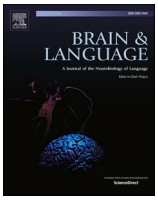
Brain & Language 223 (2021) 105031

Contents lists available at ScienceDirect

Brain and Language

journal homepage: www.elsevier.com/locate/b&l



Idiosyncratic use of bottom-up and top-down information leads to differences in speech perception flexibility: Converging evidence from ERPs and eye-tracking

Efthymia C. Kapnoula [a,b,c,*], Bob McMurray [a,b,d,e]

[a] Dept. of Psychological and Brain Sciences, University of Iowa, United States
[b] DeLTA Center, University of Iowa, United States
[c] Basque Center on Cognition, Brain and Language, Spain
[d] Dept. of Communication Sciences and Disorders, DeLTA Center, University of Iowa, United States
[e] Dept. of Linguistics, DeLTA Center, University of Iowa, United States

ARTICLE INFO

Keywords:
Speech perception
Categorization
Gradiency
Categorical perception
Individual differences
N100
P300
EEG
Visual World Paradigm
Visual analogue scale

ABSTRACT

Listeners generally categorize speech sounds in a gradient manner. However, recent work, using a visual analogue scaling (VAS) task, suggests that some listeners show more categorical performance, leading to less flexible cue integration and poorer recovery from misperceptions (Kapnoula et al., 2017, 2021). We asked how individual differences in speech gradiency can be reconciled with the well-established gradiency in the modal listener, showing how VAS performance relates to both Visual World Paradigm and EEG measures of gradiency. We also investigated three potential sources of these individual differences: inhibitory control; lexical inhibition; and early cue encoding. We used the N1 ERP component to track pre-categorical encoding of Voice Onset Time (VOT). The N1 linearly tracked VOT, reflecting a fundamentally gradient speech perception; however, for less gradient listeners, this linearity was disrupted near the boundary. Thus, while all listeners are gradient, they may show idiosyncratic encoding of specific cues, affecting downstream processing.

1. Introduction

To perceive speech, listeners map continuous acoustic cues onto categories. For example, voice onset time (VOT) is the delay between the release of the articulators (mouth opening) and the onset of voicing (vocal cords vibrating). It is the primary cue that contrasts /b/ and /p/: in English, VOTs near 0 ms indicate a voiced sound like /b/, and VOTs near 60 ms, a voiceless sound like /p/. VOT varies continuously, even as the subjective percept is more or less discrete. One simple way of mapping continuous cues to discrete categories would be to impose a threshold or boundary; however, VOT varies as a function of place of articulation (Lisker & Abramson, 1964), talker (Allen & Miller, 2004), dialect (Walker, 2020), speaking rate (Miller et al., 1986), and coarticulation (Nearey & Rochet, 1994), and similar factors affect virtually all speech categories. As a result, different boundaries would be needed in different contexts. Thus, to correctly perceive speech, it appears that listeners must be sensitive to fine-grained differences and use context flexibly.

Early views of speech perception postulated that listeners perceive speech sounds categorically. This was motivated by evidence that listeners discriminate sounds from different phoneme categories better than equivalent acoustic differences within the same category (Liberman et al., 1961; Pisoni & Tash, 1974; Repp, 1984; Schouten and van Hessen, 1992). This empirical phenomenon is known as categorical perception (CP) and it was thought to reflect the fact that listeners' perceptual encoding of speech is fundamentally shaped by the speech categories of their language. To be clear, the claim of CP was not just that the categories themselves are discrete, but that listeners encode continuous cues like VOT in a somewhat discrete way (see Fig. 1A), collapsing regions of the acoustic space that fall into the same category. CP was argued to make speech more efficient as it was presumed to derive from rapid –perhaps modular– processes that discard irrelevant variance in the speech signal, and because it offered a stable code to help listeners and speakers attain parity (Liberman et al., 1961).

However, mounting evidence suggests that listeners perceive speech sounds in a continuous or gradient manner. First, a number of studies

* Corresponding author at: Basque Center on Cognition, Brain and Language, Mikeletegi Pasealekua, 69, 20009, Donostia, Gipuzkoa, Spain.
E-mail address: kapnoula@gmail.com (E.C. Kapnoula).

https://doi.org/10.1016/j.bandl.2021.105031
Received 15 March 2021; Received in revised form 29 July 2021; Accepted 22 September 2021
Available online 8 October 2021

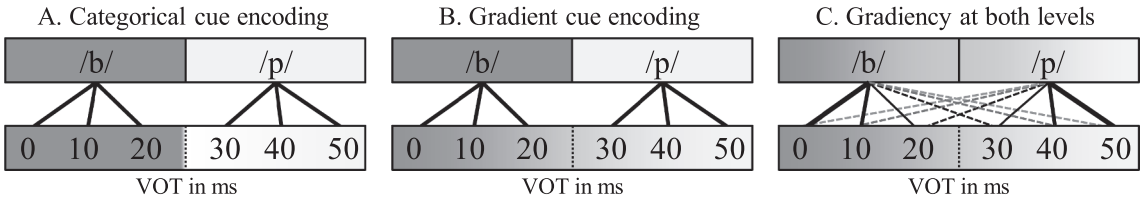


Fig. 1. Examples of gradient and categorical mapping of speech cues. The bottom level represents encoding of a continuous acoustic cue such as VOT. The top level represents speech categories; in classic views this may be phonemes or phonological features, but this could also be syllables or even lexical items. The important thing is that information is more or less discretely represented. A) Categorical encoding of cues leads to a sharp category boundary at both levels. B) Cue encoding is gradient, but speech categories are activated discretely, leading to an abrupt boundary only for the latter. C) Both cue encoding and activation of speech categories are gradient.

suggest that CP may derive from memory demands of discrimination tasks, and when less biased tasks are used, people can discriminate within-category differences as well as between-category differences (Gerrits & Schouten, 2004; Massaro & Cohen, 1983; Schouten et al., 2003). Moreover, as we describe later, event-related potential (ERP) work on the earliest stages of perception, the encoding of continuous speech cues like VOT, suggests a linear—not categorical—mapping (Frye et al., 2007; Sarrett et al., 2020; Toscano et al., 2010). It is possible that even with a gradient cue encoding, categories could still impose discreteness at later levels such as at the phoneme- or lexical-level (e.g., see Fig. 1B). But this too has been ruled out: within-category differences affect encoding at multiple levels: sublexical (e.g., phonemes or syllables; McMurray et al., 2008; Miller, 1997; Samuel, 1982; Toscano et al., 2010), lexical (Andruski et al., 1994; McMurray et al., 2002; Utman et al., 2000), and semantic (Sarrett et al., 2020).

For example, McMurray et al. (2002) used the Visual World Paradigm (VWP) along with VOT continua spanning two words (e.g., *beach/peach,*). They found that continuous differences in VOT lead to gradient activation of competing lexical items: Even when participants clicked on the target (e.g., *beach*), the probability of looking to the competitor (*peach*) was linearly related to VOT; as VOT approached the category boundary (i.e., when the stimulus was more ambiguous), participants had a higher probability of fixating the competitor. This was taken as evidence that words are activated gradiently as a function of VOT, reflecting the probability of an input being the target (see also McMurray et al., 2008).

As a result, it is now commonly accepted that speech perception is fundamentally gradient. This gradiency may be functionally useful, helping listeners to be more flexible in the face of contextual variation for at least two reasons (Clayards et al., 2008; Kapnoula et al., 2021; McMurray et al., 2009). First, differences in VOT due to talker gender, coarticulation, place of articulation, and rate can often be on the order of 5–10 ms; yet listeners have been shown to adjust their boundaries for such factors. In order to accomplish this, they must track fine-grained differences in cues like VOT *within* a category (McMurray & Jongman, 2011).

Second, gradiency may make listeners more flexible when they make an error. For example, McMurray et al. (2009) examined listeners' ability to recover from lexical garden paths: participants heard words like *þarricade*, where the initial sound came from a /b/ to /p/ continuum (i.e., *þ* stands for /b/, /p/, or anything in between). Critically, if the VOT was 40 ms, listeners might initially favor *parakeet,* and then have to revise this decision when *-cade* arrives. However, if the degree of commitment is gradiently tuned to the phonetic detail, when the VOT is 25 ms (still consistent with /p/, but closer to the boundary), they ought to recover faster, because /b/ (and *barricade*) would still be somewhat active. This contrasts with a categorical listener, who would fully activate /p/ (and fully suppress /b/) for both VOTs. A VWP experiment showed evidence consistent with gradient predictions, suggesting that a partial commitment can be helpful for maintaining flexibility (Brown-Schmidt & Toscano, 2017; see also Gwilliams et al., 2018; Szostak & Pitt, 2013).

1.1. Individual differences in speech perception

As we described, there is overwhelming evidence for gradiency in the *modal* listener—the *average* performance in the commonly studied population of normal-hearing, monolingual adults. However, recent studies have also revealed substantial individual differences (for a review, see Yu & Zellou, 2019). This has important consequences for how we think about the necessity and utility of gradiency in achieving a flexible speech perception system.

Kong and Edwards (2016) measured how gradiently listeners categorize speech sounds using a visual analogue scaling (VAS) task (see also Massaro & Cohen, 1983, and Munson et al., 2010). In this task, participants heard speech sounds from a */da/* to */ta/* continuum and responded by clicking on a line to rate how *da*-like or *ta*-like each sound was. Some listeners used the entire scale to respond (reflecting a more linear relationship between VOT and rating), while others used mostly the endpoints of the line, following a more step-like response pattern.

Kapnoula et al. (2017) used a similar method to ask how gradiency is related to other aspects of speech perception and to non-linguistic cognitive processes. They showed that higher gradiency was linked to higher utilization of a secondary cue (see also Kim et al., 2020, and Kong & Edwards, 2011, 2016), pointing to a functional role of gradiency in speech perception. Indeed, a follow-up study by Kapnoula et al. (2021) provides direct support for the idea that such gradiency can be useful; they used a VAS task, measuring listeners' speech categorization gradiency, along with a lexical garden path task modeled after McMurray et al. (2009). They found that all listeners showed a similar level of initial commitment to lexical competitors. However, more gradient listeners were more likely to recover from errors, particularly when the stimulus was acoustically distant from the target (e.g., when the VOT mismatched what would have been expected for that word). This supports the notion that gradiency can make speech perception more flexible, although it also suggests that listeners may vary in the degree to which they adopt this approach.

In sum, listeners vary in how gradient they are in categorizing speech sounds (Kapnoula et al., 2017; Kim et al., 2020; Kong & Edwards, 2016; see also Fuhrmeister & Myers, 2021 for structural MRI evidence) and these differences have functional consequences for the flexibility of speech perception (Kapnoula et al., 2021). The first goal of this study was to further validate the VAS task as a way of documenting this kind of individual differences in speech processing. However, the presence of individual differences also gives rise to other important questions that remain unanswered.

1.2. Individual differences vs. modal gradiency

How can individual differences in gradiency be reconciled with the robust evidence for gradiency in the modal listener (e.g., McMurray et al., 2002)? It is tricky to compare these two lines of research because the tasks that have identified modal gradiency differ substantially from the VAS in important ways. For example, VWP-based measures assess specifically within-category gradiency, which may be a fundamental

aspect of speech perception. Similarly, Toscano et al. (2010) used a P3 ERP component to index late phonological or lexical categorization. They found that –much like in the VWP– the P3 gradiently tracks within-category VOT changes; as the VOT approached the category boundary, the P3 was reduced –even when controlling for the participants' ultimate response.

In contrast, VAS measures reflect gradiency across the entire continuum, including between-category differences close to the boundary, and those may be the regions that show the most variability across listeners. In addition, both the VWP and ERP/P3 paradigms reflect *real-time* differences in the activation of phonological or lexical representations. In contrast, the VAS may rely on listeners' ability to maintain a gradient representation of the signal in memory before responding. Finally, neither the VWP nor the P3 studies examined listeners at an individual level, but instead focused on group-level averages.

A critical inconsistency between these two paradigms is highlighted by a recent study by Kapnoula et al (2021). In this study, individual differences in gradiency (measured with VAS) did *not* moderate the degree of initial commitment to lexical competitors in the garden-path paradigm (e.g., relative commitment to *barricade* vs. *parakeet* based on the VOT). However, gradiency did moderate later recovery. This raises the possibility that the VAS measures something distinct from the lexical processing assessed by the VWP. However, it is also possible that the specifics of that paradigm made it difficult to detect an effect. In the garden-path paradigm, competitors (e.g., *parakeet*, when hearing *barricade*) are only briefly active before disambiguating information arrives. Thus, it may be that competitors were just not active long enough to detect this effect. It is possible that VAS gradiency could be observed in a VWP task that uses a minimal pair continuum (e.g., *bear/pear*) where the lack of disambiguating context allows competitors to remain active for longer.

A study that relates different measures of gradiency in the same individuals may help us understand how individual differences can be reconciled with robust evidence for a fundamentally gradient system. This is the second goal of the our study, where we deploy a VAS measure following Kapnoula et al. (2017), a VWP task similar to McMurray et al. (2002), and a P3/ERP paradigm based on Toscano et al. (2010).

1.3. Sources of gradiency in speech categorization

Our third goal was to assess a set of possible sources of individual differences in speech gradiency. We considered four potential sources: secondary cue use, non-linguistic cognitive differences, lexical competition, and early encoding of acoustic cues.

1.3.1. Cue integration

In searching for the source of individual differences in gradient speech perception, an obvious direction is the well-established link between gradiency and multiple cue integration (Kapnoula et al., 2017; Kim et al., 2020; Kong & Edwards, 2016; Ou et al., 2021). As described above, there is a positive correlation between gradiency and secondary cue use. While multiple cue integration has often been seen as an *outcome* of gradiency, it could also reflect a causal link in the other direction. For example, the ability to integrate multiple cues may help listeners form a more precise estimate of the degree to which the input varies continuously between two phonemes.

Importantly, this positive relationship has only been found for some sets of cues: VOT/F_0 (Kapnoula et al., 2017, 2021; Kong & Edwards, 2016), and formant frequency/vowel length (Kim et al., 2020; Ou et al., 2021); but not others such as VOT/vowel length (Kapnoula et al., 2017) and frication spectrum/formant transitions (Kapnoula et al., 2021). Independently of the reason behind this discrepancy (which remains unclear), this pattern speaks against the idea that differences in cue integration drive differences in gradiency; we see evidence that cues may not be (well) integrated, but perception is still gradient, or cues may be adequately integrated, but cue integration does not predict gradiency. Thus, it seems more likely that the causal link is in the opposite direction (i.e., as originally thought): higher gradiency allows listeners to be more sensitive to small acoustic differences, permitting better cue integration (though gradiency is clearly not sufficient to predict integration in all cases[1]). Alternatively, a gradient representation could also help listeners avoid making a strong commitment based on one cue, allowing them to use multiple cues more effectively. Lastly, a third factor could drive both. While it was not the goal of the present study to rigorously test this hypothesis, we do include a cue integration measure for replication purposes, and we return to this hypothesis in the Discussion.

1.3.2. General cognitive differences

Previous studies have also examined the link between gradiency and domain-general cognitive differences such as working memory, inhibitory control, cognitive flexibility, and sustained attention (Kapnoula et al., 2017; Kim et al., 2018; Kong & Edwards, 2016). These studies show few correlations, with the exception of a weak link between gradiency and working memory (Kapnoula et al., 2017). This effect could reflect the degree to which within-category information is maintained up to the response stage. That is, there may be high gradiency for all listeners at earlier perceptual stages of encoding, but working memory limitations prevent some listeners from maintaining gradient information long enough to affect their response.

An alternative view is that individual differences in speech gradiency reflect a broader tendency to perceive the world discretely or continuously. This could derive from differences in things like inhibition that affect decision making across domains. To test this hypothesis, Kapnoula et al. (2021) used a VAS task with two speech continua and a visual continuum (apple to pear). They found a very weak relationship between gradiency in the visual and speech tasks, ruling out a general tendency for gradiency (or categoricity). Importantly, they also found a very weak correlation between the two *speech* tasks. This further supports the idea that individual differences in gradiency are not the result of a general trait.

