What defines a category? Evidence that listeners' perception is governed by generalizations

Rachael R. Richardson[1] **(rachaelr@umd.edu)**
Naomi H. Feldman[1,2] **(nhf@umd.edu)**
William Idsardi[1] **(idsardi@umd.edu)**
[1]Department of Linguistics and [2]Institute for Advanced Computer Studies
1401 Marie Mount Hall, University of Maryland, College Park, MD 20742

Abstract

Listeners draw on their knowledge of phonetic categories when identifying speech sounds, extracting meaningful structural features from auditory cues. We use a Bayesian model to investigate the extent to which their perceptions of linguistic content incorporate their full knowledge of the phonetic category structure, or only certain aspects of this knowledge. Simulations show that listeners are best modeled as attending primarily to the most salient phonetic feature of a category when interpreting a cue, possibly attending to other features only in cases of high ambiguity. These results support the conclusion that listeners ignore potentially informative correlations in favor of efficient communication.

Keywords: speech perception; categorization; voice onset time; speaking rate

Introduction

Identifying a category presumes a structural knowledge of the category, i.e., of the perceptual cues associated with the category and of how these perceptual cues relate to its more abstract features. In the domain of speech sounds, the phonetic category is a latent, linguistic variable explaining the observable variation in the signal. Knowing the structure of phonetic categories enables listeners to extract a message from the speech they hear. These categories are thought to be structured in terms of features, such as voicing, place, or manner, which facilitate generalization (e.g., Cristià & Seidl, 2008; Maye, Weiss, & Aslin, 2008). This paper explores the way in which listeners use phonetic features during perception.

We probe listeners' categorization of English stop consonants, which are typically characterized by their *voicing* and *place* features. Voiced stops /b,d,g/ differ from voiceless stops /p,t,k/ in the voicing feature,[1] whereas labials /b,p/, alveolars /d,t/, and velars /g,k/ differ in the place feature. Categorical perception has been found along acoustic dimensions relevant to both voicing and place (Liberman, Harris, Hoffman, & Griffith, 1957; Wood, 1976), suggesting that both types of features contribute to the intrinsic identity of a category. Knowing a stop consonant category entails knowing both its voicing and its place.

We employ a computational model to assess the explanatory power of different possible category encodings of these features during a perception task. Our focus is on perception of a durational cue, voice onset time (VOT), that is widely attested cross-linguistically as a cue to voicing contrasts in initial stop consonants (Lisker & Abramson, 1964). VOT is defined as the amount of time between the release of a stop consonant and the onset of glottal phonation. In voiceless stops, phonation substantially lags the release, whereas in voiced stops, phonation closely follows the release.

Although VOT serves primarily as a cue to voicing, it varies as a result of other factors as well. For example, place of articulation affects the distribution of VOT: Consonants articulated at the back of the mouth (e.g. velars) have significantly longer VOTs, whereas consonants articulated in the front of the mouth (e.g. labials) have shorter VOTs. This pattern is largely owed to phonetic universals, but there is enough cross-linguistic variation to require language-specific components in a complete account (Cho & Ladefoged, 1999).

Speaking rate also affects the distribution of VOT, as it affects all durational cues. However, unlike place of articulation and voicing, speech rate is not an intrinsic cue to category membership. Previous analyses have shown that listeners' adjustments to speech rate variation are robust when measured at different scales, with variation in both target syllable rate and target sentence rate contributing to changes in internal category structure (Wayland, Miller, & Volaitis, 1994).

We test three models of category encoding on their ability to predict listeners' perception of stop consonant categories from a single acoustic cue. In the first model, All Available Features (AAF), the likelihood function is generated from Gaussians jointly conditioned on both place and voicing features. The second model encodes categories as a distribution conditioned on a single feature: Voicing Only (VO). The third model is designed for Effective Ambiguity Resolution (EAR), and conditions recruitment of the place feature on the amount of uncertainty that remains after taking account of the voicing feature. Simulations show that listeners' behavior is better fit by a model that defines phonetic categories according to a single feature: voicing, with possible recruitment of the place feature when the distribution of cues defined by voicing is maximally uninformative. We argue that this behavior is consistent with that of an optimal listener who partitions perceptual space to maximize their ability to communicate efficiently.

