such model is Lacerda's (1995) model of categorical effects as emergent from exemplar-based categorization, which has a direct mathematical link to our model, as described by Feldman et al. (2009). Shi et al. (2010) also provide an

implemention of our model in an exemplar-based framework, which is closely related to the neural model proposed by Guenther and Gjaja (1996). In that model, there are more exemplars, or a higher level of neural firing, at category centers than near category boundaries. This occurs even when a percept is closer to the boundary, simulating a partial weighing of the category center on top of the acoustic signal.

There is also a second relevant neural network model that relates Gaussian distributions of speech sounds and neural firing preferences (Vallabha & McClelland, 2007). The perceptual and categorical level had bidirectional links, allowing the category to not only be influenced by the perception, but to in turn influence the perceived sound. This also leads to a bias toward category centers. While we are not suggesting that any of the aforementioned models are the proper representative of how our Bayesian model is implemented, the links between them are worth further investigating to see if together, they can combine to explain the range of categorical effects on perception at all levels of analysis. Critically, however, it is possible that these different models may provide an account of how listeners arrive at the computation described in the present model we put forward.

At a neural level, it is possible that different implementations are indeed playing a role in perception of different phonemes. The nature of the cues being perceived makes this likely, as different classes of phonemes make use of different cues for encoding phonetic distinctions. Particularly, vowels primarily employ static spectral cues such as formant frequencies while consonants primarily employ dynamic spectral and static temporal cues such as formant transitions and voice onset time. Correspondingly, studies have found that consonants and vowels are represented differently in the brain both in terms of physical location as well as the associated patterns. Obleser et al. (2010) found that vowel and consonant categories are discriminated in different regions in the entire superior temporal cortex. While the activation patterns were complex for both stimuli, they were sparsely overlapping and the local patterns were distributed across many regions of the auditory cortex. Perez et al. (2013) found that different coding strategies were used for vowels and consonants. They found vowel discrimination corresponded largely to neural spike counts in certain brain regions, while consonant discrimination correlated to neural spike timing and not to counts if timing information was eliminated. This suggests a difference between temporal or dynamic encoding and static spectral encoding. Therefore, while by a type of Ockham's razor account we would prefer a more parsimonious solution, we leave the question of whether categorical effects in different types of sounds necessarily require different implementations as an open empirical question. On the basis of existing evidence, however, this question seems likely to be answered by an appeal to different implementations.

The meaning of tau

We have shown that a single free parameter τ, which quantifies the relative contributions of meaningful and noise variance in the model, allows our model to capture differences across a range of perceptual data from consonants and vowels. We remain agnostic as to the specific factors that contribute to a phoneme having a particular level of meaningful or noise variance. However, our simulations provide evidence related to this question, as vowels were found to have high meaningful variance and low noise variance, whereas consonant were found to have high noise variance and low meaningful variance. In this section we consider the question of what exactly these mean, why one type of variance might be meaningful while another is not, and what epistemological status they have.

In plain terms, meaningful variance should be predictable. It may be directly related to the mapping of phonetics to phonology, thereby facilitating the use of fine acoustic detail in the retrieval of phonological representations, and it may also serve for extracting indexical information. Noise variance, on the other hand, should represent variability in the signal that is a side-effect of the way a sound is produced and properties of the sound that make it prone to misinterpretation. These meaningful and noise variances both contribute to the overall variability associated with a phoneme. By keeping the target productions narrow, we prevent perception from being too difficult once noise is added to the signal. In other words, the tighter underlying categories provide a stronger perceptual anchor that helps bias perception toward the category centers in the presence of interfering noise.

Given the differences across phoneme classes in the ratio between meaningful and noise variance observed in our simulations, we can ask whether there might be a mapping between certain features of the phonemes and where they lie on the τ continuum. One possibility is that the ratio between meaningful and noise variance depends on the type of acoustic cue that is used to identify a phoneme. To explore this possibility in more detail, we consider the static-dynamic and the spectral-temporal dimensions of acoustic cues to each type of phoneme. Static properties are ones that can be identified by looking at a small time slice of the spectral makeup of the sound and dynamic ones are ones that refer to change over time in the signal. Spectral properties refer to the frequencies that make up the signal, as seen in a spectrogram, while temporal properties are time

locked and depend on relative timing of parts of the signal. Most phonemes are identified by some combination of these types of properties. While there is evidence that all types of cues contribute to the identification and discrimination of all types of phonemes, certain cues have been found to be the most critical, and in many instances, sufficient for the behavior seen in studies with natural stimuli. We consider those that are specifically relevant to the types of stimuli we consider in the studies in this paper, particularly consonants in /CV/ syllables and isolated vowels.

Stop consonants have largely been found to be identified by static temporal cues like VOT for the voicing distinction (Carney et al., 1977) and dynamic spectral cues such as rapid changes in spectra at release (Stevens and Blumstein 1978), time-varying spectral features (Kewley-Port, 1983), locus equations (Sussman et al. 1991), and changes in moments (Forrest et al., 1988). Vowels, at least when they are isolated without surrounding consonants, can be largely identified by the first and second formants, representing the steady state peaks of resonant energy with certain bandwidths, a static spectral cue. However, while the simulations we consider are based on isolated vowels, it has been shown that whenever examining natural human-produced stimuli, coarticulated vowels would always be identified more accurately than isolated vowels (Strange, 1989). When considering identification of coarticulated vowels in a normal speech stream, various measures of vowel inherent spectral change such as formant transitions play a role, changing the cues to identification to also include dynamic spectral cues and increasing reliance on temporal cues such as duration (see Hillenbrand (2013) for a thorough review of the history of research on vowel perception). We give further consideration to the difference between isolated and coarticulated vowels as well as consonants below. Fricatives have the widest array of critical cues (see (McMurray and Jongman, 2011) for an overview), including dynamic spectral cues in the form of locus equations (Jongman et al. 2000) and changes in moments (Forrest et al. 1988), static temporal noise duration (Strevens, 1960; Jassem, 1962), and static spectral cues including F2 onset frequency, spectral moments, and spectral peak location (Jongman et al. 2000).