In sum, there is little evidence that non-linguistic, higher cognitive functions drive individual differences in speech gradiency. Instead, it seems more probable that such factors moderate effects of gradiency on downstream language processing (Kapnoula, 2016). Even so, we continued this investigation by including (a) a spatial Stroop task, to extend the assessment of cognitive control, and (b) a visual VAS task, as a control task to rule out that VAS gradiency is due to domain-independent categorization gradiency.

1.3.3. Lexical competition

A third possibility is that individual differences in speech perception gradiency are driven by the dynamics of lexical competition. Words suppress their competitors during spoken word recognition (Dahan et al., 2001; Luce & Pisoni, 1998) and this lateral or lexical inhibition may help "sharpen" decisions between words, committing more strongly to the target over competitors. There is also evidence for feedback from lexical to sublexical levels of processing (Elman & McClelland, 1988; Getz & Toscano, 2019a; Luthra et al., 2021; Magnuson et al., 2003; NOE & Fischer-Baum, 2020; Sarrett et al., 2020). This top-down flow of information influences speech perception in real time (Magnuson et al., 2003; McClelland et al., 2006) and drives perceptual learning (Davis et al., 2005; Kraljic & Samuel, 2006; Leach & Samuel, 2007; but see Norris et al., 2000).

The combination of these mechanisms raises the possibility that sharpening at the lexical level (via lateral inhibition) cascades via

[1] The discrepancy between findings in whether they show a link between gradiency and cue integration could be due to differences in temporal proximity across different cue pairs. That is, VOT and vowel length, as well as frication and transition are temporally separated, which may hinder their integration.

feedback to sharpen categorization at lower (sublexical) levels. Stronger lateral inhibition may lead to greater and/or faster suppression of competing words, which in turn leads to greater and/or faster deactivation of competing phoneme categories. For example, consider a situation in which a partially ambiguous lexical item is heard (e.g., *beach* with a VOT of 15 – a /b/, but near the boundary). In a system with strong inter-lexical inhibition, the more active word (e.g., *beach*) inhibits the competitor (*peach*), leading to faster suppression of /p/. In contrast, a system with weaker inter-lexical inhibition would take more time to settle, allowing gradient activation of more than one phoneme categories in parallel. This mechanism can potentially explain why listeners' degree of gradiency is not correlated across different continua (as found by Kapnoula et al., 2021), as the lexical properties (frequency, cohort density etc.) of the items used to construct the continua could impact the observed gradiency. We test this prediction here by relating VAS gradiency to lexical inhibition using a variant of the Visual World Paradigm.

1.3.4. Cue encoding

Finally, differences in speech categorization gradiency may have an early, perceptual locus. Here we assume for simplicity a two-stage process (Fig. 1): First listeners encode continuous cues (such as VOT and F_0), next these are mapped to speech (e.g., phoneme) categories (and later to lexical representations or task responses). The VAS task measures gradiency at the second level: listeners are not aware of individual speech cues, and the instructions (and the task layout) ask listeners to focus on how /b/-like or /p/-like a sound is.

In this light, there are a few possibilities. If listeners encode cues gradiently, this should allow for either a graded or a categorical activation of phoneme categories, depending on how cues are mapped to categories (Fig. 1B; C). However, if listeners encode cues categorically, this should limit their sensitivity to within-category phoneme differences, which would be reflected in a more categorical/step-like categorization pattern (Fig. 1A). Thus, gradient responding in the VAS task could only reflect gradient cue encoding, whereas categorical responding in the VAS reflects categorical activation at the phoneme level, but it does not tell us much about cue encoding. Two of the prior hypotheses (general cognitive differences and lexical competition) focus on processes that take place downstream from category representations. In contrast, if gradiency reflects differences in early cue encoding, this could explain why listeners can be gradient in one continuum (voicing), but not another (fricative place of articulation; Kapnoula et al., 2021).

The VAS task may not be sufficient to isolate a pre-categorical locus of gradiency. In fact, such representations have been notoriously difficult to directly assess. However, ERP work points to a possibility. Toscano et al. (2010) measured the amplitude of the fronto-central auditory N1, a negative ERP component that is thought to be generated in Heschl's gyrus ~ 100 ms post stimulus onset (see also Sharma et al., 2000; Sharma & Dorman, 1999). They presented stimuli that varied continuously in VOT (*beach*-to-*peach* and *dart*-to-*tart* continua) and observed a linear relationship between VOT and N1 amplitude with lower VOTs triggering larger N1. To rule out the possibility that this pattern of results was an artifact of averaging across participants with different boundaries, they also compared two mixed effects models: a linear and a categorical model, taking into account any differences between individual participants' category boundaries. In line with their prediction, the linear model was a better fit to the data, pointing to a fundamentally gradient encoding at the cue level. This N1 effect has been replicated several times (Getz & Toscano, 2019a, 2019b; Sarrett et al., 2020; see Getz and Toscano, 2021 for a review and Frye et al., 2007 for analogous findings in the M100).

In the present study, the N1 offers a useful tool for asking whether individual differences in speech categorization reflect differences in the early perception of acoustic cues. For example, individuals that give more gradient ratings may show a more linear relationship between VOT and N1 amplitude, whereas more categorical listeners may show more of a step function.

1.4. Present study

The present study has three aims: (1) to further validate the VAS paradigm as a tool for measuring individual differences in speech categorization gradiency; (2) to ask how these individual differences in gradiency map to the robust evidence for sensitivity to within-category acoustic differences in the modal listener (as seen in VWP tasks and in the ERP/P3 paradigm); and (3) to assess a set of factors as possible sources of speech perception gradiency.

First, we extracted a measure of how gradiently each participant responded in the VAS task across two speech (voicing) and one visual continuum. As in previous work, we also assess secondary cue use, aiming at replicating the well-established positive link between voicing gradiency and integration of VOT and F_0 (Kapnoula et al., 2017, 2021; Kim et al., 2020; Kong & Edwards, 2016). This set of tasks allowed us to further validate the VAS paradigm and use it to document the presence of individual differences in *speech categorization gradiency*.

Second, we asked whether VAS gradiency was correlated to performance in two experimental tasks that have been used to document gradiency in the modal listener. We assessed sensitivity to within-category differences using a VWP task like that of McMurray et al. (2002) with continua based on real words (e.g., *bear/pear*). The intuitive prediction is that the two measures would be related, with listeners who show more gradiency in the VAS also showing a more gradient pattern of competitor activation in the VWP. However, it was unclear if we would find this. For example, in Kapnoula et al.'s (2021) lexical garden path VWP task, the degree of initial commitment to a lexical competitor was gradient in all listeners independent of their gradiency in the VAS task. So, if we observe a relationship between the VWP and VAS measures of gradiency, this would suggest that this null effect may have been due to the unique nature of the stimuli in that garden-path paradigm. In addition to the VWP task, we also assessed individual differences in the P3 ERP component, which offers an alternative measure of gradiency at the level of speech categories. This was intended to provide additional converging evidence related to our second research aim.

Our third goal was to examine what makes listeners more or less gradient when categorizing speech sounds. As previewed in the Introduction, we focused on three possible sources related to perceptual, cognitive, and language processing.

First, given the weak but seemingly persistent link between gradiency and non-linguistic processes, it seemed prudent to continue this investigation. Here, we used the spatial Stroop to assess *inhibitory control*. Among the different aspects of cognitive control, inhibitory control is particularly relevant for suppressing competing responses and sharpening decisions. It is currently unclear whether domain-general inhibitory control contributes to the resolution of competition in spoken word recognition (Zhang & Samuel, 2018; Zhao et al., 2020; but see dissociation between "automatic"/"obligatory" inhibition and attention-based inhibition in Burke & Shafto, 2008). Nonetheless, domain-general inhibition could play a similar role in promoting gradiency. Two prior studies (Kapnoula et al., 2017; Kim et al., 2020) assessed this (using different tasks), but neither observed a correlation between speech perception gradiency and inhibitory control. However, given concerns about the reliability of cognitive control tasks (Enkavi et al., 2019), it was worth assessing this with a new task.

Second, we assessed *lexical inhibition*. As described above, weaker lexical inhibition could help maintain parallel activation of targets (e.g., *beach*) and competitors (*peach*), leading to longer-lasting gradient activation of both speech categories (/b/ and /p/). To test this, we used a task designed to assess inhibition between words (Dahan et al., 2001; Kapnoula & McMurray, 2016). Coarticulatory information in the first two phonemes of a target word (e.g., *net*) is manipulated to briefly boost a competitor (e.g., ne_k boosts *neck*), which, in turn, inhibits the target. Consequently, when the final phoneme (/t/) is heard, the target (*net*)

Table 1
Order and descriptions of tasks.

Order: Day	Task	Duration (min)	Construct	Measure (s)	Research aim*
1: 1	Phoneme VAS	15	Speech categorization gradiency	VAS slope	All
			Secondary cue use	Theta angle	1
1: 1	Visual VAS	10	Domain-general categorization gradiency	VAS slope	1
2: 1	Spatial Stroop	5	Inhibitory control	Congruency	3
3: 1	*net-neck* VWP	20	Lexical inhibition	Splice effect	3
4: 1	*beach-peach* VWP	25	Within-category lexical gradiency	Competitor looks	2
5: 2	EEG/ERP	90	Perceptual encoding of speech cues	N1	3
			Within-category speech gradiency	P3	2

* (1) Validation of the VAS paradigm; (2) relationship between VAS gradiency and within-category sensitivity in the modal listener; and (3) potential sources of gradiency.

may have a hard time being activated. We used this task to extract a measure of the overall strength of inter-lexical inhibition within an individual (Kapnoula & McMurray, 2016; Li et al., 2019).

Lastly, we asked whether differences in gradiency are due to differences in the *early encoding of acoustic cues*. We used the same ERP paradigm as Toscano et al. (2010) to assess how individuals perceive a primary acoustic cue (VOT; i.e., more or less gradiently). Given that N1 generally reflects encoding of VOT in a continuous/linear way, we asked whether this relationship is distorted for categorical listeners. That is, if we find a link between listeners' response pattern in the VAS task and their encoding of VOT (as reflected in the N1), this would suggest that gradiency has an early perceptual locus.

2. Methods

2.1. Participants

Seventy-one (71) monolingual American English speakers participated in this study (age: 25.4 ± 4.7, 33 male). They had typical hearing, normal/corrected-to-normal vision, and no neurological disorders. Participants received monetary compensation for their participation in the study, and underwent informed consent in accord with University of Iowa IRB policies. Because of technical problems not all subjects had data for all tasks: this averaged between two and 13 participants depending on the analysis and is described with each analysis in the results. A minimum detectable effect analysis suggested that with 71 subjects, and assuming $\alpha = 0.05$, this sample should be sufficient to detect a correlation greater than 0.320 with 80% power. This effect size is in line with our prior work.

2.2. Design and overview of tasks

Participants came to the lab twice to perform five tasks assessing different forms of speech gradiency and its possible sources. Table 1 lists the tasks in the order in which they were administered along with relevant information. In the interest of consistency and presentation ease, the Methods section describes tasks in the order in which they were administered, while the Results section is ordered by research question.

The visual analogue scaling task (VAS; Kapnoula et al., 2017, 2021; Kong & Edwards, 2011, 2016; Munson and Urberg Carlson, 2016; Schellinger et al., 2008) assessed speech categorization gradiency using two VOT × F_0 continua (/b/ to /p/ and /d/ to /t/). From this task, we also extracted a measure of secondary cue use to assess its relationship to speech gradiency. Following Kapnoula et al. (2021), we included a visual version of the VAS task using an *apple/pear* continuum to assess participants' overall tendency to use the endpoints versus the entire line. Since we used the VAS to extract our principle measure of gradiency, we ran it first to minimize any contamination/fatigue effects from other tasks.

Next, we administered a spatial version of the Stroop task (Stroop, 1935) to assess domain-general inhibitory control. Then, we used a VWP variant of the subphonemic mismatch paradigm (Dahan et al., 2001; Marslen-Wilson & Warren, 1994) to assess lexical inhibition. At the end of the first session, we assessed gradiency at the lexical level using a VWP task with speech continua (McMurray et al., 2002). The two VWP tasks were done consecutively to minimize the time participants wore the head-mounted eye-tracker. The Stroop was performed second rather than last because we thought it was more likely to be susceptible to fatigue.

Participants came to a different lab on a different day to perform the ERP task. During this session, we collected electrophysiological responses to stimuli varying in voicing using an ERP paradigm developed by Toscano et al (2010; see also Sarrett et al., 2020, and Getz and Toscano, 2021, for a review), which allowed us to estimate individual differences in speech processing; the N1 was used to index early cue encoding, and the P3 to index within-category speech gradiency (as a converging measure to the *beach-peach* VWP).

2.3. Speech gradiency and secondary cue use (VAS)

2.3.1. Logic and design

In the VAS task, participants used a continuous scale to rate tokens from two speech and one visual continuum. The three sets were presented in separate blocks the order of which was counterbalanced between participants.

2.3.2. Stimuli

Speech stimuli were two continua, one labial-onset: *bill-pill*, and one alveolar-onset: *den-ten*. For each set, we constructed a 7 VOT × 5 F_0 continuum. All stimuli were based on natural speech recordings spoken by a male monolingual speaker of American English. First, for each word pair, we extracted the pitch contour of the voiced endpoint (*bill* and *den*). We then constructed two new contours of identical shape that were shifted upwards and downwards so that the mean pitch would be 95 Hz and 145 Hz respectively, giving us four new contours (2 words × 2 F_0). For each word, the two extreme contours were used as endpoints to create three intermediate pitch contour steps. The resulting pitch contours were approximately 12.5 Hz apart. We then replaced the original contours of the two words using the pitch-synchronous overlap-add (PSOLA) algorithm in Praat (Boersma & Weenink, 2016). This yielded 10 new items (2 words × 5 F_0). Next, we constructed a voicing continuum for each pair of words using the progressive cross-splicing method described by Andruski et al. (1994) and McMurray et al. (2008); progressively longer portions of the onset of the voiced sound were replaced with analogous amounts taken from the aspirated period of the corresponding voiceless sound. VOT steps varied from 7 ms to 43 ms and were 6 ms apart. Each continuum step was presented 3 times, resulting in 210 trials (7 VOTs × 5 F_0s × 2 words × 3 reps).

Visual gradiency was assessed using a two-dimensional *apple/pear* continuum spanning color and shape in a 7 × 5 matrix. The endpoints of the visual continuum were two pictures downloaded from a commercial clipart database. These were edited to intensify prototypical characteristics. To manipulate the shape dimension, we morphed these pictures using the Fantamorph (ver. 5) software to create a seven step

continuum. These were then recolored in a five step color continuum from yellow-ish (prototypical *pear*) to red (prototypical *apple*). Each picture was presented five times, resulting in 175 trials.

2.3.3. Procedure

Participants saw a line labeled with one word at each end. They listened to or saw each stimulus and clicked on the line to indicate the corresponding position of the stimulus. As soon as they clicked, a rectangular bar appeared at that location and then they could either change their response (by clicking elsewhere) or press the space bar to verify it. The task took approximately 25 min.

2.3.4. Quantifying gradiency and secondary cue use

As in Kapnoula et al (2017, 2021), we used the rotated logistic function (Eq. (1)) to fit participants' VAS responses. Unlike standard logistic regression, this provides orthogonal measures of gradiency and secondary cue use (i.e., use of F_0).

$$p(resp) = b_1 + \frac{(b_2 - b_1)}{1+e^{\left(\frac{-4\cdot s\cdot 2\cdot \nu(\theta)}{(b_2-b_1)}\right)\cdot\left(\frac{\tan(\theta)\cdot(x_0-VOT)-F_0}{\sqrt{1+\tan(\theta)^2}}\right)}} \quad (1)$$

As in the four-parameter logistic, b_1 and b_2 are the lower and upper asymptotes. The category boundary is handled differently. The rotated logistic assumes a diagonal boundary in a two-dimensional (VOT × F_0) space that is described as a line with some crossover point (along the primary cue, VOT) and some angle, θ. A θ of 90° indicates exclusive use of the primary cue (the x axis), while a θ of 45° reflects use of both cues. After the boundary vector is identified, this equation rotates the coordinate space to be orthogonal to this boundary (the tan(θ) term) and the slope (s) of the function is then perpendicular to the diagonal boundary. Lastly, $\nu(\theta)$ switches the direction of the function, if θ is less than 90, to keep the function continuous. This function is superior to the standard logistic in that it 1) allows for asymptotes that are not 0/1; 2) does not conflate the boundary along each dimension and the slope; and 3) allows a single estimate of slope that pools across both dimensions.

