Gradient activation of speech categories facilitates listeners' recovery from lexical garden paths, but not perception of speech-in-noise

Efthymia C. Kapnoula
Dept. of Psychological and Brain Sciences
DeLTA Center
University of Iowa
Basque Center on Cognition, Brain and Language,

Jan Edwards
Hearing and Speech Science Department
Language Science Center
University of MD - College Park

and

Bob McMurray
Dept. of Psychological and Brain Sciences
Dept. of Communication Sciences and Disorders
Dept. of Linguistics
DeLTA Center
University of Iowa

Running Head: GRADIENT CATEGORIZATION FACILITATES GARDEN PATH RECOVERY

Corresponding Author:
Efthymia C. Kapnoula
Basque Center on Cognition, Brain and Language
Mikeletegi Pasealekua, 69
20009, Donostia, Gipuzkoa, Spain
kapnoula@gmail.com

Abstract

Listeners activate speech sound categories in a gradient way and this information is maintained and affects activation of items at higher levels of processing (McMurray et al., 2002; Toscano et al., 2010). Recent findings by Kapnoula, Winn, Kong, Edwards, and McMurray (2017) suggest that the degree to which listeners maintain within-category information varies across individuals. Here we assessed the consequences of this gradiency for speech perception. To test this, we collected a measure of gradiency for different listeners using the visual analogue scaling (VAS) task used by Kapnoula et al. (2017). We also collected two independent measures of performance in speech perception: a visual world paradigm (VWP) task measuring participants' ability to recover from lexical garden paths (McMurray et al., 2009) and a speech perception task measuring participants' perception of isolated words in noise. Our results show that categorization gradiency does not predict participants' performance in the speech-in-noise task. However, higher gradiency predicted higher likelihood of recovery from temporarily misleading information presented in the VWP task. These results suggest that gradient activation of speech sound categories is helpful when listeners need to reconsider their initial interpretation of the input, making them more efficient in recovering from errors.

Keywords: speech perception; gradiency; categorical perception; individual differences; visual world paradigm

Public Significance Statement

This study examined the role of individual differences in speech perception. Participants were asked to report their perception of speech sounds on a continuous scale (e.g., from *bin* to *pin*). In addition, we collected a number of other measures in order to assess how they process spoken language. Our results show that individuals differ in how sensitive they are to fine acoustic information (i.e., some can discriminate better between two different occurrences of *bin*). Such acoustic details are commonly considered noise that listeners should ignore. However, our results show that individuals who are more sensitive to such information are better in comprehending ambiguous utterances. In other words, maintaining ambiguity, whenever it exists, allows for more flexible speech perception. This finding goes against the common idea that efficient speech perception depends on discrete, step-like categorization of speech sounds.

Introduction

During speech perception, listeners rapidly make fine judgements about sounds on the basis of continuous and variable perceptual dimensions. For example, voice onset time (VOT) is the delay between the articulators' release and voicing onset. It is the primary cue contrasting /b/ from /p/; VOTs near 0 msec indicate a /b/, and VOTs near 60 msec a /p/, while VOT differences as small as 10 msec can be meaningful in assigning speech sounds to categories. However, VOT varies continuously as a function of talker (Allen & Miller, 2004), speaking rate (Miller et al., 1986), and coarticulation (Nearey & Rochet, 1994). Thus, the ability to rapidly categorize this continuous cue into meaningful units is a central challenge in speech perception.

Classic views of speech perception suggest that listeners' category decisions are discrete, in that each segment is strictly assigned to one category[1]. This direct mapping of continuous signal onto discrete categories concords with the idea that the encoding of continuous cues (like VOT) is categorical (Liberman et al., 1957; Liberman & Whalen, 2000). According to this view, listeners do not discriminate VOT differences within a category (or they do so less well), and this enables faster categorization (Liberman & Harris, 1961; Pisoni & Tash, 1974; Repp, 1984; Schouten & Hessen, 1992). This empirical phenomenon, known as *categorical perception* (CP), is thought to reflect listeners' adjustment of the sensory encoding of speech to the phoneme contrasts that are linguistically relevant in their native language.

In contrast to this categorical view, there is now accumulated evidence that listeners are sensitive to within-category differences at both sublexical (McMurray et al., 2008; Miller, 1997; Samuel, 1982; Toscano et al., 2010), and lexical (Andruski et al., 1994; McMurray et al., 2002; Utman et al., 2000) levels of processing. This has led to the broad acceptance (at least within the

[1] We use speech category rather than phoneme or feature as we do not wish to assume any particular form of the representation.

speech perception community) that gradiency is a fundamental aspect of speech perception. A critical corollary of this gradiency is that the system makes only a partial commitment to any category; a given speech input can be perceived as partially /b/ and partially /p/.