Mirman et al. (2004) showed that the type of acoustic cue used to distinguish a contrast affects listeners' discrimination behavior, using evidence from listeners' identification and discrimination of non-speech stimuli. They trained listeners to categorize stimuli along a continuum that was differentiated either by steady state spectral cues (similar to isolated vowels), or by rapidly changing acoustic information (similar to stop consonants). Stimuli with steady state spectral cues were discriminated more accurately than would be predicted by the categorization-based model from Liberman et al. (1957), whereas stimuli with rapidly changing cues were discriminated at a level approximating the predictions of the categorization-based model. Mirman et al. hypothesized that the perceptual traces of rapidly changing acoustic cues decay faster than those of steady state cues, consistent with the idea that there is more perceptual noise associated with the perception of rapidly changing cues. Static spectral cues, with their longer duration and continuous presence of the signal to be detected, would be less prone to interference. This parallels ideas from Pisoni (1975), who considers two modes of perception: acoustic and phonetic. If static spectral cues trigger acoustic processing and dynamic spectral and static temporal cues trigger phonetic processing, then we get a pattern that corresponds closely to our findings and those of Mirman et al. (2004). The match is not perfect: although fricative perception is typically sensitive to dynamic cues, the Lago et al. (2015) experiment modeled here used only static spectral cues to fricative identity. However, it is possible that listeners in their experiment may have estimated the noise variance based on cues that are typically present in fricatives, including dynamic cues. Examining a greater variety of phonemes would help establish the strength of the correlation between cue type and strength of categorical effects, and would allow us to examine which acoustic cues to perception are most strongly correlated with the degree of categorical effects.

At a higher level, a factor that may serve as meaningful variance is indexical variability. Listeners use vowel variation for identifying speakers, dialects, and accents. Vowel inherent spectral change has been found to be central for differentiating regional variations of American English (Jacewicz and Fox, 2013), and vowel spectra also effectively differentiate American, British, and Australian English (Ghorshi et al., 2008). Listeners are most sensitive to indexical information in vowels rather than consonants: they are more effective at discriminating between speakers based on vowels than consonants (Owren & Cardillo, 2006), and are more likely to notice when vowels, rather than consonants, are replaced with tokens from other speakers (Hertz et al., 2004; Hertz, 2006). If vowels are naturally more prone to slight differences across accents and speakers than consonants are, and listeners are sensitive to such differences, this could lead listeners to learn to treat vowels as though they have high meaningful variance.

We next consider the notion of restructuring in the speech stream, or coarticulation effects, whereby sounds exhibit varying acoustic realizations as a result of varying context. Previous studies have found a large amount of restructuring for stop consonant and much less for steady-state vowels (Liberman et al. 1967, Lisker & Abramson, 1964b, Liberman et al. 1954). This led researchers to propose that the degree of restructuring correlated with categorical effects on perception, namely that those sounds that

exhibited a large amount of restructuring were perceived categorically and those that did not are perceived continuously. At first pass, this could mean that restructuring is associated with unexplained noise variance, since an increase in this variance would lead to lower τ values, and in turn to more categorical perception. However, there is reason to believe that there is more to the story. Fricatives do not follow the given pattern, patterning with vowels on restructuring effects (Harris, 1958; Hughes & Halle, 1956) but exhibiting categorical effects closer to stop consonants. In addition, listeners would not want to discount coarticulatory effects, which have been shown to be used by both infants and adults to help identify upcoming words (Gow, 2001; Salverda et al., 2014; Mahr et al., 2015). Interpreting coarticulation as meaningful would not necessarily predict that stop consonants have more meaningful variance than vowels, as vowels carry large amounts of meaningful information related to speaker identity and dialect. Further research can address this question of coarticulation and restructuring and its association with meaningful and noise variance.

Finally, many phonological theories hypothesize binary distinctive features. While we know of no a priori reason to link particular abstract phonological features with meaningful or noise variability, it is possible that these phonological distinctive features are related to the degree of categorical effects of a category. In terms of the model, this would mean that sounds that possess or lack a certain feature might be more or less prone to meaningful or noise variance. In terms of features, stop consonants and fricatives share certain feature values to the exclusion of vowels, such as [-syllabic], [-approximant], and [-sonorant]. Fricatives also share feature values with vowels, to the exclusion of stop consonants, namely being [+continuant]. If having positive values for these features insulates a phoneme from noise interference (or, likewise, having negative values for these features exposes a phoneme to noise interference), then we would predict continuous perception of vowels (positive for all four features), the most noise interference, and hence categorical perception, for stop consonants (negative for all four features), and fricatives somewhere between the end points (positive for one feature). If we consider the cluster of consonants as a whole on the strong end of the τ continuum, it leaves [-syllabic], [-approximant], and [-sonorant] as the features likely relates to strong categorical effects (or, depending on the viewpoint, positive values for these features may be linked to more continuous perception). Here, again, testing a broader set of phonemes, particularly liquid and nasal consonants, would shed more light on this question.

One way to test how correlated properties such as binary features, acoustic cues, coarticulation, and indexical variability are with τ is to vary them explicitly and see how this affects perception behavior, and in turn the retrieved value of the variance ratio. The model makes a prediction that if listeners are trained on data in which the meaningful and unexplained noise variance are varied, then they should eventually exhibit the predicted categorical perception effects. One example of such an experiment is Mirman et al. (2004), who showed that familiarization with varying acoustic cues correlated with identification and discrimination behavior. Another way to see the effect of varying the variances on τ is to manipulate the presence of noise variance in the signal during testing. Feldman et al. (2009) conducted such an experiment with vowel perception where they had a no noise and a noise condition, in which they artificially added noise to the signal. This changed the ratio of meaningful to noise variance, and they saw results that correlated with the predictions of the model. Conducting experiments such as these can help us tease apart the contribution of each of these factors to listeners' perceived meaningful and noise variance.

Individual differences

In this paper, all the modeling is based on data aggregated across participants in experiments. However, it is known that speakers of a language exhibit individual differences for a variety of phonetic categories, for both production (Hillenbrand et al. 1995, Peterson and Barney 1952, Newman et al. 2001, Allen et al. 2003, Byrd 1992, Zue & Laferriere 1979, Espy-Wilson et al. 2000, Hashi et al. 2003) and perception (Liberman et al., 1957; Theodore et al., 2013). Fitting the model to aggregate rather than individual participants' data could affect our inferences about where individual phonemes are located on the τ continuum. These τ values are contingent on the steepness of the ID curve as well as the peakedness of the discrimination curve. Both of these curves will appear shallower than they should be in the aggregate form if individuals have shifted boundaries and peaks. This would lead the model to overestimate the meaningful variance and find smaller biases toward category centers than individual listeners actually have. This leads to phonemes falling higher on the continuum than they might when fitting to individuals' data.

This may at least partially explain the finding for the voiceless stop consonant category /p/, which was found to be substantially higher on the τ continuum than its voiced counterpart /b/. Voiceless stop consonants such as /p/ have been found to vary in their VOT values and the effects of speech rate on these values (Allen et al., 2003; Theodore et al., 2009). These individual differences in production may be related to perceptual judgments, which would in turn yield a more continuous profile for the category than any given individual may actually have.

Applying the model to individual data collected from a new set of experiments would allow us to examine individuals' perception patterns and validate the findings reported here from group data. While the particular quantitative estimates of τ may change if estimated from individual data, the overall pattern of results would likely be similar since all the phonetic categories examined here show individual differences to some degree.