This function was used to quantify: 1) gradiency, reflected by the slope parameter which describes the derivative of the function orthogonal to the (diagonal) boundary, with steeper slopes indicating more step-like responses, and 2) secondary cue use which is reflected in the θ parameter, where proximity to 90° indicates lower secondary cue use. Our prior work has shown that the θ parameter is mathematically independent of the slope (Kapnoula et al., 2017, Supplement S2) and it is correlated with an independent measure of secondary cue use (Kapnoula et al., 2017, Supplement S6).

The equation was fit to each participant's VAS responses using a constrained gradient descent method implemented in Matlab that minimized the least squared error (free software available at McMurray, 2017). Fits were good ($R^2 = 0.96$ and $R^2 = 0.97$ respectively[2]).

2.4. Inhibitory control (Spatial Stroop)

2.4.1. Logic and design

To assess inhibitory control independently of language, we adopted a spatial variant of the Stroop task (Wühr, 2007). Participants saw an arrow, located on the left/right side of the screen and pointing to the left or right and responded based on the direction of the arrow (ignoring the irrelevant cue, the side of the screen). Individuals tend to respond faster and more accurately when the direction of the arrow is congruent with its location and the magnitude of the congruency effect reflects the individual's ability to suppress irrelevant information (i.e., the location of the arrow). To intensify the congruence effect, we used 64 congruent and 32 incongruent trials (Logan & Zbrodoff, 1979).

2.4.2. Procedure

At the beginning of each trial, a fixation point appeared at the screen center for 200 ms. The point then disappeared, and an arrow was presented on one side (left/right). Arrows were 300 × 150 pixels in size and presented 100 pixels away from the corresponding edge of the display (centered vertically) on a 19″ monitor operating at 1280 × 1024 resolution. The arrow stayed on the screen until the participant pressed one of two keys (left/right) to report which direction the arrow was pointing to. After the response, there was a 1,000 ms pause (blank screen), at the end of which the next trial began. The task took approximately 5 min.

2.4.3. Quantifying inhibitory control

Inhibition in the Stroop task was quantified as the difference in RT between the congruent and incongruent conditions, excluding incorrect trials.

2.5. Inter-lexical inhibition (net-neck VWP)

2.5.1. Logic and design

The first VWP task assessed individual differences in lexical inhibition. Following previous experiments (Dahan et al., 2001; Kapnoula & McMurray, 2016; Marslen-Wilson & Warren, 1994), we manipulated auditory stimuli such that the onset of each word was either consistent (matching) with the coda consonant or temporarily boosted activation for a competitor (to observe its inhibition on the target). Each target word (e.g., *net*) appeared in three conditions. In the *matching-splice* condition, both the onset (*ne*) and release burst (*t*) came from the same word ($ne_t t$), though from different recordings. This should lead to rapid activation of the correct word. In the *word-splice* condition, the onset of a competing word (e.g., *ne-* from *neck*) was spliced onto the release burst of the target word ($ne_{ck}t$). This should briefly over-activate the competitor (*neck*), and inhibit the target (*net*); then, once the release burst arrives, it would be more difficult to fully activate the target (due to its prior inhibition). To ensure that inhibition effects in the *word-splice* condition were not simply due to the cross-spliced stimuli being poorer exemplars of the target, we also included the *nonword-splice* condition ($ne_p t$). Here, the onset of the stimulus was taken from a nonword such that a bottom-up mismatch is still present, but the onset does not activate a competitor.

We used the Visual World Paradigm (VWP) to measure target activation over time for each splicing condition. Participants saw four pictures (a picture of the target, e.g., *net*, along with two unrelated words and one word with an initial-phoneme overlap with the target, e.g., *nurse*). The competitor was never displayed. There were 28 target-competitor pairs (e.g., *net-neck*). Participants heard each of the four words in all three splice conditions. Therefore, each set of four pictures was presented 12 times (4 pictures × 3 splice conditions), and each picture in a set had an equal probability of being the target.

2.5.2. Stimuli

To construct the auditory stimuli, we excised the coda release burst from each target word (e.g., *net*), starting at the onset of the burst and until the end of the recording, and spliced it onto the onset portion[3] of: (1) another recording of the target ($ne_t t$), (2) its competitor ($ne_{ck}t$), and (3) the nonword ($ne_p t$); the full list of experimental triplets is presented in Table A.1 in the Appendix. Stimuli were recorded by a male native speaker of American English (different than the speaker used to record the VAS stimuli) in a sound attenuated room at 44,100 Hz. The splice point was at the zero crossing closest to the release onset. We also created three spliced versions of each filler word with each target word spliced with itself and one of two nonwords.

A total of 112 pictures (28 target words × 4 pictures in each set) were

[2] Five fit sets (3 labials and 2 alveolars) were excluded due to problematic fits.

[3] The onset was taken from the beginning of each recording up to the release, and thus included the closure.

Table 2
List of stimuli presented in the within-category lexical gradiency task.

Set	Labials		Alveolars	
	Voiced	Voiceless	Voiced	Voiceless
1	bath	path	deer	tear
2	beach	peach	drain	train
3	bear	pear	dot	tot
4	bees	peas	dent	tent
5	bowl	pole	dart	tart

developed using a standard lab procedure (Apfelbaum et al., 2011). A full list of the visual stimuli is presented in Table A.2 in the Appendix. For each word, we downloaded 5–10 candidate images from a clipart database, which were viewed by a group of 3–5 lab members. One image was selected and was subsequently edited to ensure a prototypical depiction of the target word. The final images were approved by a lab member with extensive experience using this paradigm.

2.5.3. Procedure

Participants were first familiarized with the 112 pictures by seeing each picture along with its orthographic label. Then they were fitted with the eye-tracker. After calibration, participants were given instructions for the task.

At the beginning of the trial, participants saw a red circle at the screen center along with four pictures in the corners of a 19″ monitor operating at 1280 × 1024 resolution. At this point, participants could briefly look at the pictures before hearing anything, thus minimizing eye movements due to visual search rather than lexical processing (see Apfelbaum et al., 2021, for discussion and validation of this approach). After 500 ms, the circle turned blue, prompting the participant to click on it to start the trial. The circle then disappeared, and the target was played. Participants clicked on the picture that matched the word they heard, and the trial ended. There was no time limit and participants were encouraged to take their time and perform the task as naturally as possible. Participants typically responded in less than 2 secs ($M = 1{,}216$ ms, $SD = 109$ ms) and the entire task took approximately 20 min.

2.5.4. Eye-tracking recording and analysis

We recorded eye-movements at 250 Hz using an SR Research Eyelink II head-mounted eye-tracker. Both corneal reflection and pupil were used whenever possible. Participants were calibrated using the standard 9-point Eyelink procedure. Eye movements were automatically parsed into saccades and fixations using the default psychophysical parameters, and adjacent saccades and fixations were combined into a "look" that began at the onset of the saccade and ended at the offset of the fixation (McMurray et al., 2002).

Eye movements were recorded from the onset of the trial to the participants' response (mouse click) and were time-locked to the auditory stimulus onset. This variable-time offset makes it difficult to analyze results late in the time course. To address this, we adopted the object padding approach of many prior studies (Allopenna et al., 1998; McMurray et al., 2002) by setting a fixed trial duration of 2,000 ms (relative to stimulus onset). For trials that ended before 2,000 ms we extended the last eye-movement; trials which were longer than 2,000 ms were truncated. This assumes that the last fixation reflects the word that was "settled on", and therefore should be interpreted as an approximation of the final state of the system and not necessarily what the participant was fixating at that time. For assigning fixations to objects, boundaries around the objects were extended by 100 pixels in order to account for noise and/or head-drift in the eye-track record. This did not result in any overlap between the objects; the neutral space between pictures was 124 pixels vertically and 380 pixels horizontally.

2.5.5. Quantifying inter-lexical inhibition

Inter-lexical inhibition was quantified by first computing the average proportion of fixations to the target between 600 ms and 1,600 ms post stimulus onset (logit-transformed); with inhibition reflected in the difference in target fixations between the *word-splice* and *nonword-splice* conditions.

2.6. Within-category lexical gradiency (beach-peach VWP)

2.6.1. Logic and design

This VWP task assessed the degree to which lexical activation is sensitive to within-category differences in VOT (McMurray et al., 2002). Listeners heard a token from a speech continuum (e.g., *beach/peach*) and selected the corresponding picture in a VWP task. Critically, to assess specifically within-category sensitivity, our analysis quantifies the degree to which competitor fixations are sensitive to VOT after accounting for each listener's own boundary and ultimate response.

2.6.2. Stimuli

Stimuli consisted of 10 monosyllable CVC word pairs beginning with a stop consonant (Table 2); five labial and five alveolar. The two words in each pair were identical except for the voicing of the initial consonant (e.g., *bear-pear*). Auditory stimuli were recorded in a sound attenuated room at 44,100 Hz by the same male native speaker of American English as the one used for the VAS stimuli. For each of the 10 minimal pairs, we constructed a 7 VOT × 2 F_0 continuum following the same procedures used to make the continua for the VAS task. This resulted in 140 auditory stimuli (10 pairs × 2 F_0 × 7 VOT).

Each labial-initial pair was paired with an alveolar-initial pair with a different vowel, making a quadruplet (e.g., *bath-path*, *deer-tear*; see Table 2); this allowed the d/t items to serve as unrelated items for the b/p continua and vice versa—a more efficient design than that of McMurray et al. (2002). The four images corresponding to the each of the items in a quadruplet were presented together throughout the task and across participants. Visual stimuli consisted of 20 pictures (each corresponding to one stimulus in Table 2) developed using the procedure described above (Apfelbaum et al., 2011).

2.6.3. Procedure

Procedures were identical to the VWP task described above. Participants typically responded in less than 2 secs ($M = 1{,}038$ ms, $SD = 105$ ms) and the task took approximately 25 min.

2.6.4. Quantifying within-category sensitivity

On each trial, the image that the participant selected was treated as the target (e.g., *beach*). Sensitivity to within-category information was quantified based on the average fixations to its competitor (*peach*) from 300 ms to 2,000 as a function of VOT. In this scheme, if listeners are sensitive to within-category detail, competitor fixations should rise as VOT approaches the boundary, and the slope of the function is therefore a useful metric of gradiency.

2.7. Early cue encoding and within-category speech gradiency (EEG/ERP)

2.7.1. Logic and design

This task assessed individual differences in brain responses to continuous acoustic cues. We used the ERP paradigm of Toscano et al. (2010), where early encoding of VOT is linearly reflected in the amplitude of the N1 ERP component, and modal gradiency at the level of categories is observed in the P3. Stimuli were items from two voicing continua (*bill/pill*, *den/ten*). Secondarily, like Toscano et al. (2010), we used the P3 to assess category-level brain responses. Traditionally, the P3 is elicited in "oddball" tasks, in which participants respond to infrequent targets (Polich & Criado, 2006). Thus, participants responded with one button if they heard a target word (e.g., *bill*) and a different button for any of the other three (*pill*, *den*, *ten*). Since all stimuli were equally frequent, participants were expected to make a "target" response

on approximately 25% of trials, and a "non-target" response on about 75% of trials. The target word rotated through each of the four words across blocks.

2.7.2. Stimuli

Stimuli were based on the same recordings and stimulus construction steps as in the VAS task; however, here we only used the two extreme F_0 values (95 and 145 Hz). This allowed us to increase the number of VOT steps from seven to nine (4.5 ms apart), thus capturing more precisely the location of each listener's category boundary along the VOT dimension. Each of the four words (*bill*, *pill*, *den*, *ten*) served as the target on a different block of trials. Each stimulus × target word × target location combination was repeated seven times. Therefore, each of the 36 (2 continua × 9 VOT steps × 2 F_0 steps) auditory stimuli was presented 56 times (4 target words × 2 target locations × 7 repetitions), giving a total of 2,016 trials. The trials were split into eight blocks of 252 trials, one block for each of the 4 target × 2 target location conditions. The order of blocks was pseudorandomized but constant for all participants.

2.7.3. Procedure

Participants were seated in a grounded and electrically-shielded booth and the EEG recording equipment was set up. Next, electrode impedances were minimized, and the earphones were inserted. Preparation took approximately 30 min. At the beginning of the task, participants read the instructions and performed a few trials to familiarize themselves with the task, while the experimenter remained outside the booth and monitored their responses to ensure they performed the task as instructed. After practice, they began the task.

Auditory stimuli were presented over earphones (ER-1 by Etymotic Research) connected to an amplifier located outside the booth. Instructions and visual stimuli were presented on a computer monitor located approximately 75 cm in front of the participant. Instructions, stimulus presentation, and sending of event codes to the EEG amplifier were handled by Presentation® (by Neurobehavioral Systems).

At the beginning of each trial, participants saw a black fixation cross at the center of the screen. The cross stayed on the screen for 700–1,300 ms (jittered) and then they heard a word over the earphones. After the word was played, there was a 200 ms silence and then the cross was replaced with a green circle and two words, one each side of the circle, indicating which button corresponded to which response. One word was always the target for that block (e.g., *bill*) and the other was the word *other*. Participants had 2,000 ms to make a response (by pressing one of the two buttons) and the trial ended. The next trial began 200 ms later.

Average trial duration (including RT) was ~ 2,240 ms. With 2,016 total trials, the task took approximately 75 min. Participants were given an opportunity for a break every 36 trials and were encouraged to take a break and ask for water half-way through the experiment. They usually completed the task within 90 min.

2.7.4. EEG recording and preprocessing

ERPs were recorded from 32 electrode sites (International 10–20 System). EEG channels were collected using the reference-free acquisition provided by Brain Products actiCHamp and were referenced to the average of the two mastoids after recording. Horizontal electrooculogram (EOG) recordings were collected via two electrodes located approximately 1 cm lateral to the external canthus of each eye. Vertical EOG recordings were made using an electrode located approximately 1 cm below the lower eyelid of the left eye. Recordings were made with the Brain Products actiCHamp amplifier system at 500 Hz. Reception and storing of the recordings, as well as linking them to the event codes sent by Presentation® were handled by Brain Vision PyCorder. No filter was applied during recording.

Data were analyzed using Brain Vision Analyzer 2. A 1 Hz 48 dB/octave low cut-off filter, a 30 Hz 48 dB/octave high cut-off filter, and a 60 Hz notch filter were applied to the data prior to processing. We first removed eye blinks by checking the three EOG channels starting at 400 ms before the stimulus onset to 900 ms after. If voltage shifted by more than 50 μV/ms (in either direction), or if voltage shifted by more than 75 μV (in either direction) within any 100 ms part of that segment, then that part (as well as 100 ms before and 100 ms after) was marked as bad. Any segments containing marked portions were excluded from further processing.

To remove other artifacts (e.g., due to movement, muscle tension, or sweat), we examined all channels from 300 ms before to 800 ms after stimulus onset. For each one, if voltage shifted by more than 50 μV/ms (in either direction), then a marker was placed at the time of the voltage shift and a portion of the segment (200 ms before the marker to 200 ms after that marker) was marked as bad. If voltage shifted by more than 75 μV (in either direction) within any 100 ms portion of that segment, then that part (as well as 100 ms before and 100 ms after) was marked as bad. Lastly, if amplitude was higher than 150 μV or lower than −150 μV, then a marker was placed at the time of the voltage divergence and a portion of the segment (200 ms before the marker to 200 ms after that marker) was marked as bad. This was done with the "individual channel mode" option, which allows us to exclude segments for specific channels (while retaining unaffected channels). On average 7.3% of the trials (i.e., 15 trials) included at least one bad segment (3.9% were blink removals and 3.4% other artifacts). Lastly, each trial was baselined using as a baseline the average voltage within a time window starting 100 ms before the onset of the auditory stimulus up until its onset.