We begin by reviewing data from Volaitis and Miller (1992) showing that differences in VOT distributions arise from variation not only in voicing, but also in speaking rate and place of articulation. The following section introduces our category encoding models. We then present simula-

[1]We refer to these categories as voiced and voiceless, despite the fact that word-initially, they are better characterized phonetically as being voiceless unaspirated and aspirated stops, respectively.

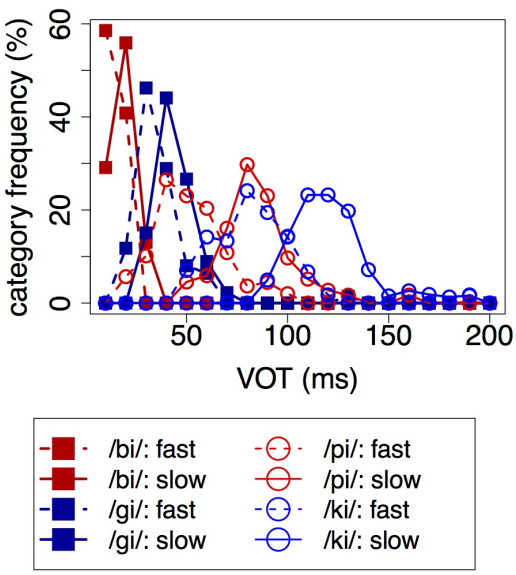

Figure 1: Digitized plot of selected production study data from Volaitis and Miller (1992)

tions showing that, excepting near the category boundary, the model that attends preferentially to the voicing feature provides a better account of listeners' perceptual behavior. We conclude by summarizing our findings and discussing implications for theories of speech perception.

Volaitis and Miller's Experiments

Volaitis and Miller (1992) investigated whether internal category structure could be considered context independent, exploring the effect of syllabic speaking rate on category structure in two experiments.

In a production experiment, participants were recorded speaking six syllables: three beginning with voiced stop consonants, /b,d,g/ and three beginning with their voiceless counterparts /p,t,k/. All ended with the vowel /i/. Each participant produced six instances of each syllable in the order /bi/, /pi/, /gi/, /ki/, /di/, /ti/ at eight different speech rates. This was repeated four times for each participant. To examine the effect of rate categorically, the syllables were then divided into three duration categories: 100-299 ms, 300-499 ms, and 500-799 ms. Volaitis and Miller found that VOT systematically increased with syllable duration for all three places of articulation. All speakers showed the same pattern of increasing VOT with increasing syllable duration. In addition, all speakers showed a pattern of increased VOT for stops articulated farther back in the mouth. Aggregated data for these four speakers from eight of these conditions are shown in Figure 1.

Volaitis and Miller then conducted a perception experiment to investigate how listeners adjust to this apparent systematic variation in VOT. A new group of participants was presented with a forced-choice categorization task. Participants were tested on four synthetically generated series of consonants along the voicing continuum. These continua were synthesized to have identical onsets but overall durations of 125 ms and 325 ms: a fast and slow condition, respectively. Two of the continua were synthesized as velar stop consonants, /gi/ and /ki/, while the other two were synthesized as labial stop consonants, /bi/ and /pi/. Participants were given options of identifying the stimuli as either /gi/ and /ki/ (for velar continua) or /bi/ and /pi/ (for labial continua),[2] and were asked to identify the sound they heard. Results confirmed a large and reliable effect of syllable duration on the location of the category boundary, with longer syllables more often evoking voiceless responses.