Extensions and future directions

Thus far, we have shown that our model accounts for phonetic perception, but only for a specific set of phonemes, namely bilabial stop consonants, sibilant fricatives, and front non-low tense vowels. Considering additional classes of phonemes would allow us to consider in more depth the correlation between these categorical effects and the related phonological and perceptual cues discussed above, such as phonetic features. In particular, a logical next step would be to apply this model to behavioral data from nasal and liquid consonant perception, to get representation from all the major classes of consonants. For nasals, we could vary the initial formant transitions to create a place of articulation continuum from /m/ to /n/, testing the model on a place of articulation dimension. For liquids, continua could be constructed by varying either the spectral cues of frequency onset or F_2/F_3 transition, or the temporal cues of initial steady state duration or F_1 transition duration. These cues exhibit trading relations with each other, and all contribute to identification and discrimination of liquids (Polka & Strange, 1985).

The case of the liquids warrants a consideration of the number of dimensions that the model handles. It is a simplification to treat any continua of sounds as purely varying in a single dimension. Even for vowels, if we wanted to consider a case with more than two phonemes, it would be impossible to have them lie along a single line. An initial multidimensional version of the model we use here is described by Barrios (2013), and this can be applied to perception of an arbitrary number of classes of sounds along arbitrary number of dimensions. Expanding the model to account for multiple classes and dimensions would also allow us to study more ecologically valid sound perception in situations where identification is not constrained to a binary choice and multiple sounds affect the representations that underlie discrimination ability.

A second future direction concerns the source of phonetic representations, how they come online in speakers and listeners, and how they can change over time with exposure or varying input, such as when a person learns a second language. If we view the source of categorical effects as a black box that takes in parameters for meaningful and noise variances, the category centers, and the actual acoustics being received and outputs a weighted target, then we need to ask how children learn to separate the two sources of variance for the separate phonemes. We can potentially get insight into exactly what is being learned by considering the time course of learning that a child goes through. We know from previous work that children acquire linguistically relevant categories for their language very early in life (Werker et al., 1981), but we do not yet know when the perceptual warping associated with categorical effects on perception becomes stable. How these effects correlate with children's phonological development in terms of both perception and production can be informative as to the source of the effects and how they are learned, making this a prime area for consideration in further study.

A particularly appealing possibility, initially proposed by Barrios (2013), is to use this model to quantify effects found in initial second language speech perception and changes in second language learners' performance over time. Two classic models that have been widely used to explain these perception effects in second language learning rely on relationships between categories in the first language (L1) and second language (L2) to predict which are easier or harder to acquire: Flege's Speech Learning Model (Flege, 1987) and Best's Perceptual Assimination Model (Best et al., 1988). Flege's model uses equivalence classification to explain why sounds in L2 that are similar to sounds in L1 will be difficult to both perceive and produce, whereas sounds that are considered new, or different from any classes in the L1, will be easier to process. Best's model also uses similarity between sounds in L1 and L2 to predict how they will be processed. Best discusses three options for processing of non-native speech contrasts: sounds can be assimilated to a close native category, sounds may be assimilated to a new category that doesn't exist in L1, or sounds may not be recognized as speech and stay unassimilated to any category. The models make predictions about initial perception as well as eventual acquisition, critically relying on categorical representations in the two languages to do so. However, neither of these models makes a quantitative statement as to the source of this effect nor the representations involved.

By having the speaker and listener in our model employ different sets of categories, we can simulate the case of the L2 learner. Using continua of stimuli, and looking at changes over time, we can infer what the assumed means and variances are for various phonemes and how these parameters change over time. The same τ parameter could explain why effects are greater or smaller across languages for different phonemes and we can provide a quantitative account for the well-understood, but underspecified, findings in second language perception and acquisition. There is already some indication that this can be done, with an initial attempt to apply this model to native Spanish speakers' perception of English appearing in Barrios (2013).

Conclusion

Native language categories influence our perception of speech sounds, and it is often assumed that there are different types of categorical effects for different phonemes, particularly stop consonants and vowels. This paper has instead investigated a unified explanation by applying a Bayesian computational model to the problem of speech perception. The model treats sound perception as optimal statistical inference of an intended target production in the presence of two sources of variance in the signal: meaningful category variance and articulatory and perceptual noise variance. We derive τ, the ratio of meaningful to noise variance, that corresponds to the degree of categorical effects for a given phoneme. Our results demonstrate that different quantitative findings in categorical effects on phoneme perception are in fact qualitatively similar for stop consonant, fricative, and vowel perception and that they are attributable to variation of a single parameter in a unified model of speech perception.

Author Note This research was supported in part by NSF IGERT grant DGE-0801465 and NSF grant BCS-1320410. We thank Bill Idsardi for his ongoing input on the project, Sol Lago and Mathias Scharinger for allowing us to use raw data from experiments conducted jointly with them, and audiences at the 2012 Cognitive Science Society conference (Kronrod et al. 2012) and the University of Maryland Language Science Lunch Talk series for helpful comments on earlier versions of this research. In addition, we thank two anonymous reviewers for helpful comments on an earlier version of this paper.

Appendix A: Identification derivation

This appendix provides a detailed derivation of the equations used to fit behavioral identification data. We assume that the data collected in the behavioral forced-choice identification experiment is an empirical measure of the model probability $p(c_1|S)$, where we observe the rate at which the listener chooses one of the two categories, say c_1, upon observing a speech stimulus, S. According to Bayes' rule (1) we can write this quantity as

$$p(c_1|S) = \frac{p(S|c_1)p(c_1)}{p(S|c_1)p(c_1) + p(S|c_2)p(c_2)} \tag{12}$$

Applying the distributions from the probabilistic model, we have

$$p(c_1|S) = \frac{N(\mu_{c1}, \sigma_{c1}^2 + \sigma_S^2)p(c_1)}{N(\mu_{c1}, \sigma_{c1}^2 + \sigma_S^2)p(c_1) + N(\mu_{c2}, \sigma_{c2}^2 + \sigma_S^2)p(c_2)} \tag{13}$$

We make the assumption that the two categories in the forced choice identification trial have equal prior probabilities, so we replace both prior probabilities with 0.5

$$p(c_1|S) = \frac{N(\mu_{c1}, \sigma_{c1}^2 + \sigma_S^2)(0.5)}{(0.5)N(\mu_{c1}, \sigma_{c1}^2 + \sigma_S^2) + (0.5)N(\mu_{c2}, \sigma_{c2}^2 + \sigma_S^2)} \tag{14}$$

and divide through by $N(\mu_{c1}, \sigma_{c1}^2 + \sigma_S^2)(0.5)$, giving us

$$p(c_1|S) = \frac{1}{1 + \frac{N(\mu_{c2}, \sigma_{c2}^2 + \sigma_S^2)}{N(\mu_{c1}, \sigma_{c1}^2 + \sigma_S^2)}} \tag{15}$$