We extracted two measures for analysis: the N1 and the P3 amplitude. For both, we started with a visual inspection of the waveform to identify the time regions over which these deflections were exhibited. This was done averaging across stimuli and subjects, and thus independently of any factors of experimental interest. The N1 time-window was set between 115 and 170 ms post stimulus onset, and the P3 time-window between 400 and 730 ms (see Toscano et al., 2010, for a similar range of 300–800 ms post stimulus onset).

Next, we selected the channels to include in the analyses. Our goal was to identify the channels that showed the characteristic morphology of each component (N1/P3) in the broadest terms; again, this was done independently of VOT and VAS slope (the critical measures here). For the N1, we focused on 20 fronto-central sites (Näätänen & Picton, 1987) and computed the average voltage across all trials over the 115 to 170 ms range to identify the channels showing a negative amplitude on average. This yielded 12 channels: Cz, CP1, CP2, C3, C4, FC2, FC1, CP5, CP6, Fz, FC5, and FC6 (see heat maps in the corresponding analysis). For the P3, we took a similarly broad approach identifying channels that showed an average positive amplitude in the 400 to 730 ms range. Here, we included only trials in which the participant responded that the target word was present, as the P3 generally appears as a positive deflection on infrequent trials (i.e., "target" trials, in our case). Further, we considered only 12 central and parietal sites (i.e., locations that are associated with the P3; Nasman & Rosenfeld, 1990). For five adjacent channels this number was positive (Pz, P7, P3, CP2, and CP1; see corresponding heat maps in results).

2.7.5. Quantifying cue- and category-level gradiency

Two measures were extracted from the EEG data. First, we used the auditory N1 amplitude to capture cue-level gradiency. Here, the N1 should decrease as the VOT increases (i.e., from /b/ to /p/), with gradient individuals showing a more linear VOT/N1 relationship. Second, the P3 was used to capture sensitivity to within-category differences. In this case, more extreme VOTs (i.e., more prototypical tokens) were expected to elicit a higher P3, with this pattern being more robust for gradient individuals due to their higher sensitivity to within-category differences.

2.8. Statistical methods

Statistical analyses were conducted in R (version 4.0.4; R Core Team,

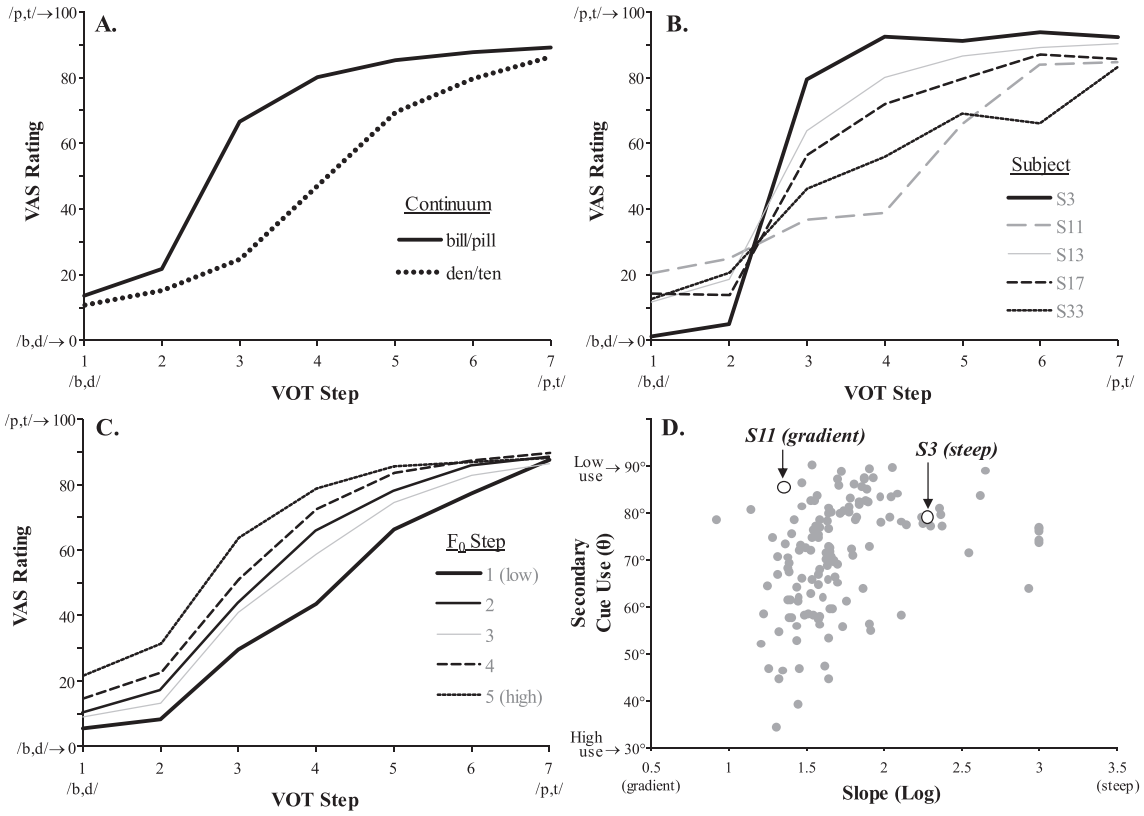


Fig. 2. A) VAS rating as a function of VOT step (averaged across F_0) and continuum. Here, a low VAS rating reflects a more voiced percept (/b/ or /d/). B) VAS rating as a function of VOT (averaged across F_0 steps) for 5 representative subjects. C) VAS rating as a function of both VOT and F_0 averaged across both continua. D) Variability.

2016). For mixed effects analyses, we used the lme4 (version 1.1–26; Bates et al., 2015) and lmerTest (version 3.1–3; Kuznetsova et al., 2020) packages. To determine the random effects included in each model, we compared models with increasing complexity using likelihood ratio tests (LRTs). The most complex random effects structure supported by the data was adopted.

3. Results

Our results are organized into three broad sections, each focusing on one of our three research aims. We start with preliminary analyses of the VAS slopes to document our basic dependent measure, and we replicate the well-established link to cue integration to validate the paradigm. Next, we examine the degree to which individual differences in the VAS slope are reflected in standard measures of gradiency in the modal listener (the VWP and the P3 ERP component). Finally, we assess three potential loci of gradiency: domain-general inhibition, inter-lexical inhibition, and cue-level encoding.

All of these analyses are necessarily correlational, and the complexity of some of the measures (e.g., the VWP and ERP measures) makes it challenging to use them in approaches like hierarchical regression to pinpoint causal factors. Instead, we focus on each task individually, and make cautious inferences about cause based on: (a) the correlation of key measures with VAS gradiency (convergent and discriminant validity), (b) their links to known stages of speech processing, and (c) the broader pattern of results across tasks.

3.1. Documenting phoneme categorization gradiency using the VAS task

We started by verifying that individuals performed the VAS task as expected and that our measure was sensitive to individual variability. Fig. 2A shows VAS responses as a function of VOT for each continuum. While the continua differed (consistent with the well-known effect of place of articulation on VOT), in both cases, responses to low VOTs started near the low extreme of the VAS scale, and transitioned smoothly as VOT increased. Critically, Fig. 2B shows that, as expected, there were marked individual differences, with some subjects showing nearly step-like responses (e.g., S3), and others (S11, S33) showing a more linear response. A closer look at the distribution of slopes (the X axis in Fig. 2D) confirms that there was considerable variation between these two extremes (S3 and S11 are marked with open circles). Finally, Fig. 2C shows the expected effect of secondary cue, with increased voiceless responses for higher F_0s.

We first examined the degree to which our measures of speech categorization gradiency for the two speech contrasts (labial and alveolar) are related to each other. In line with Kapnoula (2016), the VAS slope for labial stimuli was moderately correlated with that of alveolar stimuli, $r(64) = 0.407$, $p = .001$. This correlation was lower than expected, raising the possibility that a composite or average of the two may have reduced sensitivity. Consequently, in our subsequent analyses we used the place-specific VAS slopes, wherever applicable (e.g., in the *beach/peach* VWP task, we related trials with labial stimuli to labial VAS slope and trials with alveolar stimuli to alveolar VAS slope). We used the average of the two slopes to analyze data from more general tasks (e.g., spatial Stroop) and to split participants into gradiency groups. We return to the theoretical importance of this moderate correlation in the Discussion.

Next, we examined the relationship between speech and visual VAS tasks. Visual VAS slope was weakly correlated with labial, $r(67) = 0.208$, $p = .089$, and not correlated with alveolar VAS slope, $r(67) = 0.163$, $p = .182$. These results suggest that participants' tendency to use the entire VAS range (versus the endpoints) is not likely to drive differences in our VAS measure of speech gradiency. To be conservative, we extracted the standardized residual of the speech VAS slopes after partialing out the

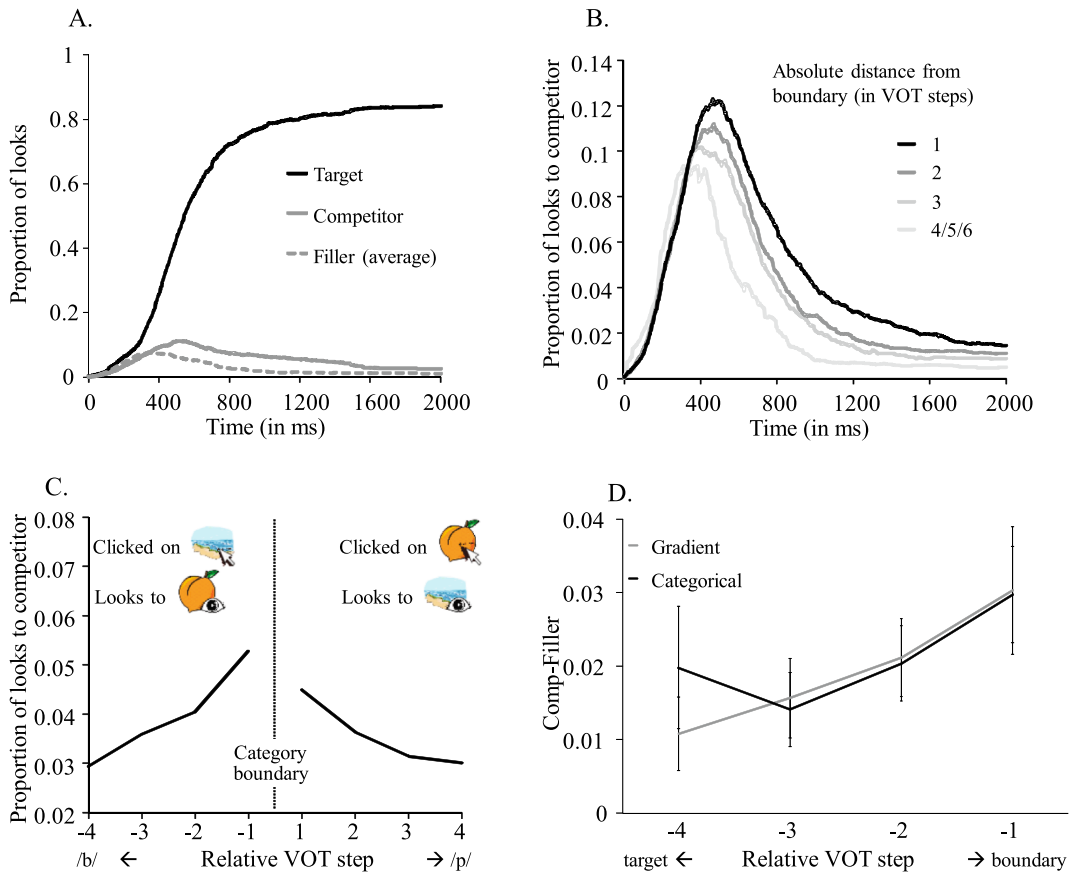


Fig. 3. A) Proportions of looks to the picture of the target, the competitor, and the filler. B) Looks to competitor as a function of (absolute) distance from category boundary. C) Average proportions of looks to the competitor as a function of distance from category boundary per voicing. D) Comp-Filler (i.e., looks to competitor adjusted for overall looking) as a function of distance from the boundary (rVOT) split by VAS gradiency.

variance explained by the visual VAS slope. All analyses on VAS slope were conducted on both measures (raw and residualized slopes), generally leading to identical results. Here, we report the analyses using raw slopes, but provide detailed results for both raw and residualized slopes in the Supplemental analyses corresponding to each primary analysis.

Finally, we assessed the relationship between gradiency (VAS slope) and secondary cue use (the θ parameter estimated from the same task; see Fig. 2C, D). We conducted a hierarchical regression with VAS slope as the dependent variable. Place of articulation (PoA; effect-coded) and secondary cue use (θ angle) were entered as predictors. In the first step, PoA was entered alone and significantly accounted for 10.7% of the variance, $\beta = -0.328$, $F(1,131) = 15.76$, $p < .001$, with higher VAS slopes (lower gradiency) for labial-initial stimuli. In the second step, secondary cue use (θ) was entered, which explained a significant portion of the VAS slope variance, $\beta = 0.277$; $R^2_{change} = 0.045$, $F_{change}(1,130) = 6.97$, $p = .009$. Specifically, higher F_0 use predicted higher gradiency (for complete results see Supplement S1). The results are in line with previous work (Kapnoula et al., 2017; Kim et al., 2020; Kong & Edwards, 2016) in showing that individuals with greater speech gradiency also make greater use of a secondary cue.

As a whole, these results replicate prior work and validate the VAS approach as a way of assessing individual differences in speech categorization gradiency.

3.2. *Relating individual differences in VAS slope to standard measures of gradiency*

We next turned to our second research aim, asking whether and how individual differences in the VAS slope are related to within-category sensitivity in the VWP task and the P3 ERP component.

3.2.1. *Phoneme categorization gradiency and lexical gradiency (VWP)*

The lexical gradiency/ VWP task was modeled after McMurray et al, (2002) to probe the degree to which sensitivity to within-category differences in VOT were reflected at the level of lexical activation. Problems with the eye-tracking led to the exclusion of eight participants from the analyses of fixations (but they were included in the analyses of mouse-click responses). The response functions looked typical with a smooth transition between /b/ and /p/ responses. Statistical analyses of responses are reported in Supplement S2.

Fig. 3A shows the likelihood of fixating the target, competitor, and unrelated fillers as a function of time. Generally, participants looked more to the target, but they also fixated its competitor more than the fillers. Critically, the goal of this task was to assess specifically how *within-category* acoustic differences affect competitor activation. To accomplish this, we took two steps. First, data were split by participants' identification responses (e.g., whether they clicked on the picture of the *beach* or the *peach*). This allowed us to define the target and competitor items in each trial and analyze the data accordingly. Second, if listeners have different category boundaries, a difference between two adjacent VOT steps could be within- category for one subject, but between-categories for another. Thus, as in prior work, we treated VOT relative to each subject's boundary and adjusted for the effects of place of articulation, F_0, and item. The steps for computing relative VOT (rVOT) are presented in the Supplement S3.