Gradiency as a useful feature of speech perception

Gradiency—and by extension, partial commitment to a category—has been argued to have several functional roles (Clayards et al., 2008; Kleinschmidt & Jaeger, 2015; McMurray et al., 2009; McMurray & Farris-Trimble, 2012; Oden & Massaro, 1978). First, gradiency may allow listeners to take advantage of fine-grained (within-category) acoustic differences. For example, processes like coarticulation lead to fine acoustic modifications that can predict upcoming speech sounds (Beddor et al., 2002; Daniloff & Moll, 1968), potentially speeding processing (Gow, 2001; Mahr et al., 2015; McMurray & Jongman, 2015; Salverda et al., 2014; Yeni–Komshian, 1981). Since these modifications can be within-category (i.e., don't shift the percept to a distinct category), this is only possible if listeners are sensitive to fine-grained acoustic detail, requiring gradiency at the level of cue encoding.

Second, if listeners encode cues continuously, this may allow for greater flexibility in how cues are combined (e.g., (Massaro & Cohen, 1983a; Toscano & McMurray, 2010). Similarly, continuous encoding of acoustic cues could be necessary in compensating for sources of variance like talker voice—for example recoding vowel formants relative to the talker's mean formants (McMurray & Jongman, 2011, 2015) require formants to be encoded continuously.

Third, listeners may better cope with phonetic uncertainty if they only make a partial commitment to one category over another. In a gradient system, the degree of commitment to a category depends on how prototypical the signal is (Andruski et al., 1994; McMurray et al., 2002; Miller, 1997). For example, a labial stop with a VOT of 5 msec and one with a VOT of 15 msec are both /b/. However, the former is a better example of /b/ and, thus, activates /b/-onset words more strongly; while the latter may partially activate both /b/ and /p/. This has been shown with goodness ratings (Allen & Miller, 1999; Miller & Volaitis, 1989), cross-modal priming (Andruski et al., 1994; Utman et al., 2000), and the visual world paradigm (VWP; McMurray et al., 2002). In line with this, Clayards et al., (2008) showed that listeners adopt a more gradient representation in response to uncertainty. They manipulated the variability of VOTs during a brief training session and found that when VOTs were more variable, listeners' responses were more gradient. The authors interpreted this finding as supporting an ideal observer model in which listeners faced with high uncertainty change the way they map cue values onto phoneme categories—when the signal is ambiguous it is better to keep options open until more information arrives.

One way to capture this flexibility are so-called garden-path words like *þumpernickel* (where /þ/ has a VOT of 10 msec, which is closer to /b/, but somewhat ambiguous between /b/ and /p/). In this case, listeners may initially activate /b/-initial words like *bumpercar* and *butter*. However, when they hear -*nickel*, they revise their initial interpretation. If listeners fully commit to /b/, they have made a garden path error and may be slow to recover (if at all). In contrast, if they keep /p/-initial items partially active, they may be able to reactivate them more quickly. Crucially, when the VOT is ambiguous (e.g., around 10-15 msec), a misperception is most likely. It is exactly in these situations that the likelihood of needing to revise is the greatest.

McMurray et al. (2009) argued that if the gradient activation of phonemes is reflected in lexical activations, then recovery from such garden paths should be related to within-category differences in VOT. To examine this, they constructed VOT continua ranging from a well-articulated word (e.g., *bumpercar*) to an overt misarticulation (e.g., *pumpercar*). This manipulation resulted in ambiguous stimuli (e.g., *pumpercar*), the onset of which (*pumper*) was partially consistent with two words (*bumpercar* and *pumpernickel*). As the VOT of the initial consonant approached the misarticulated endpoint, this induced lexical garden paths: listeners temporarily activated the competitor word (in this case, *pumpernickel*) and only later (upon hearing -/kar/), was there evidence in favor of the target (*bumpercar*).

Listeners' responses were examined in a VWP task which used eye-movements to monitor commitment to the target and competitor over time. As expected, both the probability of a lexical garden path (initial fixations to the competitor), as well as the time to recover from it (fixate the target) were linearly related to the magnitude of the acoustic discrepancy between target and signal. This suggests that an early graded commitment may permit more flexible updating (see also Gwilliams, Linzen, Poeppel, & Marantz, 2018). Further studies have extended this using sentences to force listeners to revise their initial decision (e.g., *The* $_{r/w}$*ing had an exquisite set of feathers*). Benefits of partial commitment were observed more than a second later (Brown-Schmidt & Toscano, 2017; Connine et al., 1991; Szostak & Pitt, 2013).