If we focus on the term in the denominator, $\frac{N(\mu_{c2}, \sigma_{c2}^2 + \sigma_S^2)}{N(\mu_{c1}, \sigma_{c1}^2 + \sigma_S^2)}$, and replace the normal distributions with the full form of a Gaussian distribution, $N(\mu, \sigma^2) = \sqrt{\frac{1}{2\pi\sigma^2}} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$, we obtain (just for this term)

$$\frac{N(\mu_{c2}, \sigma_{c2}^2 + \sigma_S^2)}{N(\mu_{c1}, \sigma_{c1}^2 + \sigma_S^2)} = \frac{\sqrt{\frac{1}{2\pi(\sigma_{c2}^2 + \sigma_S^2)}} exp\left(-\frac{(S-\mu_{c2})^2}{2(\sigma_{c2}^2 + \sigma_S^2)}\right)}{\sqrt{\frac{1}{2\pi(\sigma_{c1}^2 + \sigma_S^2)}} exp\left(-\frac{(S-\mu_{c1})^2}{2(\sigma_{c1}^2 + \sigma_S^2)}\right)} \tag{16}$$

Simplifying the square root term, applying the division rule for exponents we get

$$\sqrt{\frac{(\sigma_{c1}^2 + \sigma_S^2)}{(\sigma_{c2}^2 + \sigma_S^2)}} exp\left(-\frac{(S-\mu_{c2})^2}{2(\sigma_{c2}^2 + \sigma_S^2)} + \frac{(S-\mu_{c1})^2}{2(\sigma_{c1}^2 + \sigma_S^2)}\right) \tag{17}$$

Expanding the squares and simplifying we get

$$\sqrt{\frac{(\sigma_{c1}^2 + \sigma_S^2)}{(\sigma_{c2}^2 + \sigma_S^2)}} exp\left(-\frac{S^2 - 2S\mu_{c2} + \mu_{c2}^2}{2(\sigma_{c2}^2 + \sigma_S^2)} + \frac{S^2 - 2S\mu_{c1} + \mu_{c1}^2}{2(\sigma_{c1}^2 + \sigma_S^2)}\right) \tag{18}$$

Converting to a common denominator, simplifying, and plugging back into the original equation gives us

$$p(c_1|S) = \frac{1}{1 + \sqrt{\frac{\sigma_1^2}{\sigma_2^2}} \times exp\frac{(\sigma_2^2 - \sigma_1^2)S^2 + 2(\mu_{c2}\sigma_1^2 - \mu_{c1}\sigma_2^2)S + (\mu_{c1}^2\sigma_2^2 - \mu_{c2}^2\sigma_1^2)}{2\sigma_1^2\sigma_2^2}} \tag{19}$$

where $\sigma_1^2 = \sigma_{c1}^2 + \sigma_S^2$ and $\sigma_2^2 = \sigma_{c2}^2 + \sigma_S^2$.

Fitting this equation to the behavioral data from our experiments, along with the value for one of the two category means set before the fitting procedure, we are able to derive optimal values for the following four parameters: μ_{c_1}, μ_{c2}, $\sigma_1^2 = \sigma_{c1}^2 + \sigma_S^2$, and $\sigma_2^2 = \sigma_{c2}^2 + \sigma_S^2$.

Appendix B: Discrimination derivation

This appendix gives the detailed derivation of the expected value of the posterior of the target production given the

speech sound, $E[T|S]$. By definition of expected value, this is equal to

$$E[T|S] = \int Tp(T|S)dT \tag{20}$$

In our paradigm, the intended target production, T, is potentially derived from a number of categories, particularly in our case, two of them, c_1 and c_2. Hence, we need to marginalize the posterior $p(T|S)$ over the possible categories as $p(T|S) = \sum_c p(T|S,c)p(c|S)$,

$$E[T|S] = \int T\sum_c p(T|S,c)p(c|S)dT \tag{21}$$

Since the term $p(c|S)$ does not depend on T, we can rewrite the expected values as

$$E[T|S] = \sum_c p(c|S)\int Tp(T|S,c)dT \tag{22}$$

We already know the term outside of the integral, $p(c|S)$, from the derivation in Appendix A (13–15). We need to now figure out a way to calculate the inside term, $\int Tp(T|S,c)dT$.

Again by definition of expected value, we know that $\int Tp(T|S,c)dT = E[T|S,c]$, so we can rewrite the equation as

$$E[T|S] = \sum_c p(c|S)E[T|S,c] \tag{23}$$

Hence, we need to calculate the expected value of the intended target production for a specific perceived category. We can do this by considering the distribution $p(T|S,c)$.

Applying Bayes rule, we get the following:

$$p(T|S,c) = \frac{p(S|T)p(T|c)}{\int_T p(S|T)p(T|c)} \tag{24}$$

If we rewrite using the actual probability distributions from the model, and remove the normalizing term in the denominator, since we are interested in relative values and not absolutes, we get the following proportional value for $p(T|S,c)$:

$$p(T|S,c) \propto N(T,\sigma_S^2)N(\mu_c,\sigma_c^2) \tag{25}$$

We can replace the normal distribution with the equation for the Gaussian, $N(\mu,\sigma^2) = \sqrt{\frac{1}{2\pi\sigma^2}}exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$, to get

$$p(T|S,c) \propto \frac{1}{\sqrt{2\pi\sigma_c^2}}\exp\left(-\frac{(T-\mu_c)^2}{2\sigma_c^2}\right)\frac{1}{\sqrt{2\pi\sigma_S^2}}\exp\left(-\frac{(S-T)^2}{2\sigma_S^2}\right) \tag{26}$$

Since we are considering the proportional value we can remove constants that do not depend on T. Removing the two square root normalizing terms and combining exponents we get:

$$p(T|S,c) \propto \exp\left(-\frac{(T-\mu_c)^2}{2\sigma_c^2} - \frac{(S-T)^2}{2\sigma_S^2}\right) \tag{27}$$

Expanding the squared terms and separating the components in the exponent we get the form

$$p(T|S,c) \propto \exp\left(-\frac{T^2}{2\sigma_c^2} + \frac{2T\mu_c}{2\sigma_c^2} - \frac{\mu_c^2}{2\sigma_c^2} - \frac{S^2}{2\sigma_S^2} + \frac{2ST}{2\sigma_S^2} - \frac{T^2}{2\sigma_S^2}\right) \tag{28}$$

We can take the two parts of the exponent that do not depend on T and move them into a separate exponent term,

$$p(T|S,c) \propto \exp\left(-\frac{T^2}{2\sigma_c^2} + \frac{2T\mu_c}{2\sigma_c^2} + \frac{2ST}{2\sigma_S^2} - \frac{T^2}{2\sigma_S^2}\right)\exp\left(-\frac{\mu_c^2}{2\sigma_c^2} - \frac{S^2}{2\sigma_S^2}\right) \tag{29}$$