We then computed the average competitor fixations from 300 ms to 2,000 ms as a function of rVOT. As expected, even when participants clicked on the target, the proportion of competitor looks increased as the rVOT approached the boundary (see Fig. 3B, C). Lastly, we accounted for

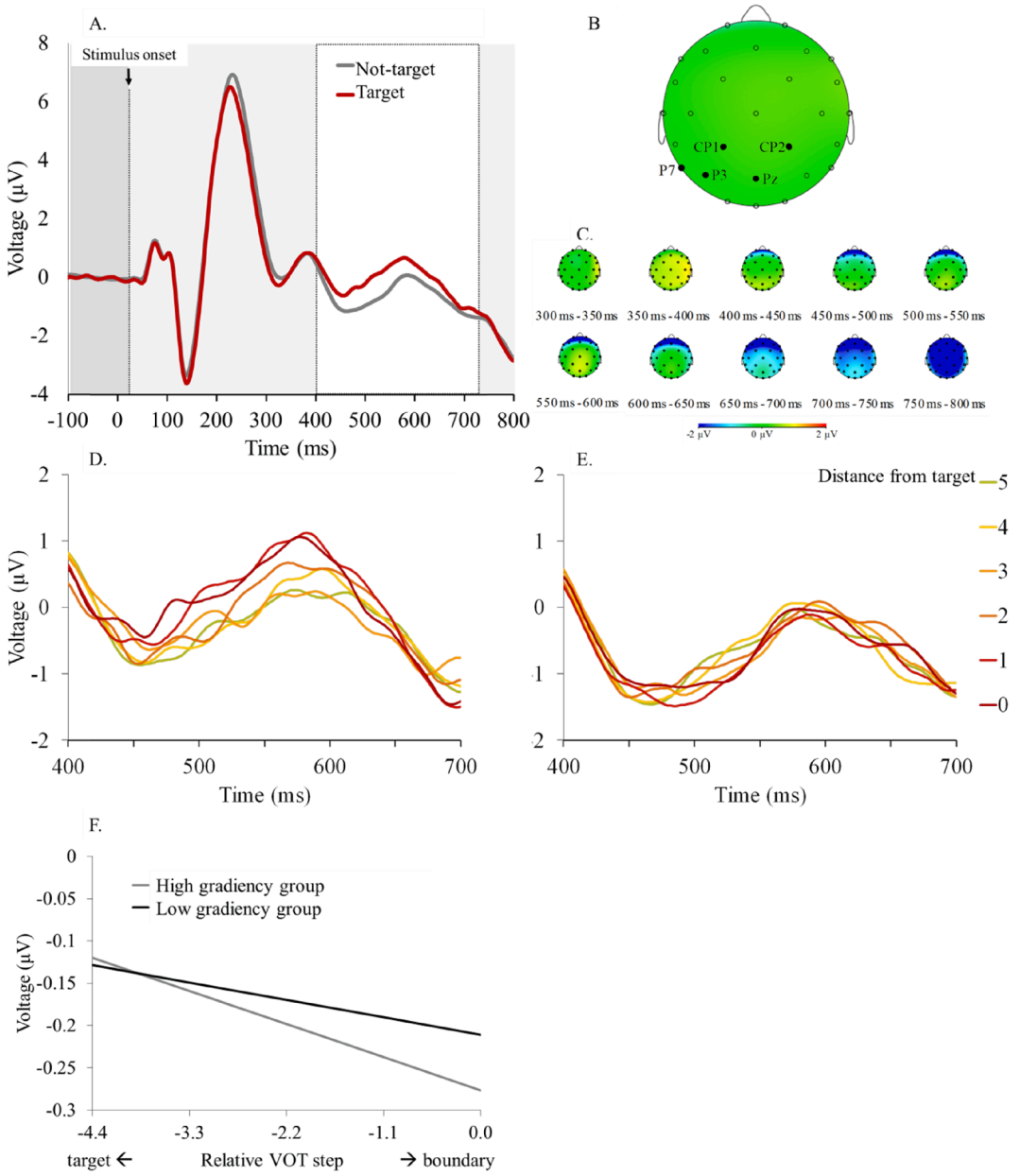


Fig. 4. A) Voltage as a function of time and response. B) Electrode positions. C) Voltage fluctuation as a function of time per electrode site. D) Voltage in time as a function of distance from target for the relevant continuum (i.e., when stimulus PoA matched target) and E) for the irrelevant continuum. F) Model-estimated effect of rVOT per gradiency group on P3.

the possibility that raw looks to the competitor may reflect differences in overall looking; in addition to competitor activation (e.g., looks to *peach* when participants clicked on *beach*), we also calculated the proportion of looks to filler items, and used this difference as our dependent variable (henceforth *Comp-Filler*). This difference-based measure was used to reflect competitor activation independently of individual differences in the overall quantity of fixations to anything.

We then examined the effect of rVOT on this adjusted estimate of competitor fixations using a mixed effects model. To combine both sides of the continuum (voiced and voiceless), rVOT was reversed for voiceless stimuli, so that high rVOT indicated high distance from the target, regardless of whether the target was voiced or voiceless. The fixed effects included rVOT, VAS slope (labial VAS slope, if the target was labial, and alveolar, if it was alveolar), and their interaction. Voicing (voiced/voiceless target), place of articulation (PoA; labial/alveolar target), and their interaction were entered as covariates. All continuous measures were centered. Random effects included random intercepts for subjects and a random slope of rVOT for items (model and results in Supplement S4). There was a significant main effect of rVOT, $B = 0.01$, $t(9.48) = 6.62$, $p < .001$, indicating more competitor looks as stimuli diverged from the target (as expected). There was no main effect of VAS slope, $B = -0.005$, $t(250.7) = -1.55$, $p = .122$, but there was a small but significant rVOT × VAS slope interaction, $B = -0.004$, $t(5871) = -2.25$, $p = .025$, indicating stronger rVOT effect for more gradient listeners (see Fig. 3D, where the function for categorical listeners flattens at extreme rVOTs).

In sum, there was a robust (and expected) effect of rVOT on looks to the competitor across participants, responses, and time windows, according to which the more a stimulus deviated from the target, the more listeners looked to the competitor. This supports the idea that gradient lexical activation is a fundamental aspect of spoken word recognition. In addition, the VAS slope × rVOT interaction suggests that this effect was

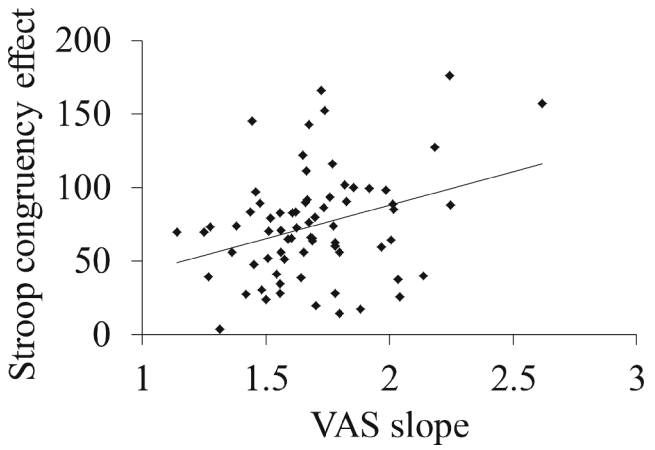


Fig. 5. Relationship between congruency effect (i.e., RT in incongruent trials – RT in congruent trials) and speech gradiency (VAS slope).

moderated by listeners' gradiency in the VAS task, pointing to a larger rVOT effect for gradient listeners.

3.2.2. *Phoneme categorization gradiency and ERP indices of categorization gradiency*

Next, we examined the P3 ERP component, which is thought to be a marker of post-perceptual categorization. Two participants did not return for the second day. Due to a programming error, one participant was not exposed to all conditions and was excluded from analyses. One participant felt discomfort after ~ 10 min in the ERP booth and was let go. Thus, we excluded four participants' data leaving data from 67.

Recall, that the ERP task used an oddball task to elicit the P3. In this task, subjects monitor for one target (e.g., *bill*) and respond "non-target" for the other three stimuli (*pill, den, ten*). Detailed results on the identification task are reported in Supplement S5. The P3 is typically elicited on the infrequent "target" trials (which comprise ~ 25% of trials). Thus, we expected to find a P3-like deflection in trials with a "target" response, and a higher P3 for more prototypical (target-like) stimuli (Toscano et al., 2010). Indeed, a difference was observed between "target" and "other" response trials in the expected direction (see Fig. 4A).

We first examined the P3 as a function of stimulus prototypically (i.e., how target-like it was), with the goal of replicating Toscano et al. (2010), before examining individual differences. The average P3 voltage (from 400 to 730 ms, averaged over Pz, P7, P3, CP2, and CP1; see Fig. 4B,C) was the dependent variable in a linear mixed effects model. Two measures of distance from the target (in VOT and F_0) were entered as fixed effects together with a factor reflecting whether the stimulus and the target matched in place of articulation (PoA; coded as 1 for Match and −1 for Mismatch) and all interactions. Voicing of the target was added as a covariate. The random effects included random VOT and F_0 slopes (and their interaction) for subjects and random channel intercepts. We excluded any cells reflecting 6 or fewer trials to eliminate noise due to the low number of contributing trials (each cell should have had 14 trials: 7 repetitions × 2 target locations). All models and results are reported in Supplement S6.

Distance from the target was significant both when measured in VOT, B = -0.025, t(66) = − 4.06, p < .001, and in F_0, B = -0.083, t(66) = -3.17, p < .001, in the expected negative direction; smaller acoustic distance from the target predicted stronger P3. Moreover, the effect of VOT distance from the target was stronger when the stimulus was relevant to the task (i.e., when the stimulus had the same PoA as the target), as indicated by the significant 2-way interaction, B = -0.024, t(66) = -8.70, p < .001 (compare Fig. 4D, and 4E). These results replicate previous findings showing a negative link between distance from the target (in VOT) and P3 amplitude. They also extend this work by showing that this link also applies to distance from the target in terms of F_0. This suggests that the relationship between distance from target and P3 may hold true independently of how acoustic distance is measured.

Next, we asked whether the effects of VOT on the P3 are modulated by gradiency. To test this, we added place-specific VAS slope and its interactions with distance from the target (in VOT and F_0) and relevancy of continuum to the fixed effects of the model. In this model, VAS slope was significant, B = 0.184, t(45040) = 4.82, p < .001 with the direction of the effect pointing to a negative link between gradiency and P3 amplitude; steep-slope categorizers had overall larger P3s. In addition, the VOT distance × VAS slope interaction was significant, B = 0.047, t(772) = 3.815, p < .001, suggesting a stronger VOT distance effect on P3 for gradient categorizers (Fig. 4F). The main effect of PoA Match as well as the VOT distance × VAS slope × PoA Match interaction were also significant, B = 0.183, t(45560) = 25.73, p < .001, B = 0.028, t(45560) = 3.25, p = .001, respectively. When data were split by PoA Match, the main effects of VOT distance and F_0, as well as the VOT distance × VAS slope interaction were significant only for stimuli matching the target in place of articulation, B = -0.049, t(66) = -5.07, p < .001, B = 0.095, t(64) = -2.89, p = .005, B = 0.123, t(22400) = -2.31, p = .021, B = 0.063, t(977) = 3.55, p < .001, respectively. This pattern was expected, and it validates our paradigm in showing that the P3 is sensitive to the acoustic distance between stimulus and target –in terms of VOT, F_0, and place of articulation (see Fig. 4D; E).

In sum, P3 amplitude was affected by the acoustic distance between stimulus and target (in terms of both VOT and F_0), and this effect was more robust in task-relevant trials (i.e., where the stimulus onset matched the target in place of articulation). Critically, this effect was also modulated by the degree of gradiency as measured in the VAS task; gradient listeners (i.e., with shallower VAS slopes) showed smaller P3 and were more strongly affected by VOT distance from the target, compared to steep-slope categorizers.

3.2.3. *Summary*

The foregoing analyses unify several lines of work by demonstrating that individual differences in speech categorization gradiency that can be robustly detected with the VAS task are also reflected in two prominent approaches that were first used to demonstrate gradiency in the modal or average listener. In the VWP, competitor fixations increase as VOT approaches the category boundary (even accounting for the ultimate response); here we show that the steepness of this increase is enhanced in more gradient listeners. In the P3 ERP paradigm, the P3 is typically strongest at prototypical VOTs, and falls off toward the category boundary. Here we show steeper fall off in more gradient listeners. This pattern suggests that all three measures are tapping something fundamental to how listeners categorize speech.

3.3. *The locus of individual differences in gradiency*

We next turn to our third research question, addressing three hypothesized loci for the observed individual differences in speech categorization.

3.3.1. *Phoneme categorization gradiency and inhibitory control*

Participants performed the spatial Stroop task with high accuracy (*M* = 96.4%, *SD* = 4%) and speed (*M* = 441 ms, *SD* = 67 ms). To assess the congruency effect, we ran paired-samples t-tests with RT and accuracy as dependent measures. Accuracy was logit-transformed. Only correct trials were included in the analyses of RT and trials with RT > 2,000 ms (3 trials) were excluded from both analyses. Participants were significantly faster in congruent trials (*M* = 419 ms, *SD* = 63 ms) compared to incongruent trials (*M* = 492 ms, *SD* = 77 ms), t(70) = 15.52, p < .001. They were also significantly more accurate in congruent (*M* = 99%, *SD* = 2%) than incongruent trials (*M* = 91%, *SD* = 10%), t(70) = 8.28, p < .001. Thus, there was a congruency effect in the expected direction for both accuracy and RT.

We then conducted hierarchical regression to ask whether the

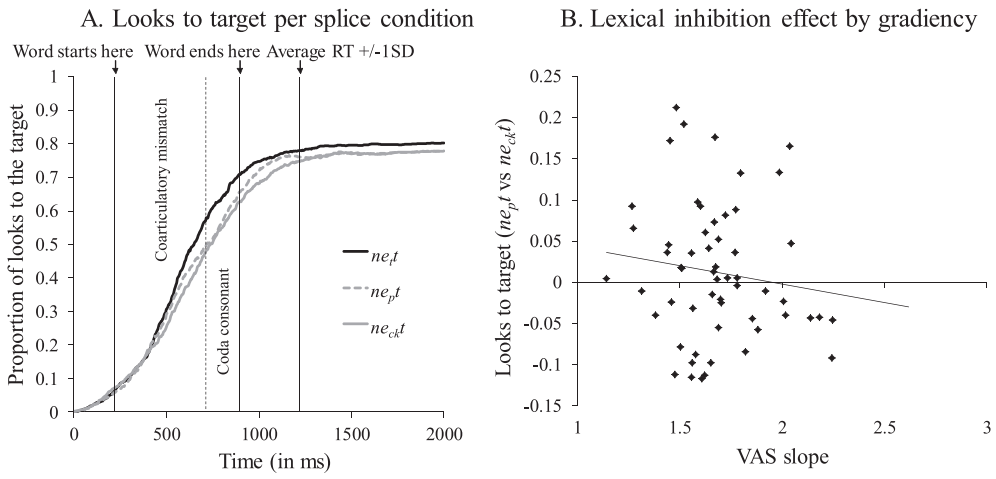


Fig. 6. Looks to the target per splice condition (panel A). Relationship between lexical inhibition effect (i.e., proportion of looks to target in nonword-splice versus word-splice condition) and speech gradiency (VAS slope; panel B).

magnitude of the congruency effect predicted speech gradiency (i.e., VAS slope averaged across labial and alveolar continua). Average Stroop accuracy and RT (across conditions) were added as predictors on the first level to account for effects of overall speed and/or accuracy; congruency was then added on the second step. Overall accuracy and RT accounted for < 1% of the VAS slope variance, $F < 1$. In the second step, the congruency effect accounted for 11.8% of the variance in VAS slope, $\beta = 0.367$, $F_{change}(1,67) = 8.65$, $p < .01$ (results in Supplement S7). The direction of the effect (Fig. 5) points to more categorical VAS performance in listeners with higher congruency effects. Thus, in contrast to our prediction, our results point to a link between inhibitory control and VAS. The direction of the effect indicates that subjects with better inhibitory control showed more gradient VAS responding. We return to this finding in the Discussion.

3.3.2. *Phoneme categorization gradiency and lexical inhibition*

We next investigated lexical inhibition using the subphonemic mismatch task. Participants performed the word recognition task promptly ($M = 1{,}216$ ms, $SD = 109$ ms) and accurately ($M = 99.6\%$, $SD = 1\%$). We excluded 13 participants from the analyses due to problems with the eye-tracking. Analyses on accuracy and RTs are reported in Supplement S8.