Volaitis and Miller conclude from this evidence that the perceptual mapping between acoustic structure and phonetic category is comprehensively altered with changes in speech rate. However, they did not ask whether differences in place of articulation alter the perceptual mapping between VOT and phonetic category identity in the same way. If place of articulation and speaking rate behave similarly, we would expect that examining listeners' behavior with respect to changes in place of articulation would reveal the same changes in listeners' internal category structures as were found in response to changes in speaking rate. On the other hand, if extrinsic factors such as speaking rate behave differently from intrinsic features such as place of articulation, we might expect different patterns of behavior in each case.

Our simulations of these data test three hypotheses. Our first hypothesis is that to complete the forced choice listening task, listeners recruit all available information. In this case, they should jointly infer both available features to stop consonant identity: place of articulation and voicing, and interpret the cue with respect to both at all places along the VOT continuum. Our second hypothesis is that we can more accurately describe listeners as preferentially inferring only the single most salient feature in this task. Listeners' responses are uniformly voiced for VOT below a threshold of about 35 ms, and universally voiceless for those above 80 ms. The place feature may therefore only be useful in categorizing stimuli between these values. Our third simulation tests the hypothesis that listeners use information adaptively, inferring the featural knowledge that will best help them solve the task effectively and efficiently, depending on the position in the VOT continuum.

Toscano and McMurray (2010) investigated a similar question using a learning model to categorize word-initial stop consonants. They found that discrepancies between production and perception data can be described as resulting from preferential down-weighting of cues which are less informative at a given position in the voicing continuum. Here we investigate a more abstract characterization of listeners' behavior, investigating categorization using a single cue with

[2]Participants were presented with three options: they could label the sound as the voiced category, the voiceless category, or a third unnatural voiceless category with an extremely long VOT. In our analysis we have collapsed the natural and unnatural voiceless categories, counting both as corresponding to a voiceless response.

multiple possible abstract featural specifications.

Model of Sound Categorization

Our model adopts a framework introduced by Nearey and Hogan (1986) and used in several recent models of speech perception (e.g., Clayards, Tanenhaus, Aslin, & Jacobs, 2008; Feldman, Griffiths, & Morgan, 2009; Kleinschmidt & Jaeger, 2015; Sonderegger & Yu, 2010). The model characterizes perception of speech sounds as a statistical inference problem. The goal of listeners, in perceiving a speech sound, is to infer the category of the sound using the information available to them from the speech signal and their prior knowledge of phonetic categories.

Following previous literature, we define phonetic categories as Gaussian distributions. In producing a VOT, speakers first select a stop consonant category and a speech rate, then articulate a production. If, at rate r, phonetic category c has mean μ_{cr} and variance σ^2_{cr}, speakers generate production x from that phonetic category with probability

$$x|c,r \sim \mathcal{N}(\mu_{cr}, \sigma^2_{cr}) \tag{1}$$

Inverting production data to produce perception data, we apply Bayes' rule. The posterior probability of perceiving a particular speech sound category from a given VOT at a particular rate is equal to the probability that the VOT was produced by that category and rate, weighted by the prior probability of that category occurring, normalized according to the probability of that VOT occurring,

$$p(c|x,r) = \frac{p(x|c,r)p(c)}{p(x)} \tag{2}$$

We begin by contrasting the All Available Features (AAF) model, which defines a phonetic category as a Gaussian distributions over equally weighted voicing and place features, with the Voicing Only (VO) model, which defines categories using only the voicing cue. The difference in the number of features affects the likelihood function given in Equation 1, which becomes

$$x|v,p,r \sim \mathcal{N}(\mu_{vpr}, \sigma^2_{vpr}) \tag{3}$$

for AAF and

$$x|v,r \sim \mathcal{N}(\mu_{vr}, \sigma^2_{vr}) \tag{4}$$

for VO.

For VOTs far from the categorical boundary, judgments are uniformly either voiced or voiceless. We investigated the possibility that for such values, a listener can be described as relying on the voicing feature alone, recruiting the less reliable place feature only when necessary to resolve a significant level of ambiguity between categories. Accordingly, our third model, Efficient Ambiguity Resolution (EAR) is an interpolation the AAF and VO models. EAR uses the uncertainty in the posterior distribution from the VO model to gate the recruitment of the AAF place-specific category knowledge.