Since this separate term is now just a scalar that does not depend on T, we can remove it entirely, while preserving proportionality:

$$p(T|S,c) \propto \exp\left(-\frac{T^2}{2\sigma_c^2} + \frac{2T\mu_c}{2\sigma_c^2} + \frac{2ST}{2\sigma_S^2} - \frac{T^2}{2\sigma_S^2}\right) \tag{30}$$

With a common denominator we get:

$$p(T|S,c) \propto \exp\left(\frac{-\sigma_S^2T^2 + 2\sigma_S^2T\mu_c + 2\sigma_c^2ST - T^2\sigma_c^2}{2\sigma_c^2\sigma_S^2}\right) \tag{31}$$

We would like to get this into a form that looks like a Gaussian, so we will need to complete the square. In order to see how to do this we want to group the T^2 and the T terms and see what we need to complete the square. Grouping the terms we get

$$p(T|S,c) \propto \exp\left(-\frac{\sigma_c^2+\sigma_S^2}{2(\sigma_c^2\sigma_S^2)}T^2 + \frac{2(\sigma_c^2S+\sigma_S^2\mu_c)}{2(\sigma_c^2\sigma_S^2)}T\right) \tag{32}$$

Now we can isolate the T^2 term by diving through by $\sigma_c^2 + \sigma_S^2$ to get

$$p(T|S,c) \propto \exp\left(-\frac{T^2 - 2\frac{\sigma_c^2S+\sigma_S^2\mu_c}{\sigma_c^2+\sigma_S^2}T}{2\frac{\sigma_c^2\sigma_S^2}{\sigma_c^2+\sigma_S^2}}\right) \tag{33}$$

We can now complete the square in the numerator by multiplying the equation by a scalar not dependent on T. If we

multiply the proportion by $\exp\left(-\frac{\frac{(\sigma_c^2 S+\sigma_S^2\mu_c)^2}{(\sigma_c^2+\sigma_S^2)^2}}{2\frac{\sigma_c^2\sigma_S^2}{\sigma_c^2+\sigma_S^2}}\right)$ and move the exponent term inside we get the form

$$p(T|S,c) \propto \exp\left(-\frac{T^2 - 2\frac{\sigma_c^2 S+\sigma_S^2\mu_c}{\sigma_c^2+\sigma_S^2}T + \frac{(\sigma_c^2 S+\sigma_S^2\mu_c)^2}{(\sigma_c^2+\sigma_S^2)^2}}{2\frac{\sigma_c^2\sigma_S^2}{\sigma_c^2+\sigma_S^2}}\right) \tag{34}$$

This can be rewritten as the complete square

$$p(T|S,c) \propto \exp\left(-\frac{(T - \frac{\sigma_c^2 S+\sigma_S^2\mu_c}{\sigma_c^2+\sigma_S^2})^2}{2\frac{\sigma_c^2\sigma_S^2}{\sigma_c^2+\sigma_S^2}}\right) \tag{35}$$

This now looks precisely like a normal distribution,

$$p(T|S,c) = N\left(\frac{\sigma_c^2 S+\sigma_S^2\mu_c}{\sigma_c^2+\sigma_S^2}, \frac{\sigma_c^2\sigma_S^2}{\sigma_c^2+\sigma_S^2}\right) \tag{36}$$

where the mean is $\frac{\sigma_c^2 S+\sigma_S^2\mu_c}{\sigma_c^2+\sigma_S^2}$ and the variance is $\frac{\sigma_c^2\sigma_S^2}{\sigma_c^2+\sigma_S^2}$.

The expected value $E[T|S,c]$ is precisely the mean of the distribution $p(T|S,c)$. From this normal distribution we have the mean so we have found the expected value

$$E[T|S,c] = \frac{\sigma_c^2 S+\sigma_S^2\mu_c}{\sigma_c^2+\sigma_S^2} \tag{37}$$

We can plug this back into Eq. 23 to get

$$E[T|S] = \sum_c p(c|S)\frac{\sigma_c^2 S+\sigma_S^2\mu_c}{\sigma_c^2+\sigma_S^2} \tag{38}$$

For the case of two categories we consider throughout this paper, this has the following form:

$$E[T|S] = p(c_1|S)\frac{\sigma_{c1}^2 S+\sigma_S^2\mu_{c1}}{\sigma_{c1}^2+\sigma_S^2} + p(c_2|S)\frac{\sigma_{c2}^2 S+\sigma_S^2\mu_{c2}}{\sigma_{c2}^2+\sigma_S^2} \tag{39}$$

The value that is fit from this equation is σ_S^2. Since we already have the sums $\sigma_S^2+\sigma_{c_1}^2$ and $\sigma_S^2+\sigma_{c_2}^2$ from the identification part of the simulations, we can subtract the σ_S^2 term in order to get the individual category variances σ_{c1}^2 and σ_{c2}^2. This gives us the final parameters needed to derive the fit of the model to the behavioral data for any set of identification and discrimination experiments.

References

Aaltonen, O., Eerola, O., Hellström, A., Uusipaikka, E., & Lang, A. H. (1997). Perceptual magnet effect in the light of behavioral and psychophysiological data. *Journal of the Acoustical Society of America*, *101*(2), 1090–1105.

Abramson, A. S. (1961). Identification and discrimination of phonemic tones. *Journal of the Acoustical Society of America*, *33*, 842.

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

Anderson, J. A., Silverstein, J. W., Ritz, S. A., & Jones, R. S. (1977). Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychological Review*, *84*(5), 413–451.

Andruski, J. E., Blumstein, S. E., & Burton, M. (1994). The effect of subphonemic differences on lexical access. *Cognition*, *52*, 163–187.

Angeli, A., Davidoff, J., & Valentine, T. (2008). Face familiarity, distinctiveness, and categorical perception. *The Quarterly Journal of Experimental Psychology*, *61*(5), 690–707.

Barrios, S. (2013). *Similarity in l2 phonology*. PhD thesis, University of Maryland.

Bastian, J., & Abramson, A. S. (1962). Identification and discrimination of phonemic vowel duration. *Journal of the Acoustical Society of America*, *34*, 743.

Bastian, J., Delattre, P. C., & Liberman, A. M. (1959). Silent interval as a cue for the distinction between stops and semivowels in medial position. *Journal of the Acoustical Society of America*, *31*, 1568.

Bastian, J., Eimas, P. D., & Liberman, A. M. (1961). Identification and discrimination of a phonemic contrast induced by silent interval. *Journal of the Acoustical Society of America*, *33*, 842.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society*, *53*, 370–418.

Beale, J. M., & Keil, F. C. (1995). Categorical effects in the perception of faces. *Cognition*, *57*, 217–239.

Best, C. T., McRoberts, G. W., & Sithole, N. M. (1988). Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(3), 345–360.