Previous studies found that listeners look less to the target picture in the word-splice condition ($ne_{ck}t$) than the matching- (ne_it) and nonword-splice (ne_pt) conditions (Dahan et al., 2001; Kapnoula et al., 2015; Li et al., 2019). We analyzed the fixation data to verify the presence of this lexical inhibition effect in our data and to ask whether the magnitude of the effect was modulated by gradiency. We first computed the proportion of trials on which participants fixated the target for each of the three splicing conditions at each point in time. Fig. 6A shows an effect of splicing, with the highest likelihood of fixating the target in the matching-splice condition, followed by the nonword- and word-splice conditions.

For statistical analyses, we computed the average proportion of fixations to the target between 600 ms and 1,600 ms post stimulus onset (logit-transformed). This value was the dependent variable in a linear mixed effects model. Splice conditions were contrast-coded; 1) matching- versus nonword-splice (+/-0.5) and 2) nonword- versus word-splice (+/-0.5). We also added VAS slope and its interactions with the splicing contrasts to the fixed effects. The most complex model supported by the data included random intercepts for subjects and items.

Participants looked significantly more to the target in the matching-splice than the nonword-splice condition, $B = 1.045$, $t(4602) = 7.75$, $p < .001$, reflecting sensitivity to the subphonemic mismatch. The word- and nonword-splice conditions also differed, $B = 0.628$, $t(4602) = 4.65$, $p < .001$, with more looks to target in the nonword-splice condition, as expected. This effect replicates prior work which used this contrast as a key indicator of inhibition. VAS slope did not predict looks to the target, $B = 0.328$, $t(56) = 0.642$, $p = .524$ (Fig. 6B), and none of the interactions was significant, $t < 1$. Thus, differences in speech gradiency are not likely to be driven by lexical competition.

3.3.2.1. Top-down inhibition and lexical competition. Although not the goal of the present study, there is an increasing interest in the idea that inhibitory control is involved in lexical competition (e.g., Zhang & Samuel, 2018; Zhao et al., 2020). We thus asked whether lexical competition (specifically, our measure of lateral inhibition between words) is correlated with top-down inhibition. To examine this, we added the Stroop congruency effect to the model along with the splice-condition contrasts and their interactions. Neither the Stroop score, $t < 1$, nor any of the interactions, $t < 1$, $t < 1$, had a significant effect on looks to the target, suggesting that the two kinds of inhibition rely on different mechanisms.

3.3.3. *Perceptual encoding differences and phoneme categorization gradiency*

Finally, we examined cue-encoding using the N1 ERP component. As with the P3, four participants' data were excluded leaving 67. The ERP showed a clear negative deflection at around 150 ms, with the characteristic morphology of the N1 (Fig. 7A). As expected, smaller VOTs (more voiced sounds) evoked stronger (more negative) N1 (Getz and Toscano, 2021; Toscano et al., 2010).

As before, our first analysis was intended to replicate the VOT effect on the N1, before moving to individual differences. For each experimental cell, we computed the N1 amplitude as the average voltage from 115 to 170 ms over the channels Cz, CP1, CP2, C3, C4, FC2, FC1, CP5, CP6, Fz, FC5, and FC6 (see Fig. 7B, C)[4]. This was used as the dependent variable in a linear mixed effects model with VOT step, F_0, and their interaction as fixed effects. Place of articulation of the stimulus and the target (PoA; alveolar: 1, labial: −1) and voicing of target (voiceless: 1, voiced: −1), and their interactions were added as covariates. The random effects included random VOT and F_0 slopes for subject (and their interaction) and random VOT slope for channel. Lastly, similarly to the P3 analyses, we excluded any cells with 6 or fewer trials in order to eliminate any noise due to the low number of contributing trials. All N1 models and results are reported in Supplement S9.

[4] Analyses including a) all channels and b) only the three frontal channels used in Toscano et al (2010) led to qualitatively identical results.

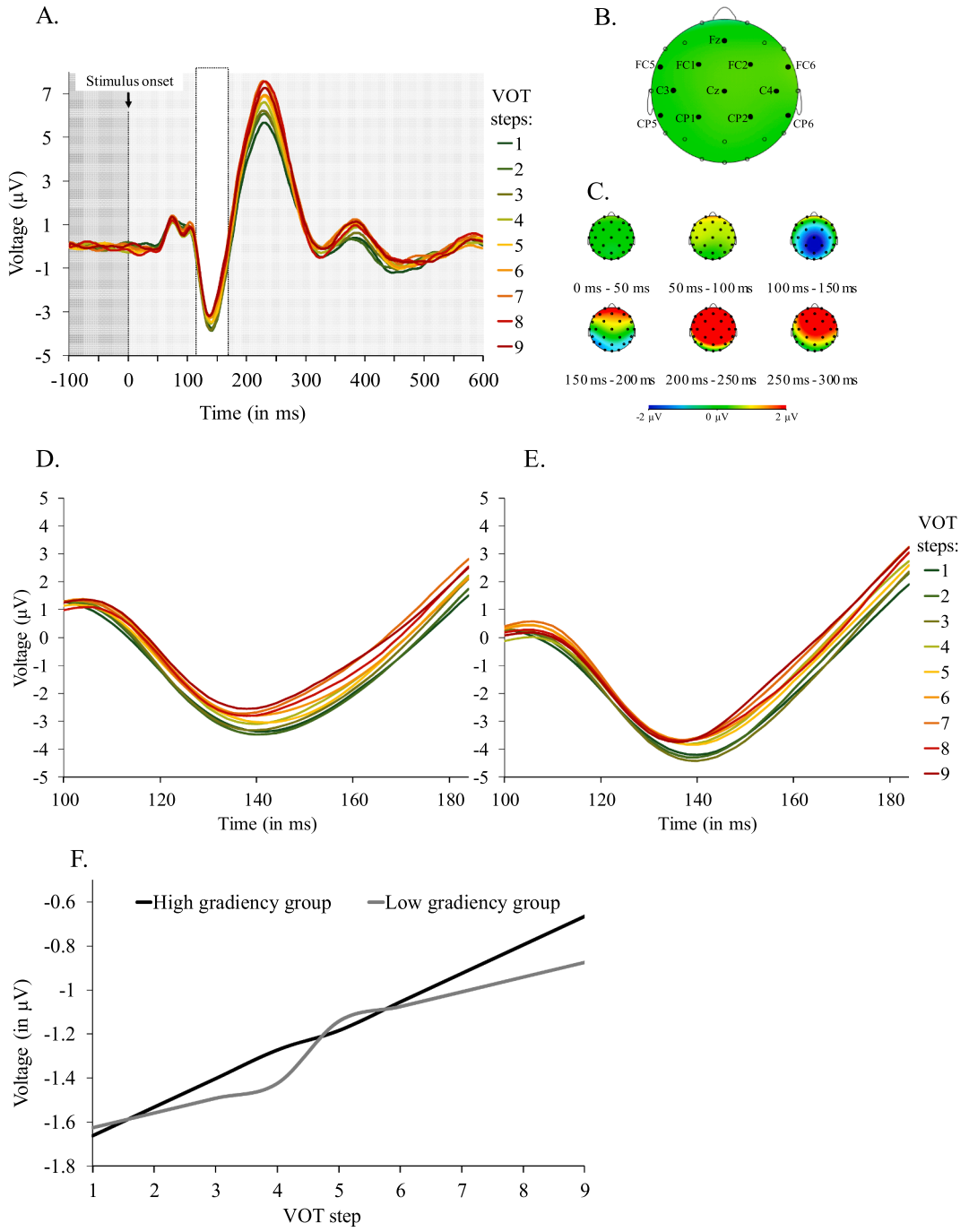


Fig. 7. A) Voltage as a function of time and VOT step. B) Electrode positions. C) Voltage as a function of time per electrode site. D) Voltage as a function of time and VOT step for the high gradiency group and E) for the low gradiency group. F) Model-estimated effect of VOT per group on N1 with stepVOT in the model.

There was a significant main effect of VOT, B = 0.114, t(43) = 8.17, $p < .001$, and F_0, B = 0.310, t(67) = 6.24, $p < .001$. As expected, higher VOTs and F_0s predicted higher average voltage (i.e., smaller N1). There was also a significant VOT × F_0 interaction, B = -0.040, t(67) = -2.31, $p < .005$, pointing to a stronger VOT effect for stimuli with low F_0. These results are consistent with previous findings showing that word-initial speech sounds with lower VOTs (i.e., more voiced) elicit stronger N1 (Toscano et al., 2010); the F_0 effect is also in line with that, as lower F_0s (more voiced) show stronger N1.

In the next model, we added subjects' place-specific VAS slope and its interaction with VOT as fixed effects. VAS slope was significant B = -0.159, t(107200) = -3.28, p = .001, with higher gradiency predicting smaller N1 amplitude. Crucially, there was a significant interaction of VOT and VAS slope, B = -0.066, t(3478) = -4.49, $p < .001$. The direction of the interaction points to a stronger VOT effect on N1 for gradient individuals. As shown in Fig. 7D, there seems to be a more robust linear relationship between VOT and N1 amplitude for gradient listeners. In contrast, for the categorical group (Fig. 7E), there was a separation between VOT steps 3 and 4, around the N1 peak (140 ms) pointing to a more step-like function.

To test this, we computed a *binary* variable reflecting stimulus identity (voiced/voiceless) for each subject (i.e., adjusted for their individual boundary). This variable (*step*VOT) was a step function. To compute it, we fit each participant's behavioral responses in the ERP

Table A1
Stimuli triplets (in IPA) used in the lexical competition (*net-neck* VWP) task.

Matching-splice ($ne_{t}t$ condition)	Word-splice ($ne_{ck}t$ condition)	Nonword-splice ($ne_{p}t$ condition)
beɪt (bait)	beɪk (bake)	beɪp
bæt (bat)	bæk (back)	bæp
braɪd (bride)	braɪb (bribe)	braɪg
bʌg (bug)	bʌd (bud)	bʌb
kɑrp (carp)	kɑrp (cart)	kɑrp
kæt (cat)	kæp (cap)	kæk
tʃɪk (chick)	tʃɪp (chip)	tʃɪt
dɑrt (dart)	dɑrk (dark)	dɑrp
dɑt (dot)	dɑk (dock)	dɑp
fɔrk (fork)	fɔrt (fort)	fɔrp
græd (grad)	græb (grab)	græg
hip (heap)	hit (heat)	hik
hʌb (hub)	hʌg (hug)	hʌd
dʒɑb (job)	dʒɑg (jog)	dʒɑd
nɑt (knot)	nɑk (knock)	nɑp
lip (leap)	lik (leak)	lit
mʌg (mug)	mʌd (mud)	mʌb
nɛt (net)	nɛk (neck)	nɛp
pɑrt (part)	pɑrk (park)	pɑrp
pɪk (pick)	pɪt (pit)	pɪp
poʊp (pope)	poʊk (poke)	poʊt
rɑd (rod)	rɑb (rob)	rɑg
ʃeɪk (shake)	ʃeɪp (shape)	ʃeɪt
steɪk (steak)	steɪt (state)	steɪp
sut (suit)	sup (soup)	suk
tɑrp (tarp)	tɑrt (tart)	tɑrk
wɛb (web)	wɛd (wed)	wɛg
zɪp (zip)	zɪt (zit)	zɪk

Table A2
List of images used in the lexical competition (*net-neck* VWP) task.

Target word	Cohort	Unrelated 1	Unrelated 2
bait	boot	jug	wet
bat	boat	street	drug
bride	bread	feet	yacht
bug	bark	dead	gap
cart	kid	snake	lid
cat	cord	blood	beard
chick	chart	hook	pig
dart	dog	ride	feed
dock	date	step	bulb
fork	fog	side	god
grad	gripe	stork	drop
heat	hood	maid	yard
hub	head	wreck	crib
job	jet	book	duck
knot	knight	rag	bead
leak	lark	peg	wig
mug	mit	spark	truck
net	nut	red	goat
part	pad	black	trout
pit	plug	luck	sweat
pope	plate	cube	dad
rod	root	bet	vote
shake	shed	choke	keg
steak	stick	check	milk
suit	sword	reed	flake
tarp	toad	jeep	vet
web	wood	cook	shout
zip	zap	cloud	raid

task using a four-parameter logistic and extracted the crossover parameter as a category boundary estimate, separately for each place of articulation. Based on this boundary, we created the *step*VOT variable, coded as 1/-1, depending on whether the VOT of a stimulus was higher or lower than the boundary for that combination of participant and stimulus. We then added *step*VOT and its interaction with VAS slope to the fixed effects of the model. The main effect of *step*VOT effect was significant, B = 0.053, t(109600) = 3.77, p < .001, as well as the *step*VOT × VAS slope interaction, B = 0.366, t(110200) = 8.53, p < .001. The direction of the interaction suggested a stronger *step*VOT effect on N1 for participants with steeper VAS slopes (i.e., less gradient; see Fig. 7F). Given the strong collinearity between *step*VOT and raw VOT, we also conducted a log-likelihood model comparison between the models with and without *step*VOT in the fixed effects, which allows us to ask if *step*VOT accounts for unique variance over and above raw VOT. The model including *step*VOT was significantly better, $\chi^2(2) = 84.86$, p < .001.

To further examine the *step*VOT × VAS slope interaction, we split participants by gradiency (median split using average VAS slope across continua). The same fixed and random effects structures were used as above (excluding VAS slope from the fixed effects). Raw VOT had a significant effect on N1 for both the low and the high gradiency group, B = 0.073, t(45) = 4.24, p < .001, B = 0.126, t(48) = 6.29, p < .001, respectively. However, *step*VOT was significant for the low gradiency group, B = 0.117, t(56540) = 6.22, p < .001, but not for the high gradiency group, B = -0.020, t(57540) = -1.04, p = .30.

Lastly, we asked whether raw VOT was significant over and above *step*VOT for the low gradiency group. To test this, we conducted the reverse analysis; we included *step*VOT in the first model (with the same random effects[5] structure as above). Then raw VOT was added. and the two models were compared using log-likelihood. As expected, the model with raw VOT provided a significantly better fit of the data, $\chi^2(1) = 14.73$, p < .001. This suggests that, even for low gradiency participants (i.e., listeners with more categorical cue encoding), VOT had a linear effect on N1 over and above any categorical/step-like effect.

To sum up, our N1 analyses show that listeners' sensitivity to subtle differences between speech sounds is reflected in their early brain responses. Specifically, we observed a linear relationship between N1 and VOT across listeners (replicating Toscano et al., 2010). This finding suggests that speech gradiency observed behaviorally stems from gradiency at the perceptual encoding of acoustic cues. Crucially, there was also evidence that this relationship is better described as a *combination of a linear and a step-like function*, at least for listeners who show a more categorical response pattern. When this finding is combined with the results from the other tasks it suggests that gradiency derives from the early encoding of speech cues (as indicated by the N1 results), but has consequences throughout later stages of phonological categorization (the P3) and word recognition (the VWP).

4. Discussion

We used an array of techniques and measures to study the nature of speech categorization gradiency. We first validated the VAS paradigm as a tool for documenting individual differences in speech categorization gradiency independently from domain-general categorization biases (see also, Kapnoula et al, 2021) and replicated the positive correlation between gradiency and cue integration. Second, we examined the relationship between the VAS assessment of gradiency, which is focused on individual differences, and two experimental paradigms used to establish gradiency in the modal listener. Gradiency extracted from explicit VAS ratings was associated with analogous results from both eye-tracking (VWP) and electrophysiological measures (P3), pointing to a common underlying mechanism. These results suggest that gradiency is a fundamental aspect of speech perception – all listeners are sensitive to within-category differences; but, at the same time, some listeners seem to encode speech in a way that more strongly reflects their categories – particularly for segments close to category boundaries. Third, regarding the sources of individual differences, our ERP results point to an early

[5] In this analysis, we kept raw VOT in the random effects in both models. We also ran a different set of analyses where stepVOT was included in the place of raw VOT in the random effects across both models and we again found that adding raw VOT was a better fit of the data, $\chi^2(1) = 122.81$, p < .001.

locus of gradiency, at the encoding of acoustic cues. In contrast, we found that gradiency in the VAS is not related to lexical competition, but may be modulated by cognitive control. We next discuss these contributions, focusing on our second and third (novel) research aims, and we link our results to previous research and to broader theoretical debates in speech perception.