To apply our model to the Volaitis and Miller experimental data, we estimate Gaussian distributions for each category from their production data. Given the heights of bars in their histograms, h, and the VOT values corresponding to those bars, x, we can compute maximum likelihood values for the mean and variance of a category as

$$\hat{\mu}_{cr} = \frac{\sum_i x_i \times h_i,}{\sum_i h_i} \tag{5}$$

$$\hat{\sigma}^2_{cr} = \frac{\sum_i (x_i - \hat{\mu}_{cr})^2 \times h_i,}{\sum_i h_i} \tag{6}$$

where i ranges over all productions at a given speaking rate and a given voicing value and, in AAF, a given place of articulation. We then compute the posterior distribution over category labels for each stimulus according to Equation 2 and compare it to the data from their perception experiment.

Simulations

We apply the models outlined above to the data from the Volaitis and Miller experiments. Data from different speaking rates are modeled separately, based on previous data that listeners compensate perceptually for changes in speaking rate.

We are primarily interested in the effect of differences in category structure on our model's ability to accurately portray the relationship between production and perception. The model which best preserves this relationship can be considered the better representation of listeners' phonetic cue processing.

Simulation 1: Joint Contributions of Place and Voicing

Our first hypothesis is that variation in VOT caused by both place of articulation and voicing will make significant contributions to listeners' inferences about linguistic content. To test this hypothesis, we use the AAF model.

Results are shown in Figures 2A and 2B. Dashed lines give the empirical data, and solid lines give the model predictions. The overall effects of speech rate and of place of articulation are evident in both model and data, with slower speech rates more often eliciting voiced category responses. However, the model predicts a much larger effect of place of articulation than is evinced in the perception data, with widely divergent category boundary predictions for the labial and velar sounds. For labials, the model predicts the category boundary at a shorter VOT than evinced by the behavioral experiments, whereas for velars, it predicts the category to be at a larger VOT. This may also be true of rate, which is predicted to have a larger effect than actually occurs in the perception data. However, the speaking rates in the perception experiment were far to one end of the range of speaking rates in the production data, and this prevents us from drawing strong conclusions about listeners' use of rate information.