Blumstein, S. E., Myers, E. B., & Rissman, J. (2005). The perception of voice onset time: an fMRI investigation of phonetic category structure. *Journal of Cognitive Neuroscience*, *17*(9), 1353–1366.

Boersma, P., & Weenink, D. (2005). Praat: Doing phonetics by computer. (Version 4.3.01) [Computer program]. Retrieved from http://www.praat.org/.

Byrd, D. (1992). Preliminary results on speaker-dependent variation in the TIMIT database. *Journal of the Acoustical Society of America*, *92*, 593–596.

Calder, A. J., Young, A. W., Perrett, D. I., Etcoff, N. L., & Rowland, D. (1996). Categorical perception of morphed facial expressions. *Visual Cognition*, *3*(2), 81–117.

Carney, A. E., Widin, G. P., & Viemeister, N. F. (1977). Noncategorical perception of stop consonants differing in VOT. *Journal of the Acoustical Society of America*, *62*(4), 961–970.

Chistovich, L. A. (1960). Classification Of rapidly repeated speech sounds. *Akusticheskij Zhurnal*, *6*, 392–398. (Translated in Soviet Physics-Acoustics, New York, 1961, 6, 393–398).

Crawford, J. C., & Wang, W.S.-Y. (1960). Frequency studies of English consonants. *Language & Speech*, *3*, 131–139.

Damper, R. I., & Harnad, S. R. (2000). Neural network models of categorical perception. *Perception and Psychophysics*, *62*(4), 843–867.

Davidoff, J., Davies, I., & Roberson, D. (1999). Colour categories in a stone-age tribe. *Nature*, *398*, 203–204.

Diesch, E., Iverson, P., Kettermann, A., & Siebert, C. (1999). Measuring the perceptual magnet effect in the perception of /i/ by German listeners. *Psychological Research Psychologische Forschung*, *62*, 1–19.

Eimas, P. D. (1963). The relation between identification and discrimination along speech and nonspeech continua. *Language & Speech*, *6*, 206–217.

Eimas, P. D., Siqueland, E. R., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants. *Science*, *171*(3968), 303–306.

Elman, J. L. (1979). Perceptual origins of the phoneme boundary effect and selective adaptation to speech: a signal detection theory analysis. *Journal of the Acoustical Society of America*, *65*, 190–207.

Espy-Wilson, C. Y., Boyce, S. E., Jackson, M., Narayanan, S., & Alwan, A. (2000). Acoustic modeling of american english /r/. *Journal of the Acoustical Society of America*, *108*, 343–356.

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

Flege, J. E. (1987). The production of 'new' and 'similar' phones in a foreign language: Evidence for the effect of equivalence classification. *Journal of Phonetics*, *15*, 47–65.

Fok, A. (1979). The frequency of occurrence of speech sounds and tones in Cantonese. In Lord, R. (Ed.) *Hong Kong Language Papers*. Hong Kong: Hong Kong University Press.

Forrest, K., Weismer, G., Milenkovic, P., & Dougall, R. N. (1988). Statistical analysis of word-initial voiceless obstruents: preliminary data. *Journal of the Acoustic Society of America*, *84*, 115–123.

Fry, B. (1947). The frequency of occurrence of speech sounds in Southern English archives. *Néerlandaises de Phonétique Expérimentale.*, *20*, 103–106.

Fry, D. B., Abramson, A. S., Eimas, P. D., & Liberman, A. M. (1962). The identification and discrimination of synthetic vowels. *Language and Speech*, *5*, 171–189.

Gerrits, E., & Schouten, M. E. H. (2004). Categorical perception depends on the discrimination task. *Perception and Psychophysics*, *66*(3), 363–376.

Ghorshi, S., Vaseghi, S., & Yan, Q. (2008). Cross-entropic comparison of formants of British, Australian, and American English accents. *Speech Communication*, *50*, 564–579.

Gimson, A. C. (1980). *An introduction to the pronunciation of english*, 3rd edition. London: Edward Arnold.

Goldstone, R. L., Lippa, Y., & Shiffrin, R. M. (2001). Altering object representations through category learning. *Cognition*, *78*, 27–43.

Gow, D. W. (2001). Assimilation and anticipation in continuous spoken word recognition. *Journal of Memory and Language*, *45*, 133–159.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Griffith, B. C. (1958). *A study of the relation between phoneme labeling and discriminability in the perception of synthetic stop consonants*. PhD thesis, University of Connecticut.

Guenther, F. H., & Gjaja, M. N. (1996). The perceptual magnet effect as an emergent property of neural map formation. *Journal of the Acoustical Society of America*, *100*(2), 1111–1121.

Harnad, S. (Ed.) (1987). *Categorical perception: The groundwork of cognition*. New York: Cambridge University Press.

Harris, K. S. (1958). Cues for the discrimination of American English fricatives in spoken syllables. *Language and Speech*, *1*, 1–7.

Harris, K. S., Bastian, J., & Liberman, A. M. (1961). Mimicry and the perception of a phonemic contrast induced by silent interval: Electromyographic an acoustic measures. *Journal of the Acoustical Society of America*, *33*, 842.

Hashi, M., Honda, K., & Westbury, J. R. (2003). Time-varying acoustic and articulatory characteristics of American English [[turned r]]: A cross-speaker study. *Journal of Phonetics*, *31*, 3–22.

Hertz, S. (2006). R, A model of the regularities underlying speaker variation: Evidence from hybrid synthesis. *Proceedings of Interspeech*.

Hertz, S. R., Spencer, I. C., & Goldhor, R. (2004). When can segments serve as surrogates?.

Hess, U., Adams, R., & Kleck, R. (2009). The categorical perception of emotions and traits. *Social Cognition*, *27*(2), 320–326.

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

Hillenbrand, J. M. (2013). Static and dynamic approaches to vowel perception. In Morrison, G. S., & Assmann, P. F. (Eds.) *Vowel Inherent Spectral Change,* volume 9 of *Modern Acoustics and Signal Processing*, (pp. 9–30). Berlin Heidelberg: Springer.

Hughes, G. W., & Halle, M. (1956). Spectral properties of fricative consonants. *Journal of the Acoustical Society of America*, *28*, 303–310.

Iverson, P., & Kuhl, P. K. (1995). Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling. *Journal of the Acoustical Society of America*, *97*(1), 553–562.

Iverson, P., & Kuhl, P. K. (1996). Influences of phonetic identification and category goodness on American listeners' perception of /r/ and /l/. *Journal of the Acoustical Society of America*, *99*(2), 1130–1140.

Iverson, P., & Kuhl, P. K. (2000). Perceptual magnet and phoneme boundary effects in speech perception: Do they arise from a common mechanism? *Perception and Psychophysics*, *62*(4), 874–886.

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., & Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition*, *87*, B47–B57.