Contexts such as those used by Szostak and Pitt (2013), Brown-Schmidt and Toscano, (2017), and McMurray et al., (2009) constitute rather unnatural situations. However, they capture something common. Most real world utterances are less than ideal due to reductions, speech errors, and disfluencies. Moreover, speech is often processed in poor conditions (e.g., a cellphone on a noisy bus, a zoom call with bad internet). These factors make misperceptions common. In fact, in a naturalistic corpus, Bard, Shillcock, and Altmann (1988) found that as many as 21% of words could be recognized only after their offset.

Consequently, a system that maintains competitor activation to the degree that it might be needed later would be well situated for recovering from particularly phonetic ambiguity due to reductions or speech errors (see also McMurray & Farris-Trimble, 2012). Indeed, such adaptation could underlie the findings suggesting that normal hearing listeners maintain activation for competitors when processing speech-in-noise (Brouwer & Bradlow, 2016; McQueen & Huettig, 2012), as may cochlear implant users (who face degraded input) also do (Farris-Trimble et al., 2014; McMurray et al., 2016, 2019). In contrast, a categorical system – which ignores within-category detail– may fully commit to the incorrect category and find itself in a costly garden path situation when disambiguating information arrives.

This work suggests that listeners can engage this "strategy" of maintaining partial activation for competing interpretations when the acoustic input is ambiguous. However, this functional role of gradiency is only theoretical. To our knowledge, no study has directly asked if changes in the accuracy of speech perception are related to a listeners' gradiency.

Individual differences in speech categorization gradiency

Recent work suggests there are stable individual differences in gradiency that could be harnessed to ask this question (for a review on individual differences in speech perception, see Yu & Zellou, 2019). Kong and Edwards (2011), based on work by Munson, Edwards and Schellinger (2010) and Massaro & Cohen (1983b), measured the gradiency of listeners' phoneme categories using a visual analogue scaling (VAS) task. Participants heard tokens from a speech continuum (/*da*/ to /*ta*/) in acoustically equally distant steps. Participants responded by clicking on a line whose endpoints were labeled with the endpoints of the continuum; they could click anywhere on the line to rate how *da*-like or *ta*-like it was.

Methods that allow gradient responding are a key element in studying gradiency. For example, a two alternative forced choice (2AFC) identification task forces participants to coerce an underlying gradient percept to one of two options. This leaves uncertainty as to what an ambiguous response means. If a listener categorizes a sound as /d/ on 75% of trials, this could be for at least two reasons. First, the sound could be discretely categorized as /d/ or /t/ on each individual trial; however, due to noise in the encoding of VOT, 25% of trials are miscategorized. Alternatively, they may consistently perceive the sound as 75% /d/-like and choose /d/ on 75% of trials to map their gradient percept onto the responses[2]. A VAS task can distinguish these situations. In the first case, listeners would use only the endpoints of the line, choosing /d/ 75% of the time. In the second case, they would consistently respond at around 75% of the line. Thus, VAS may offer more insight into the underlying cause of a gradient response.

Using this task, Kong and Edwards (2011) showed unexpected individual differences in how listeners categorized speech sounds. Some listeners largely used the endpoints of the line, reflecting a categorical response, while others used the entire range, continuously reflecting the actual VOT (i.e., more gradient). Listeners who responded more gradiently appeared to rely more strongly on a secondary voicing cue (F_0). This pattern suggests that gradiency may have a broader role in speech processing. Building on this approach, Kapnoula et al. (2017) examined the relationship between an individual's phoneme categorization gradiency and several other factors linked to speech perception and non-linguistic cognitive processes. This revealed a number of key findings. First, Kapnoula et al (2017) replicated Kong and Edwards' (2011) finding that more gradient listeners showed stronger use of secondary cues. Second, they did not find a robust link between gradiency and executive function. Third, the authors reported a small correlation between categorization gradiency and performance in a sentence-in-noise

[2] However, we note that under some views, if they perceive the stimulus as 75% /b/, they should always choose /b/ (Nearey & Hogan, 1986)—thus, whether not a subject adopts this further adds to the ambiguity of interpreting such data.

comprehension task. This correlation was not significant after controlling for working memory, suggesting that general cognitive function may mediate this relationship.