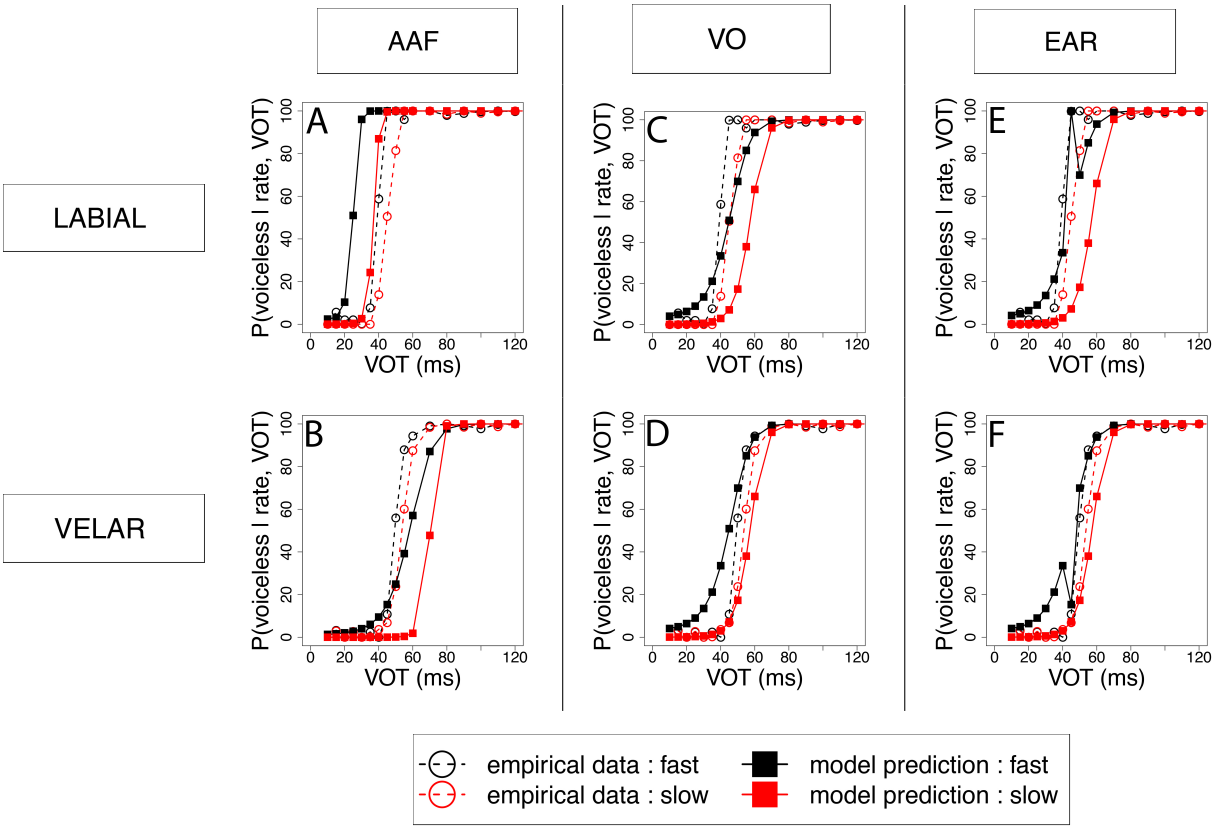

Figure 2: Empirical data and predictions of models evaluated with labial, velar data: AAF = A,B VO = C,D EAR = E,F

Simulation 2: Generalizing across Places

Our second hypothesis is that one feature, voicing, is overwhelmingly more informative to listeners, and that they primarily define their categories in terms of this feature when performing the forced choice task. To test this hypothesis, we use the VO model. Under this model, listeners are able to generalize with respect to the place feature, recruiting information about VOT variation at *all other* places of articulation to solve the inference problem.

Results are shown in Figures 2C and 2D. Dashed lines give the empirical data, and solid lines give the model predictions. Because the model's predictions are independent of place of articulation, it predicts the same identification function for both labials and velars. The solid black line shows the model's estimate of the empirical data for faster sounds, while the red line gives the model predictions for the slower sounds.

Omitting place as a relevant feature predicts that the shift due to rate will be uniform across places of articulation. This is borne out in the empirical data and model results, though the model predicts a substantially larger difference between rates than is seen empirically. The model also underestimates the slope of the categorical boundary, likely due to the increased variance of the likelihood functions that were estimated from combined labial and velar data. Nevertheless, our quantitative comparison will show that the VO model provides a better fit to the empirical data.

Simulation 3: Efficiently Resolving Ambiguity

Although listeners' inferences about linguistic content are generally dominated by voicing, variation in VOT caused by place of articulation makes a significant contribution which is most apparent near the category boundary. We hypothesize that the degree of recruitment of place features can be best described in terms of uncertainty about category identity. This strategy would facilitate efficient communication, allowing the listener to preferentially process less ambiguous cues using a simpler representation than required for more ambiguous cues. Such an account is compatible with theoretical accounts of listeners only recruiting additional information regarding category membership, including lexical status and visemes, as necessary to resolve ambiguous members (Green & Miller, 1985) and models of online perception of phonemes as underspecified lexical forms (e.g., Lahiri & Reetz, 2002). To test this hypothesis, we use a blend of the AAF and VO models. For each step along the continuum, we calculate the entropy of the category distribution as defined only by voicing. If the entropy in this distribution is below a specified threshold, then the categorization task is performed using the voicing distribution (VO). If the entropy is above that threshold, then results are calculated using a joint distribution on voicing and place (AAF).