Jacewicz, E., & Fox, R. A. (2013). Cross-dialectal differences in dynamic formant patterns in American English vowels. In Morrison, G. S., & Assmann, P. F. (Eds.) *Vowel Inherent Spectral Change, Modern Acoustics and Signal Processing*, (pp. 177–198). Berlin Heidelberg: Springer.

Jassem, W. (1962). Noise spectra of Swedish, English, and Polish fricatives. In *Proceedings of the Sepech Communication Seminar*, (pp. 1–4): Royal Institute of Technology Speech Transmission Laboratory.

Joanisse, M. F., Robertson, E. K., & Newman, R. L. (2007). Mismatch negativity reflects sensory and phonetic speech processing. *NeuroReport*, *18*(9), 901–905.

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

Kewley-Port, D. (1983). Time-varying features as correlates of place of articulation in stop consonants. *Journal of the Acoustical Society of America*, *73*, 322–335.

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), 148–203.

Kozhevnikov, V. A., & Chistovich, L. A. (1965). *Rech' Artikuljatsia i vosprijatije*: Moscow-Leningrad. Translated in Speech: Articulation and perception, Washington: Joint Publications Research Service, 1966, 30, 543.

Kröger, B. J., Birkholz, P., Kannampuzha, J., & Neuschaefer-Rube, C. (2007). Modeling the perceptual magnet effect and categorical perception using self-organizing neural networks. *International Congress of Phonetic Sciences*, 789–792.

Kronrod, Y., Coppess, E., & Feldman, N. (2012). A unified model of categorical effects in consonant and vowel perception. In *Proceedings of the Cognitive Science Society Annual Meeting*. Japan: Sapporo.

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception and Psychophysics*, *50*(2), 93–107.

Kuhl, P. K. (1993). Early linguistic experience and phonetic perception: Implications for theories of developmental speech perception. *Journal of Phonetics*, *21*, 125–139.

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*(5044), 606–608.

Lacerda, F. (1995). *The perceptual-magnet effect: an emergent consequence of exemplar-based phonetic memory* (Vol. 2, pp. 140–147). Stockholm: KTH and Stockholm University.

Lago, S., Scharinger, M., Kronrod, Y., & Idsardi, W. J. (2015). Categorical effects in fricative perception are reflected in cortical source information. *Brain and Language*, *143*, 52–58.

Liberman, A. M., Cooper, F., Shankweiler, D., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*, 431–461.

Liberman, A. M., Delattre, P. C., Cooper, F. S., & Gerstman, L. J. (1954). The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs*, *68*(8(379)).

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, *54*(5), 358–368.

Liberman, A. M., Harris, K. S., Kinney, J. A., & Lane, H. (1961). The discrimination of relative onset-time of the components of certain speech and nonspeech patterns. *Journal of Experimental Psychology*, *61*(5), 379–388.

Lisker, L., & Abramson, A. S. (1964a). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, *20*, 384–422.

Lisker, L., & Abramson, A. S. (1964b). Stop categories and voice onset time. In *Proceedings of the Fifth International Congress of Phonetic Sciences, Munster*.

Lively, S. E., & Pisoni, D. B. (1997). On prototypes and phonetic categories: a critical assessment of the perceptual magnet effect in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *23*(6), 1665–1679.

Lotto, A. J., Kluender, K. R., & Holt, L. L. (1998). Depolarizing the perceptual magnet effect. *Journal of the Acoustical Society of America*, *103*(6), 3648–3655.

Macmillan, N. A., Kaplan, H. L., & Creelman, C. D. (1977). The psychophysics of categorical perception. *Psychological Review*, *84*, 452–471.

Mahr, T., McMillan, B. T. M., Saffran, J. R., Weismer, S. E., & Edwards, J. (2015). Anticipatory coarticulation facilitates word recognition in toddlers. *Cognition*, *142*, 345–350.

Marr, D. (1982). *Vision: a computational investigation in the human representation of visual information*. San Francisco: Freeman & Co.

Massaro, D. W. (1987a). Categorical partition: A fuzzy logical model of categorical behavior. In Harnad, S. (Ed.) *Categorical perception: the groundwork of cognition*, (pp. 254–283): Cambridge University Press.

Massaro, D. W. (1987b). *Speech perception by ear and eye*. London: Lawrence Erlbaum Associates.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*, 1–86.

McMurray, B. (2009). *Klattworks: A [somewhat] new systematic approach to formant-based speech synthesis for empirical research. Manuscript*. in preparation.

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), 219-246.

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition*, *86*, B33–B42.

McQueen, J. M., Jesse, A., & Norris, D. (2009). No lexical–prelexical feedback during speech perception or: Is it time to stop playing those Christmas tapes? *Journal of Memory and Language*, *61*, 1–18.

Miller, J. L. (1994). On the internal structure of phonetic categories: a progress report. *Cognition*, *50*, 271–285.

Miller, J. L. (1997). Internal structure of phonetic categories. *Language and Cognitive Processes*, *12*(5/6), 865–869.

Miller, J. L., & Volaitis, L. E. (1989). Effect of speaking rate on the perceptual structure of a phonetic category. *Perception and Psychophysics*, *46*, 505–512.

Mines, M. A., Hanson, B. F., & Shoup, J. E. (1978). Frequency of occurrence of phonemes in conversational English. *Language & Speech*, *21*(3), 221–241.

Mirman, D., Holt, L. L., & McClelland, J. L. (2004). Categorization and discrimination of nonspeech sounds: Differences between steady-state and rapidly-changing acoustic cues. *Journal of the Acoustical Society of America*, *116*(2), 1198–1207.

Nearey, T. M., & Hogan, J. T. (1986). *Phonological Contrast in Experimental Phonetics: Relating Distributions of Production Data to Perceptual Categorization Curves*, (pp. 141–162). Orlando: Academic Press.

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

Obleser, J., Leaver, A., VanMeter, J., & Rauschecker, J. P. (2010). Segregation of vowels and consonants in human auditory cortex: Evidence for distributed hierarchical organization. *Frontiers in Psychology*, *1*(232).

Owren, M. J., & Cardillo, G. C. (2006). The relative roles of vowels and consonants in discriminating talker identity versus word meaning. *Journal of the Acoustical Society of America*, *119*(3), 1727–1739.

Perez, C. A., Engineer, C. T., Jakkamsetti, V., Carraway, R. S., Perry, M. S., & Kilgard, M. P. (2013). Different timescales for the neural coding of consonant and vowel sounds. *Cerebral Cortex*, *23*(3), 670–683.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, *24*(2), 175–184.

Pisoni, D. B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception and Psychophysics*, *13*(2), 253–260.

Pisoni, D. B. (1975). Auditory short-term memory and vowel perception. *Memory and Cognition*, *3*(1), 7–18.