4.1. Unifying different measures of gradiency

Our second goal was to examine the relationship between speech categorization gradiency, as measured by the VAS task, and sensitivity to within-category differences, as reflected in the VWP (McMurray et al., 2002, 2008) and the P3 (Toscano et al., 2010). While all listeners showed a gradient response in the VWP, as expected, gradient VAS categorizers showed higher sensitivity to within-category differences in the VWP (Fig. 3D) – though the effect was weaker with the residualized VAS measure (Supplement S4). A similar pattern was observed for the P3, which is thought to tap phoneme or lexical-level processing. Again, we saw a more robust effect of VOT on the P3 in listeners that performed more gradiently in the VAS (Fig. 4F). This provides clear converging evidence that VAS ratings reflect the core structure of downstream phonological categorization and lexical processing.

It makes sense that individual differences in gradiency would predict how much listeners use within-category differences to tune lexical activations in real time. This suggests that the two tasks tap into similar processes. However, this does not entirely concord with results of Kapnoula et al. (2021). They used a lexical garden-path paradigm in which listeners hear words like *barricade/parakeet* varying in onset VOT. They found that the level of initial commitment (indexed by eye movements) was not predicted by gradiency in the VAS; listeners' initial commitment was gradient regardless of their VAS performance. However, despite that, listeners' VAS gradiency predicted their likelihood of recovering from lexical garden-paths. This was a bit puzzling, given that likelihood of recovery should depend on the degree of initial commitment.

Our results suggest that the Kapnoula et al. (2021) study may have simply failed to detect an effect. In their study, the competitor (e.g., *parakeet* when hearing *barricade*) is only supported by the bottom-up input for a very brief period (until the listener hears *-ade*, which rules out *parakeet*). Thus, competitors may not have been active long enough for us to detect a moderation by gradiency. In contrast, when a competitor is active and there is no further bottom up information to rule it out (as in both the VWP and P3 paradigms used here; e.g., *pear* after hearing *bear*), then this potentially creates a more sensitive measure.

Despite the fact that both the P3 and the VWP measures were moderated by VAS gradiency, these effects were numerically small. This has two key implications. First, this fact may reflect differences in the psychometric properties of the measures. Both the VWP and ERP task used here require a large number of trials to achieve a reliable measure, and both were originally intended to assess the response to within-category changes at a group level (e.g., the modal listener). In contrast, the VAS task requires only a few trials at each step to achieve high reliability (Kong & Edwards, 2016), suggesting it may be more suitable as an individual differences measure. Second, at a theoretical level, the numerically small effects underscore the fact that all listeners may be generally sensitive to within-category differences, even as they differ in overall gradiency. Thus, even the listeners that we label as "categorical" are tracking within-category changes. Indeed, our electrophysiological results suggest that gradient cue encoding is observed in both subsets of listeners. This means that gradiency is a fundamental aspect of speech perception, even as there are also individual differences in the degree to which speech sounds are perceptually warped around the boundary. We discuss these results next.

4.2. Sources of gradiency

The third goal of this study was to examine potential sources of speech categorization gradiency, both within and outside the language domain.

4.2.1. Cue integration

Our empirical work was not designed to shed significant new light on the question of whether differences in cue integration are the source of differences in speech categorization gradiency. Nonetheless it is important to briefly consider this in light of prior work. Consistent with prior work we also found significant correlations between speech categorization gradiency and secondary cue use in the VAS task. However, as others have noted, these correlations are inconsistent and not observed for all sets of cues (Kapnoula et al., 2021). This suggests that cue integration is a consequence rather than a source of gradiency. This is in line with the idea we discuss shortly that differences in early cue encoding may drive differences in both speech categorization gradiency and cue integration. Critically, because this happens at the cue level, it may explain the idiosyncrasies across cues, (e.g., a listener's encoding of VOT may be distinct from their encoding of spectral cues).

4.2.2. Inhibitory control as a source of gradiency

Our results revealed a positive link between gradiency and inhibitory control, with more gradient participants showing better inhibition (smaller congruency effect). This finding was not expected given the lack of correlation between gradiency and inhibition reported by Kapnoula et al. (2017; using the Flanker task). However, we note that cognitive control tasks in general are not consistently reliable (Enkavi et al., 2019) and Kapnoula et al. (2017) used a fairly off-the-shelf version of the Flanker task (from the NIH toolbox), while here we took pains to create a more sensitive measure (by increasing the number of congruent trials; Logan & Zbrodoff, 1979).

Moreover, the direction of the effect is perhaps unexpected. A straightforward way to conceptualize the role of inhibitory control in speech gradiency is via suppressing competitors. When a listener with stronger inhibitory control hears *beach,* they are better able to suppress *peach,* leading to a more discretely activated representation. In contrast, a listener with weaker inhibitory control may leave *peach* active allowing for a more gradient response. That is not what we observed: people with stronger cognitive control showed more *gradient* representations (Fig. 5). To understand this, we must consider what exactly is reflected by the spatial Stroop congruency effect. The rationale of this task is that participants must suppress the incorrect option and activate the correct one as quickly as possible. Therefore, a higher congruency effect reveals greater difficulty in flexibly managing the activation of competing representations. How could such difficulty impact speech perception? In speech, different phoneme categories (and/or words) are the competing representations and managing this competing activation may be more difficult for individuals with poorer cognitive control.

This interpretation points to a *modulatory* rather than causal link between inhibitory control and gradiency; higher gradiency allows for multiple representations to become partly activated and—in those cases—greater flexibility in managing their activation (i.e., better inhibitory control) is necessary for gradiency to be maintained and reflected in the response. Flexible control of competing lexical representations would also explain why gradient listeners are better at recovering from lexical garden paths (Kapnoula et al., 2021). This is also in line with the results of the other tasks here: even listeners who were not gradient in the VAS task showed gradiency in the VWP, the P3, and the N1, consistent with robust evidence for gradiency in the modal listener (Andruski et al., 1994; McMurray et al., 2002, 2009). So, individuals with good inhibitory control may be better able to maintain this gradiency long enough for it to be reflected in their VAS rating.

On a related note, inhibitory control did not predict lexical inhibition. This is important not just for broader debates over how competition

is resolved in spoken word recognition (e.g., Zhang & Samuel, 2018), but also because it emphasizes that inhibitory control reflects domain-general cognitive operations, not differences in language processing. This underscores the findings from work both on individual differences (Kapnoula et al., 2017, 2021; Ou et al., 2021) and on the modal listener (Andruski et al., 1994; McMurray et al., 2002, 2009) that suggest that a gradient representation is something to be achieved and actively maintained as it is helps maintain flexibility. Cognitive control may be deployed to help achieve this goal.

4.2.3. Lexical competition as a source of gradiency

A second hypothesized source of differences in speech categorization was inhibition between words (within the mental lexicon). Our hypothesis was based on two aspects of spoken word recognition. First, words actively inhibit each other during spoken word recognition (Dahan et al., 2001). Second, activation at the lexical level flows back to the level of phonemes (Elman & McClelland, 1988; Luthra et al., 2021). Then, stronger inhibition of the competitor may lead to faster decay of competitor phonemes due to the feedback of activation. In the present context, this means that individuals with greater inter-lexical inhibition may suppress competitor words faster or more effectively, reducing any sensitivity to subtle activation differences.

This hypothesis was not confirmed (Fig. 6B). There are a number of likely reasons for this. First, we are positing a second order effect that simply may take time to percolate through the system, and which may be quite subtle. Second, it may possible that the strength of feedback also differs between individuals (e.g., Giovannone & Theodore, 2021), potentially masking the effect of lexical competition. Finally, lexical competition is not all or nothing. It is amenable to training (Kapnoula & McMurray, 2016), and in models like TRACE (McClelland & Elman, 1986) its strength varies between specific words. Consequently, it may not be a listener's general level of lexical competition that predicts gradiency, but rather the strength of lateral inhibition between specific words (e.g., *bill-pill*), which was not measured in our lexical competition task.

Thus, we cannot rule out the possibility of a mechanistic link of the sort outlined above that we simply failed to detect. However, given that we found much larger correlations between speech gradiency and other factors, it appears that individual differences in lexical competition are not a major factor (at least compared to other things) in predicting gradiency.

4.2.4. Early cue encoding as a source of gradiency

The last hypothesis was that differences between listeners in speech gradiency are due to differences in how they encode acoustic cues. To address this, we collected measures of pre-categorical encoding of VOT differences (the N1 ERP component). Our main prediction was that, if differences in speech gradiency are due to differences in the early encoding of speech cues, the linear relationship between N1 and VOT should be disrupted for individuals with steeper VAS slopes. Indeed, this is what we found.

Our results provided evidence for the first time that individual differences in speech categorization gradiency are linked to differences in how listeners encode speech cues, such as VOT. Specifically, we found stronger auditory ERP components (i.e., more negative N1s) for participants with steeper VAS slopes (i.e., lower gradiency). Second, in addition to the linear main effect of VOT on N1 amplitude, we also found evidence that, for steep-slope categorizers, the link between VOT and N1 amplitude had a step-like component (Fig. 7F). For those listeners, VOT encoding was best described by a hybrid model combining both a linear and a step-function. Critically, the step-like function was centered at each individual's category boundary – thus reflecting a category-driven warping effect. In contrast, for more gradient categorizers, the same step-function did not explain N1 variance over and above the linear model. This provides evidence for the first time that for some individuals, encoding of speech cues may be more strongly affected by category-related information and that the locus of this effect is perceptual.

More broadly, this pattern fits well with the literature on the effects of phoneme categories on speech perception; in that context, some listeners are more strongly affected by their phonological prototypes (Samuel, 1982), leading to stronger warping of their perceptual space (Kuhl, 1991). However, it is crucial to point out that the kind of warping we observed here is not entirely consistent with the kind of warping posited by the perceptual magnet effect or by categorical perception. First, warping does not appear to reduce listeners' sensitivity to fine-grained gradient changes in acoustic cues – rather it merely enhances their sensitivity to between-category differences. This is important as it means that this warping does not result in the loss of information that may be needed for dealing with context. More generally, this work (together with that by Kapnoula et al., 2021) suggests that perceptual warping is not across-the-board beneficial (or detrimental) for speech perception – rather, listeners who have less warping are more flexible in dealing with ambiguity, but this may differ across cues for any listener.

Second, even for listeners who showed this effect, a hybrid linear/step-function model was a better fit of the data compared to an exclusively step-function model. As it has been demonstrated by a number of studies, typical listeners are sensitive to within-category differences and it has been a challenge to reconcile these studies with findings showing better between-category discrimination. Our evidence for warping in some listeners may thus offer an integrative account that shows how both of these aspects of perception (better between-category discrimination and sensitivity to within-category differences) can coexist. Listeners can have enhanced discrimination at the boundary without losing the benefits of encoding fine-grained detail. In that way, our findings seem to support a type of model much like that proposed by Pisoni and Tash (1974) in suggesting that listeners use both continuous and categorical information. In addition, our results extend this account by showing that the relative contribution of each of these two facets of speech processing may differ substantially between individuals.

Finally, these findings also speak to an issue with the N1 paradigm (Getz and Toscano, 2021). Despite a large number of studies using the N1 as an index of low-level cue encoding (Getz & Toscano, 2019b; Sarrett et al., 2020; Toscano et al., 2010), it has never been clear if these N1 effects truly reflect the representations that code VOT. The alternative is that N1 effects are epiphenomenal. For example, the EEG on the scalp could entrain to the amplitude envelope of the signal (Gross et al., 2013; Keitel et al., 2018). In this case, voiceless sounds like /p/ have lower amplitudes at onset than voiced sounds (aspiration is quieter than voicing), predicting smaller N1 (exactly what is observed). The critical evidence against epiphenomenality would be a demonstration that the N1 response to VOT predicts speech perception performance. This has not been demonstrated by prior work, but here it is shown by the fact that the shape of the N1 predicts listener's responses in a completely independent task (the VAS). This offers strong evidence that the N1 is tracking representations that are causally involved in speech perception.

4.3. Gradiency and perceptual flexibility

The argument for gradient rather than categorical speech perception is typically rooted in the functional benefits for the listener (Clayards et al., 2008; Kleinschmidt & Jaeger, 2015; McMurray et al., 2002; Miller, 1997): listeners can be more flexible in accounting for the variation in the input if they make a partial, probabilistic decision. This is clearly the case as shown by work on the modal listener, where gradiency has been implicated most prominently in the ability to recover from ambiguity using later material (Brown-Schmidt & Toscano, 2017; McMurray et al., 2009; Szostak & Pitt, 2013), but also in phenomena like perceptual learning (Samuel, 2016, see McMurray & Farris-Trimble, 2012 for a discussion).

At first blush, the evidence for individual differences in gradiency shown here and elsewhere (Kapnoula et al., 2017, 2021; Kong &

Edwards, 2016; Ou et al., 2021) seems to challenge this consensus – some listeners simply do not appear to be gradient. Moreover, in two prior studies using the VAS (Kapnoula et al., 2017; Kong & Edwards, 2016), it looks like some people are behaving categorically – they do not use the middle of the VAS scale. Our results challenge this view. As we described, at a deeper level (e.g., as seen in the VWP, the P3 and the N1 response) all listeners appear to be gradient, and VAS slope appears to just moderate *how gradient they are*. In this way, the VAS task may amplify subtle differences in the degree of gradiency (which may be moderated by cognitive control). Thus, this apparent inconsistency between (1) gradiency as a fundamental aspect of speech perception, and (2) individual differences in gradiency, does not hold up.

In fact, the individual differences work described here and in our prior studies (Kapnoula et al., 2017, 2021) broadly supports the claims for a functional role of gradient representations. Listeners who are more gradient are better able to integrate multiple cues (seen here and in Kapnoula et al., 2017, 2021; Kim et al., 2020; Kong & Edwards, 2016; Ou et al., 2021), and they are better able to recover from phonetic ambiguity (Kapnoula et al., 2021). This is supported by our finding here that gradiency in the VAS task is related to gradiency at the phonological level (shown in the P3 here) and the lexical level (shown in the VWP). In addition, the positive correlation with inhibitory control suggests that listeners are actively using domain-general resources to achieve gradiency.

At the same time, this work as a whole suggests that *gradiency is not a panacea*. It is tempting to treat gradiency as a sort of "design principle" of the speech perception system that evolved or developed to achieve this kind of flexibility (Kleinschmidt & Jaeger, 2015). That is clearly not the case. Kapnoula et al. (2021) show that gradiency in one acoustic dimension (VOT) does not correlate with gradiency in a different cue (fricative spectra). Moreover, gradiency in one cue (VOT) only predicts cue integration for that cue, and even then, only certain kinds of cue integration (VOT × F_0, but not VOT × vowel length).

The present study explains why this may be the case: individual differences in gradiency derive not from some global approach to speech perception, but rather, from *idiosyncrasies in how specific cues are coded*. In this way, the fact that gradiency is beneficial for speech perception may be more of a happy accident, in which listeners can take advantage of these cue-specific idiosyncrasies of their own perceptual systems for functional benefit, when they have them. This may then explain why gradiency in one cue does not predict cue integration in other cues, or why gradiency is not correlated across cues (Kapnoula et al., 2021).