Results for the threshold which produced the best fit to the perception data (0.68 bits) are shown in Figures 2E and 2F. Dashed lines give the empirical data, and solid lines show the model predictions. While the model relies exclusively on voicing for the slower continuua, for faster sounds in the 40-50 ms VOT range, the place-specific model provides a slightly better fit. This improvement, however, is entirely owed to a change in a single point in this part of the VOT continuum, for voiced sounds only. This change is enough to effect a change in boundary slope, successfully portraying a reduced effect of the place cue near the boundary compared to what would be predicted on the basis of speakers' productions.

Quantitative comparison

We compare the success of the models using cross entropy. Entropy is a measure of information: as the model more successfully predicts listeners' perceptions from production data, fewer bits will be required to encode the perception data using an optimal code derived from the production data. We thus hold the category encoding which produces the lower cross entropy to be a closer match to the way in which human listeners perform the perceptual task.

Cross entropy is computed as

$$H(\hat{p}) = -\frac{1}{n}\sum_i \sum_c p_{data}(c|v_i)\log p_{model}(c|v_i) \qquad (7)$$

where i ranges over VOT values, c ranges over the voiced and voiceless categories, and n is the number of steps in the continuum.

For each model, our definition of c will vary according to our hypothesis concerning the internal category structure. The AAF model posits a joint distribution conditioned on voicing and place features, while the VO model represents the category as a distribution conditioned on the voicing feature alone.

	Labial		Velar	
	Fast	Slow	Fast	Slow
AAF	0.061	0.023	0.014	0.044
VO	0.012	0.018	0.014	0.012
EAR	0.010	0.018	0.013	0.012

Table 1: Average Cross Entropy of Perception Data by Model

Results are given in Table 1. The average cross entropy for AAF is 0.035 bits and for VO it is 0.014. Thus, on average, we find that VO is about 60% more efficient than AAF, requiring an average of 0.021 fewer bits to encode the distribution in the perception data. The model thereby captures the observation that participants' performance is fairly homogenous between places of articulation, with place information appearing to only play a significant role in processing for sounds near the category boundary. This pattern emerges despite the existence of potentially useful invariant auditory cues distinguishing place of articulation (Stevens & Blumstein, 1978). Using the mixed strategy of EAR reduces the average cross entropy of our model to 0.013 bits, a roughly 6% reduction over the voicing-only strategy.

Although we do not have access to the raw data that would enable us to compute the log likelihood of these models directly, note that the cross entropy for a binomial distribution is closely related to its log likelihood, with the negative log likelihood being equal to the cross entropy in nats multiplied by the number of trials in the participants' data. Thus, the differences in cross entropy between AAF and VO found here are likely to translate into non-negligible changes in log likelihood when taking into account the fact that each point in each of the four continua represents responses across 15 trials for each of 12 participants.

Discussion

This paper used a Bayesian model to explore the relationship between categorical effects and the weighting of intrinsic phonetic category features in the context of stop consonants. Using production data to model perception of voicing contrasts at different speech rates, we compared three hypotheses as to the relationship of the linguistic features. The first simulation, attributing equal explanatory power to distinct intrinsic features, predicted an exaggerated effect of place. The second simulation, although it entirely omits place of articulation, provides a more accurate description of the relative effect of rate changes on the location of the category boundary. The third simulation provided the most accurate account, with listeners relying solely on the voicing feature for most stimuli, but recruiting the place feature to disambiguate sounds near the category boundary.

The VO model substantially outperforms the place-dependent AAF, yet completely fails to account for listeners' ability to discriminate stop consonants with different places of articulation. The EAR model provides a first attempt at balancing the efficacy of generalizing across similar categories, with the acuity of a model that recruits all available structural information in interpreting the available cue. Future work should explore whether combining these models using a different mechanism, such as weighted averaging, could produce a more accurate description of the categorical listener's behavior.