If we take sentence comprehension in noise as a general measure of speech perception efficacy, the lack of a correlation with gradient categorization seems to argue against a globally beneficial role for gradiency. However, the theoretical argument for gradiency is not that it is globally better – rather that it may help listeners a) better integrate cues; and b) maintain flexibility if they need to revise a decision. That is, gradiency is likely specifically tied to *phonetic* ambiguity. In contrast, speech-in-noise perception may require a host of factors such as general cognitive processes and effort, as well as early auditory processes that segregate signal from the noise, or higher-level sentence comprehension that may help fill in the gaps when words are missed. Thus, speech-in-noise tasks may not best reflect the functional role gradiency in speech perception.

Present study

This study examines the role of speech categorization gradiency in coping with phonetic ambiguity and uncertainty in speech perception. We ask two key questions. First, is gradiency a general property of a listener – does it extend to multiple phonetic contrasts and play a role in multiple cue integration? This is critical for establishing if gradiency is tied to phonological processing more generally, or if it is a property of how we process specific cues. Second, we ask if this gradiency has functional consequences for word recognition. Our hypothesis is that gradiency may be beneficial when phonetic ambiguity leads to an incorrect early interpretation that must be revised later (i.e., a lexical garden-path).

Following Kapnoula et al. (2017), we used the VAS paradigm to measure individuals' degree of speech categorization gradiency. Listeners heard tokens from a voicing continuum (*bin* to *pin*) and responded by clicking on a line to indicate how *bin*-like or *pin*-like each stimulus was. From this, we extracted a measure of categorization gradiency. We also used a fricative continuum in the VAS task to determine if VAS gradiency is stable across different acoustic/phonetic cues. Additionally, we used a visual version of the VAS task to evaluate individuals' bias in how they use the VAS. This allowed us to determine if the VAS results in speech were specific to that domain, and to measure any general bias and partial it out of our measure of interest. Finally, we measured the use of two secondary cues for voicing (F_0 and vowel length), and one secondary cue used for distinguishing between fricatives (formant transitions). With these new continua, we sought to replicate and extend previous findings showing that gradiency is positively correlated to multiple cue integration (Kapnoula et al., 2017; Kong & Edwards, 2016).

As a whole, this set of measures was crucial for understanding what it means for listeners to be gradient: is this an idiosyncratic property of how they process specific acoustic cues, or a general property of listeners? If it is the former, one might expect a weak correlation between fricative and voicing gradiency, each should be correlated only with the related secondary cue use. If it is the latter, there should be a more uniform pattern of correlations. And if this is indeed a general property of perception (not just speech), or even a cognitive bias, the visual task should reflect this.

Next, we asked, whether gradiency predicts participants' ability to recover from lexical garden-paths after phonetic ambiguity. For this, we used a paradigm similar to the McMurray et al., (2009). We constructed VOT continua from word pairs like *bumpercar* and *pumpernickel* that ranged from a correct articulation (e.g., *pumpernickel*) to a clear misarticulation (e.g., *bumpernickel*). Our prediction was that listeners with higher gradiency should be better at perceiving and maintaining ambiguous acoustic information. Consequently, not fully committing to a speech category should allow those listeners to recover from initially misguiding information more often and/or more rapidly.

Finally, as a secondary goal, we evaluated the role of gradiency in speech-in-noise perception using a different task than Kapnoula et al., (2017). The sentence-level information available in their task may have helped listeners figure out the missing information using top-down information. Consequently, the finer grained analysis of the signal tapped by the VAS may not have been needed. Here, we assessed speech-in-noise perception using isolated words (Torretta, 1995). This task does not allow participants to take advantage of sentence-level information, forcing them to rely more on the acoustic input. If gradiency is broadly beneficial for speech perception, it should be correlated with performance in this task. In contrast, if gradiency plays a more targeted role in maintaining flexibility when phonetic cues are ambiguous; this task should not be correlated with gradiency.

Method

Participants

Sixty-seven (67) monolingual English speakers participated in this experiment. To verify these sample sizes were adequate to detect reasonable effects, we calculated minimum detectable effects (MDEs; i.e. smallest detectable effect size given a fixed alpha, power, and sample size) using G*Power (Faul et al., 2009, 2007). See Supplement 5 for analyses and results.

Participants had normal/corrected-to-normal vision and no known hearing or neurological impairments. Participants received course credit, and underwent informed consent in accord with University of Iowa IRB policies. One participant was excluded due to failure to follow the instructions. An additional 9 failed to return on the second day of the experiment. Two participants were excluded from the VWP analyses due to eye-tracking-related problems. We used all subjects that were available for a given analyses which left samples between 55 and 67 for different analyses.