But the broader zeitgeist toward gradiency as a design principle to help listeners achieve flexibility also holds a lesson. It is not the first or the last such design principle that has been posited in speech perception. In fact, looking beyond gradiency, we see that some listeners do not integrate cues like VOT and F_0 at all (Kapnoula et al., 2017); some listeners do not recover from lexical garden-paths (Kapnoula et al., 2021); listeners vary in the degree to which they can engage in perceptual retuning (Schertz et al., 2016), and listeners vary in the degree to which they can use top-down feedback (Giovannone & Theodore, 2021). Much like gradiency, all of these have been touted as critical solutions to the problem of contextual variability in speech. Looking through the lens of individual differences, any approach driven by theoretical design principles seems too simplistic. Rather, listeners appear to have an array of mechanisms in their toolkit, and which ones any given listener has may be determined by the idiosyncrasies of their perceptual system, their cognitive capacities, the kind of speech they encounter every day, and their own developmental history. These can be assembled to solve the problem, but they may be assembled differently for different individuals or even for different classes of speech sounds or different regions of the lexicon.

5. Conclusions

In conclusion, our findings are consistent with the idea that all listeners are sensitive to within-category differences. However, we also report evidence for early perceptual warping of the acoustic space close to the category boundary, leading to the amplification of between-category differences, for some listeners. More broadly, our findings speak to the flexibility of the speech perception system in using both bottom-up and top-down sources of information, but perhaps in a way that is idiosyncratic to listeners, or even to specific acoustic/phonetic dimensions. This reconciles the seemingly contradictory findings showing both gradient and categorical effects in speech perception.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors thank Jamie Klein, McCall Sarrett, and the students of the MACLab for assistance with data collection. We also thank Tanja Roembke for help with the Stroop task and Francis Smith for help with recording auditory stimuli. This project was supported by NIH Grant DC008089 awarded to BM. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 793919, awarded to EK. This work was partially supported by the Basque Government through the BERC 2018-2021 program and by the Spanish State Research Agency through BCBL Severo Ochoa excellence accreditation SEV-2015-0490.

Appendix A

See Table A1 and A2.

Appendix B. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bandl.2021.105031.

References

Allen, J. S., & Miller, J. L. (2004). Listener sensitivity to individual talker differences in voice-onset-time. *Journal of the Acoustical Society of America, 115*(6), 3171–3183.
Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the Time Course of Spoken Word Recognition Using Eye Movements: Evidence for Continuous Mapping Models. *Journal of Memory and Language, 38*(4), 419–439.
Andruski, J. E., Blumstein, S. E., & Burton, M. (1994). The effect of subphonetic differences on lexical access. *Cognition, 52*(3), 163–187.
Apfelbaum, K. S., Blumstein, S. E., & McMurray, B. (2011). Semantic priming is affected by real-time phonological competition: Evidence for continuous cascading systems. *Psychonomic Bulletin & Review, 18*(1), 141–149.
Apfelbaum, K. S., Klein-Packard, J., & McMurray, B. (2021). The pictures who shall not be named: Empirical support for benefits of preview in the Visual World Paradigm. *Journal of Memory and Language, 121*, Article 104279.
Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1–48.
Boersma, P., & Weenink, D. (2016). Praat: doing phonetics by computer [Computer program].
Brown-Schmidt, S., & Toscano, J. C. (2017). Gradient acoustic information induces long-lasting referential uncertainty in short discourses. *Lang. Cogn. Neurosci., 32*(10), 1211–1228.
Burke, D., & Shafto, M. (2008). Language and aging. *Handbook of the Psychology of Aging*.
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.
Dahan, D., Magnuson, J. S., Tanenhaus, M. K., & Hogan, E. M. (2001). Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition. *Language and Cognitive Processes., 16*(5-6), 507–534.
Davis, M., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., & McGettigan, C. (2005). Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology: General, 134*(2), 222–241.

Elman, J. L., & McClelland, J. L. (1988). Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. *Journal of Memory and Language, 27*(2), 143–165.
Enkavi, A. Z., Eisenberg, I. W., Bissett, P. G., Mazza, G. L., MacKinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2019). Large-scale analysis of test-retest reliabilities of self-regulation measures. *Proceedings of the National Academy of Sciences, 116*(12), 5472–5477.
Frye, R., Fisher, J., & Coty, A. (2007). Linear coding of voice onset time. *Journal of Cognitive Neuroscience, 19*(9), 1476–1487.
Fuhrmeister, P., & Myers, E. B. (2021). Structural neural correlates of individual differences in categorical perception. *Brain and Language, 215*, 104919. https://doi.org/10.1016/j.bandl.2021.104919
Gerrits, E., & Schouten, M. E. H. (2004). Categorical perception depends on the discrimination task. *Perception & Psychophysics, 66*(3), 363–376.
Getz, L. M., & Toscano, J. C. (2019). Semantic context influences early speech perception: Evidence from electrophysiology. *Journal of the Acoustical Society of America, 145*(3), 1789–1789.
Getz, L. M., & Toscano, J. C. (2019). Electrophysiological Evidence for Top-Down Lexical Influences on Early Speech Perception. *Psychological Science, 30*(6), 830–841.
Getz, L. M., & Toscano, J. C. (2021). The time-course of speech perception revealed by temporally-sensitive neural measures. *Wiley Interdisciplinary Reviews: Cognitive Science, 12*(2). https://doi.org/10.1002/wcs.v12.210.1002/wcs.1541
Giovannone, N., & Theodore, R. M. (2021). Individual Differences in Lexical Contributions to Speech Perception. *Journal of Speech, Language, and Hearing Research*, 1–18.
Gross, J., Hoogenboom, N., Thut, G., Schyns, P., Panzeri, S., Belin, P., … Garrod, S. (2013). Speech Rhythms and Multiplexed Oscillatory Sensory Coding in the Human Brain. *PLoS Biology, 11*(12), e1001752. https://doi.org/10.1371/journal.pbio.1001752
Gwilliams, L., Linzen, T., Poeppel, D., & Marantz, A. (2018). In Spoken Word Recognition, the Future Predicts the Past. *Journal of Neuroscience, 38*(35), 7585–7599.
Kapnoula, E. C. (2016). *Individual differences in speech perception: Sources, functions, and consequences of phoneme categorization gradiency*. University of Iowa.
Kapnoula, E. C., Edwards, J., & McMurray, B. (2021). Gradient activation of speech categories facilitates listeners' recovery from lexical garden paths, but not perception of speech-in-noise. *Journal of Experimental Psychology: Human Perception and Performance, 47*(4), 578–595.
Kapnoula, E. C., & McMurray, B. (2016). Training alters the resolution of lexical interference: Evidence for plasticity of competition and inhibition. *Journal of Experimental Psychology: General, 145*(1), 8–30.
Kapnoula, E. C., Packard, S., Gupta, P., & McMurray, B. (2015). Immediate lexical integration of novel word forms. *Cognition, 134*, 85–99. https://doi.org/10.1016/j.cognition.2014.09.007
Kapnoula, E. C., Winn, M. B., Kong, E. J., Edwards, J., & McMurray, B. (2017). Evaluating the sources and functions of gradiency in phoneme categorization: An individual differences approach. *Journal of Experimental Psychology: Human Perception and Performance, 43*(9), 1594–1611.
Keitel, A., Gross, J., & Kayser, C. (2018). Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features. *PLoS Biology, 16*(3), e2004473. https://doi.org/10.1371/journal.pbio.2004473
Kim, D., Clayards, M., & Goad, H. (2018). A longitudinal study of individual differences in the acquisition of new vowel contrasts. *Journal of Phonetics, 67*, 1–20.
Kim, D., Clayards, M., & Kong, E. J. (2020). Individual differences in perceptual adaptation to unfamiliar phonetic categories. *Journal of Phonetics, 81*, Article 100984.
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.
Kong, E. J., & Edwards, J. (2011). *Individual differences in speech perception: Evidence from visual analogue scaling and eye-tracking*. Phonetic Sci: Proc. XVIIth Int. Congr.
Kong, E. J., & Edwards, J. (2016). Individual differences in categorical perception of speech: Cue weighting and executive function. *Journal of Phonetics, 59*, 40–57.
Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin & Review, 13*(2), 262–268.
Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics, 50*(2), 93–107.
Kuznetsova, A., Brockhoff, P., Christensen, R., & Jensen, S. P. (2020). lmerTest: Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package). R Packag. Version 3.1-3.
Leach, L., & Samuel, A. (2007). Lexical configuration and lexical engagement: When adults learn new words. *Cognitive Psychology, 55*(4), 306–353.
Li, M. Y. C., Braze, D., Kukona, A., Johns, C. L., Tabor, W., Van Dyke, J. A., … Magnuson, J. S. (2019). Individual differences in subphonemic sensitivity and phonological skills. *Journal of Memory and Language, 107*, 195–215.
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: General, 61*(5), 379–388.
Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word, 20*(3), 384–422.
Logan, G. D., & Zbrodoff, N. J. (1979). When it helps to be misled: Facilitative effects of increasing the frequency of conflicting stimuli in a Stroop-like task. *Memory & Cognition, 7*(3), 166–174.
Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing, 19*(1), 1–36.
Luthra, S., Peraza-Santiago, G., Beeson, K., Saltzman, D., Crinnion, A. M., & Magnuson, J. S. (2021). Robust lexically-mediated compensation for coarticulation: Christmash time is here again. *Cognitive Science, 45*(4), e12962.
Magnuson, J. S., McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2003). Lexical effects on compensation for coarticulation: A tale of two systems? *Cognitive Science, 27*(5), 801–805.
Marslen-Wilson, W. D., & Warren, P. (1994). Levels of Perceptual Representation and Process in Lexical Access: Words, Phonemes, and Features. *Psychological Review, 101*(4), 653–675.
Massaro, D. W., & Cohen, M. M. (1983). Categorical or continuous speech perception: A new test. *Speech Communication, 2*(1), 15–35.
McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18*(1), 1–86.
McClelland, J. L., Mirman, D., & Holt, L. L. (2006). Are there interactive processes in speech perception? *Trends Cognitive Science, 10*(8), 363–369.
McMurray, B. (2017). *Nonlinear Curvefitting for Psycholinguistic (and other) Data*. OSF.
McMurray, B., Aslin, R. N., Tanenhaus, M. K., Spivey, M. J., & Subik, D. (2008). Gradient Sensitivity to Within-Category Variation in Words and Syllables. *Journal of Experimental Psychology: Human Perception and Performance, 34*(6), 1609–1631.
McMurray, B., & Farris-Trimble, A. (2012). Emergent information-level coupling between perception and production. In A. C. Cohn, C. Fougeron, & M. Huffman (Eds.), *The Oxford Handbook of Laboratory Phonology* (pp. 369–395). Oxford University Press.
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*(2), B33–B42.
McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2009). Within-category VOT affects recovery from lexical garden-paths: Evidence against phoneme-level inhibition. *Journal of Memory and Language, 60*(1), 65–91.
Miller, J. L. (1997). Internal Structure of Phonetic Categories. *Language and Cognitive Processes., 12*(5-6), 865–870.
Miller, J. L., Green, K. P., & Reeves, A. (1986). Speaking rate and segments: A look at the relation between speech production and speech perception for the voicing contrast. *Phonetica, 43*(1–3), 106–115.
Munson, B., & Urberg Carlson, K. (2016). An exploration of methods for rating children's productions of sibilant fricatives. *Journal of Speech, Language, and Hearing Research, 19*(1), 36–45.
Munson, B., Edwards, J., Schellinger, S. K., Beckman, M. E., & Meyer, M. K. (2010). Deconstructing phonetic transcription: Covert contrast, perceptual bias, and an extraterrestrial view of Vox Humana. *Clin. Linguist. Phon., 24*(4-5), 245–260.
Näätänen, R., & Picton, T. (1987). The N1 Wave of the Human Electric and Magnetic Response to Sound: A Review and an Analysis of the Component Structure. *Psychophysiology, 24*(4), 375–425.
Nasman, V. T., & Rosenfeld, J. P. (1990). Parietal P3 Response as an Indicator of Stimulus Categorization: Increased P3 Amplitude to Categorically Deviant Target and Nontarget Stimuli. *Psychophysiology, 27*(3), 338–350.
Nearey, T. M., & Rochet, B. L. (1994). Effects of place of articulation and vowel context on VOT production and perception for French and English stops. *Journal of the International Phonetic Association, 24*(1), 1–18.
Noe, C., & Fischer-Baum, S. (2020). Early lexical influences on sublexical processing in speech perception: Evidence from electrophysiology. *Cognition, 197*, 104162. https://doi.org/10.1016/j.cognition.2019.104162
Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences, 23*(3), 299–325.
Ou, J., Yu, A. C. L., & Xiang, M. (2021). Individual differences in categorization gradience as predicted by online processing of phonetic cues during spoken word recognition: Evidence from eye movements. *Cognitive Science, 45*(3). https://doi.org/10.1111/cogs.v45.310.1111/cogs.12948
Pisoni, D. B., & Tash, J. (1974). Reaction times to comparisons within and across phonetic categories. *Perception & Psychophysics, 15*(2), 285–290.
Polich, J., & Criado, J. R. (2006). Neuropsychology and neuropharmacology of P3a and P3b. *International Journal of Psychophysiology, 60*(2), 172–185.
R Core Team. (2016). R: A Language and Environment for Statistical Computing. Vienna, Austria.
Repp, B. (1984). Categorical perception: Issues, methods, findings. *Speech Lang. Adv. Basic Res. Pract., 10*, 243–335.
Samuel, A. G. (1982). Phonetic prototypes. *Perception & Psychophysics, 31*(4), 307–314.
Samuel, A. G. (2016). Lexical representations are malleable for about one second: Evidence for the non-automaticity of perceptual recalibration. *Cognitive Psychology, 88*, 88–114.
Sarrett, M. E., McMurray, B., & Kapnoula, E. C. (2020). Dynamic EEG analysis during language comprehension reveals interactive cascades between perceptual processing and sentential expectations. *Brain and Language, 211*, 104875. https://doi.org/10.1016/j.bandl.2020.104875
Schellinger, S. K., Edwards, J., Munson, B., & Beckman, M. E. (2008). Assessment of children's speech production 1: Transcription categories and listener expectations. *Poster presented at the. ASHA Conv.*.
Schertz, J., Cho, T., Lotto, A., & Warner, N. (2016). Individual differences in perceptual adaptability of foreign sound categories. *The journal Attention, Perception, & Psychophysics, 78*(1), 355–367.
Schouten, B., Gerrits, E., & van Hessen, A. (2003). The end of categorical perception as we know it. *Speech Communication, 41*(1), 71–80.

Schouten, B., & van Hessen, A. (1992). Modeling phoneme perception. I: Categorical perception. *Acoust. Soc. Am., 92*(4), 1841–1855.
Sharma, A., & Dorman, M. F. (1999). Cortical auditory evoked potential correlates of categorical perception of voice-onset time. *Journal of the Acoustical Society of America, 106*(2), 1078–1083.
Sharma, A., Marsh, C. M., & Dorman, M. F. (2000). Relationship between N1 evoked potential morphology and the perception of voicing. *Journal of the Acoustical Society of America, 108*(6), 3030–3035.
Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology: General, 18*(6), 643–662.
Szostak, C. M., & Pitt, M. A. (2013). The prolonged influence of subsequent context on spoken word recognition. *The journal Attention, Perception, & Psychophysics, 75*(7), 1533–1546.
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.
Utman, J. A., Blumstein, S. E., & Burton, M. W. (2000). Effects of subphonetic and syllable structure variation on word recognition. *Perception & Psychophysics, 62*(6), 1297–1311.
Walker, A. (2020). Voiced stops in the command performance of Southern US English. *Journal of the Acoustical Society of America, 147*(1), 606–615.
Wühr, P. (2007). A Stroop effect for spatial orientation. *The Journal of General Psychology, 134*(3), 285–294.
Yu, A. C. L., & Zellou, G. (2019). Individual Differences in Language Processing: Phonology. *Annual Review of Linguistics, 5*(1), 131–150.
Zhang, X., & Samuel, A. G. (2018). Is speech recognition automatic? Lexical competition, but not initial lexical access, requires cognitive resources. *Journal of Memory and Language, 100*, 32–50.
Zhao, L., Yuan, S., Guo, Y., Wang, S., Chen, C., & Zhang, S. (2020). Inhibitory control is associated with the activation of output-driven competitors in a spoken word recognition task. *The Journal of General Psychology*, 1–28.