Our simulations suggest that overall, listeners exhibit a preferential treatment of the voicing feature when performing the forced choice task, *ignoring* the place feature for most stimuli, despite the apparently meaningful variation in categories it reflects. Rather than than ignoring specific cues, or physical aspects of the speech signal, listeners appear to be systematically ignoring specific abstract structural aspects of the categories whose identities they are inferring.

This work inherits a limitation of the previous study: Volaitis and Miller presented participants with a forced choice between two sounds both belonging to the same place of articulation. Therefore, cues specific to place, while available, could be deemed irrelevant to the task compared to cues

which distinguish VOT across places. However, the results of Simulation 3 suggest that listeners may not be ignoring this irrelevant dimension entirely, but rather interpreting variation in VOT due to place of articulation for a specific range of stimuli. To further test whether these perceptual patterns extend to situations in which there ambiguity as to the place of articulation, we could design a new behavioral study requiring participants to make judgments between sounds with different place features. Given a task which directed the listener's attention to both place and voicing during identification, listeners may be forced to rely on a smaller, more specialized set of exemplars in their decision process, resulting in more interaction between these two features. Nevertheless, our findings suggest that listeners are able to privilege some features while ignoring others in perceptual tasks, providing support for featural representations. This type of representation could also benefit listeners by allowing them to recognize underlying phonation categories in the absence of significant cues, creating a perceptual system which retains robust recognition even in severe noise.

Although it appears that listeners are actually making use of less information than they have available to them, perhaps by treating the categorization task as less specialized, and relying on exemplars from multiple places of articulation, they are actually increasing the amount of information available to them in the decision task. Generalizing – attributing observed variation to as few features as possible – allows the listener to posit that the preferred feature is not only most informative, but on its own, *informative enough*. This powerful assumption would not only endow the perceptual system with the ability to withstand noisy input, but to effectively encode ambiguity. Effective resolution of ambiguity is a key property of linguistic processing systems, reflecting an optimization of cue interpretation in accordance with communicative pressures.

Acknowledgments We thank the Probabilistic Modeling group and the Language Science Lunch group for valuable discussion and feedback. This work was supported in part by NSF IGERT grant DGE-0801465 and NSF grant BCS-1320410.

References

Cho, T., & Ladefoged, P. (1999). Variation and universals in VOT: evidence from 18 languages. *Journal of Phonetics*, *27*(2), 207–229.

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, *108*(3), 804-809.

Cristià, A., & Seidl, A. (2008). Is infants learning of sound patterns constrained by phonological features? *Language Learning and Development*, *4*(3), 203227.

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.

Green, K. P., & Miller, J. L. (1985). On the role of visual rate information in phonetic perception. *Perception psychophysics*, *38*(3), 269-276.

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.

Lahiri, A., & Reetz, H. (2002). Underspecified recognition. *Laboratory phonology*, *7*, 637–676.

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.

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

Maye, J., Weiss, D. J., & Aslin, R. N. (2008). Statistical phonetic learning in infants: facilitation and feature generalization. *Developmental Science*, *11*(1), 122-134.

Nearey, T. M., & Hogan, J. T. (1986). Phonological contrast in experimental phonetics: Relating distributions of production data to perceptual categorization curves. In J. J. Ohala & J. J. Jaeger (Eds.), *Experimental phonology* (p. 141-162). Orlando, FL: Academic Press.

Sonderegger, M., & Yu, A. (2010). A rational account of perceptual compensation for coarticulation. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (p. 375-380). Austin, TX: Cognitive Science Society.

Stevens, K. N., & Blumstein, S. E. (1978). Invariant cues for place of articulation in stop consonants. *The Journal of the Acoustical Society of America*, *64*(5), 1358-1368.

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

Volaitis, L. E., & Miller, J. L. (1992). Phonetic prototypes: Influence of place of articulation and speaking rate on the internal structure of voicing categories. *The Journal of the Acoustical Society of America*, *92*(2), 723-735.

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

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