Pisoni, D. B. (1977). Identification and discrimination of the relative onset time of two component tones: Implications for voicing perception. *Journal of the Acoustical Society of America*, *61*(5), 1352–1361.

Pisoni, D. B., & Lazarus, J. H. (1974). Categorical and noncategorical modes of speech perception along the voicing continuum. *Journal of the Acoustical Society of America*, *55*(2), 328–333.

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

Polka, L., & Strange, W. (1985). Perceptual equivalence of acoustic cues that differentiate /r/ and /l/. *Journal of the Acoustical Society of America*, *78*(4), 1187–1197.

Repp, B. H. (1981). Two strategies in fricative discrimination. *Perception and Psychophysics*, *30*(3), 217–227.

Repp, B. H., Healy, A. F., & Crowder, R. G. (1979). Categories and context in the perception of isolated steady-state vowels. *Journal of Experimental Psychology: Human Perception and Performance*, *5*(1), 129–145.

Salminen, N. H., Titinen, H., & May, P. J. C. (2009). Modeling the categorical perception of speech sounds: a step toward biological plausibility. *Cognitive, Affective, and Behavioral Neuroscience*, *9*(3), 304–313.

Salverda, A. P., Kleinschmidt, D., & Tanenhaus, M. K. (2014). Immediate effects of anticipatory coarticulation in spoken-word recognition. *Journal of Memory and Language*, *71*, 145–163.

Sauter, D., LeGuen, O., & Huan, D. (2011). Categorical perception of emotional facial expressions does not require lexical categories. *Emotion*, *11*(6).

Shi, L., Griffiths, T. L., Feldman, N. H., & Sanborn, A. N. (2010). Exemplar models as a mechanism for performing Bayesian inference. *Psychonomic Bulletin and Review*, *17*(4), 443–464.

Smits, R. (2001). Hierarchical categorization of coarticulated phonemes: a theoretical analysis. *Perception and Psychophysics*, *63*(7), 1109–1139.

Stevens, K. N. (1966). On the relations between speech movements and speech perception. In *Meeting of the XVIII International Congress of Psychology*. Moscow: Zeitshrift fur Phonetik, Sprachwissensschaft und Kommunikations-forschung.

Stevens, K. N., & Blumstein, S. E. (1978). Invariant cues for place of articulation in stop consonants. *Journal of the Acoustical Society of America*, *64*, 1358–1368.

Stevens, K. N., Ohman, S. E. G., & Liberman, A. M. (1963). Identification and discrimination of rounded and unrounded vowels. *Journal of the Acoustical Society of America*, *35*, 1900.

Stevens, K. N., Ohman, S. E. G., Studdert-Kennedy, M., & Liberman, A. M. (1964). Cross-Linguistic study of vowel discrimination. *Journal of the Acoustical Society of America, 36*, 1989.

Stevens, S. S., Volkmann, J., & Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *Journal of the Acoustical Society of America*, *8*, 185–190.

Strange, W. (1989). Evolving theories of vowel perception. *Journal of the Acoustical Society of America*, *85*(5), 2081–2087.

Strevens, P. (1960). Spectra of fricative noise in human speech. *Language and Speech*, *3*, 32–49.

Studdert-Kennedy, M., Liberman, A. M., & Stevens, K. N. (1963). Reaction time to synthetic stop consonants and vowels at phoneme centers and at phoneme boundaries. *Journal of the Acoustical Society of America*, *35*, 1900.

Studdert-Kennedy, M., Liberman, A. M., & Stevens, K. N. (1989). (1964). Reaction Time during the discrimination of synthetic stop consonants. *Journal of the Acoustical Society of America*, 36.

Sussman, H. A., McCaffrey, H. A., & Matthews, S. A. (1991). An investigation of locus equations as a source of relational invariance for stop place categorization. *Journal of the Acoustical Society of America*, *90*, 1309–1325.

Sussman, J. E., & Gekas, B. (1997). Phonetic category structure of [I]: extent, best exemplars, and organization. *Journal of Speech, Language, and Hearing Research*, *40*, 1406–1424.

Theodore, R. M., Miller, J. L., & DeSteno, D. (2009). Individual talker differences in voice-onset-time: Contextual influences. *Journal of the Acoustical Society of America*, *125*(6), 3974–3982.

Theodore, R. M., Myers, E. B., & Lomibao, J. (2013). Listeners sensitivity to talker differences in voice-onset-time: Phonetic boundaries and internal category structure. In *Proceedings of Meetings on Acoustics (POMA) 19*.

Thorburn, W. M. (1915). Occam's razor. *Mind*, *24*, 287–288.

Thorburn, W. M. (1918). The myth of Occam's razor. *Mind*, *27*, 345–353.

Thyer, N., Hickson, L., & Dodd, B. (2000). The perceptual magnet effect in Australian English vowels. *Perception and Psychophysics*, *62*(1), 1–20.

Tomaschek, F., Truckenbrodt, H., & Hertrich, I. (2011). Processing German vowel quantity: Categorical perception or perceptual magnet effect? In *Proceedings of the 17th International Conference of Phonetic Sciences*, (pp. 2002–2005).

Toscano, J. C., McMurray, B., Dennhardt, J., & Luck, S. J. (2010). Continuous perception and graded categorization: Electrophysiological evidence for a linear relationship between the acoustic signal and perceptual encoding of speech. *Psychological Science*, *21*(10), 1532–1540.

Treisman, M., Faulkner, A., Naish, P. L. N., & Rosner, B. S. (1995). Voice-onset time and tone-onset time: The role of criterion-setting mechanisms in categorical perception. *Quarterly Journal of Experimental Psychology*, *48A*(334-366).

Treisman, M., & Williams, T. C. (1984). A theory of criterion setting with an application to sequential dependencies. *Psychological Review*, *91*, 68–111.

Vallabha, G. K., & McClelland, J. L. (2007). Success and failure of new speech category learning in adulthood: Consequences of learned Hebbian attractors in topographic maps. *Cognitive, Affective, and Behavioral Neuroscience*, *7*, 53–73.

Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences*, *104*, 13273–13278.

Werker, J. F., Gilbert, J. H. V., Humphrey, K., & Tees, R. C. (1981). Developmental aspects of crosslanguage speech perception. *Child Development*, *52*, 349–355.

Wioland, F. (1972). Estimation de la frequence des phonemes en francais parle. *Travaux de l'Institut de Phonetique*, *4*, 177–204.

Wood, C. C. (1976). Discriminability, response bias, and phoneme categories in discrimination of voice onset time. *Journal of the Acoustic Society of America*, *60*(6), 1381–1389.

Zue, V. W., & Laferriere, M. (1979). Acoustic study of medial /t,d/ in American English. *Journal of the Acoustical Society of America*, *66*, 1039–1050.

Zwicker, E. (1961). Subdivision of the audible frequency range into critical bands. *The Journal of the Acoustical Society of America*, *33*, 248.