Overview of measures

Categorization gradiency. Our critical independent variable was gradiency of speech categorization along the voicing continuum. This was assessed via a VAS task (Kapnoula et al., 2017; Kong & Edwards, 2011; Munson & Carlson, 2016; Schellinger, Edwards, Munson, & Beckman, 2008) using a voicing (/b/-to-/p/) continuum (Table 1: Voicing VAS). We used this task with a fricative continuum to examine whether gradiency is stable across contrasts (Table 1: Fricative VAS). Additionally, we used a visual continuum to assess individuals' bias in how they use the VAS and to partial this variance out of our measure of interest (Table 1: Visual VAS).

Secondary cue use. We used 2AFC tasks to assess secondary cue use (F_0 and vowel length as voicing cues; and transition/vowel as a frication cue; Table 1: 2AFC tasks). These tasks were meant to validate previous results showing that gradiency is positively correlated to multiple cue integration (Kapnoula et al., 2017; Kim et al., 2020; Kong & Edwards, 2016). If both (gradiency and secondary cue use) are general properties, then we should observe correlations between cue use measures and their corresponding gradiency measures, as well as among the three cue use measures. In contrast, if gradiency is tied to specific cues, then we should only see correlations with gradiency for those cues (e.g., VOT gradiency with voicing).

Garden-path recovery. The critical dependent measure was listeners' flexibility in recovering from lexical garden-paths, also along a voicing continuum. To assess recovery, we used a VWP task (similar to that of McMurray et al., 2009; Table 1: VWP). Our hypothesis was that more gradient listeners would maintain multiple partially active speech categories and therefore be better at recovering from the lexical garden-paths. In other words, *higher level of gradiency should predict better recovery from lexical garden- paths*.

Table 1. *Summary of tasks*

Day	Order	Task	Contrast (dimensions)	Measure
1	(internal order counterbalanced) 1	Voicing VAS	b/p (VOT × F_0)	Speech (b/p) categorization gradiency
		Fricative VAS	s/ʃ (frication × formant transition)	Speech (s/ʃ) categorization gradiency
		Visual VAS	apple/pear	VAS usage bias
	(internal order counterbalanced) 2	Voicing 2AFC	b/p (VOT × F_0)	Secondary cue use (F_0)
		Voicing 2AFC	b/p (VOT × vowel length)	Secondary cue use (vowel length)
		Fricative 2AFC	s/ʃ (frication × formant transition)	Secondary cue use (transition/vowel)
	3	Speech-in-noise		Speech perception in noise
2	4	VWP	b/p (VOT)	Flexibility in spoken word recognition

Speech-in-noise. Following up on Kapnoula et al. (2017), we asked whether gradiency is related to word recognition in noise (Table 1: Speech-in-noise). If gradiency aids speech perception globally, then higher gradiency should predict better word recognition in noise; but, if gradiency is more targeted (e.g., ambiguous cues), it may not be related to speech-in-noise.

Participants performed the VAS, 2AFC, and speech-in-noise tasks on day 1 and returned on a second day for the VWP task (Table 1).

Measuring categorization gradiency (VAS tasks)

Stimuli: Similarly to Kapnoula et al. (2017), we used a stop-onset (*bin-pin*) continuum. Stimuli were constructed from recordings of natural speech recorded by a male monolingual speaker of American English. We created a 7×5 *bin-pin* continuum by manipulating the two major cues used to distinguish /b/ from /p/: VOT and pitch (F_0). Pitch was manipulated using the Pitch Synchronous Overlap Add (PSOLA) method, in Praat (Boersma & Weenink, 2016 [version 5.3.23]). Average pitch was 138 Hz for *bin* and 146 Hz for *pin*. The pitch contours for each word were extracted and modified to create five contour steps of identical length and shape, differing only in average F_0 (138, 140, 142, 144, and 146 Hz). We then replaced the original pitch contours with each of these new contours and each token was resynthesized.

We next constructed a VOT continuum for each pitch step. VOT varied in 7 steps from 0 to 40 msec approximately 6.7 msec apart. Stimuli were constructed using the progressive cross-splicing method described by Andruski et al. (1994); progressively longer portions of the onset of the voiced sound (/b/) were replaced with analogous amounts taken from the aspirated period of the corresponding voiceless sound (/p/).

The fricative VAS task used a 7×2 *same-shame* continuum. Stimuli were constructed by manipulating the spectral peak of the frication and the formant transitions (details in Supplement S1. Lastly, for the visual VAS task, we used a 7×5 apple-pear continuum with stimuli varying in shape and color (red→yellow; details in Supplement S2).

Design. Each participant heard each b/p stimulus three times for a total of 105 trials. Fricatives were presented seven times each for a total of 98 trials. Visual stimuli were presented five times each for 175 total trials. These were done on separate blocks.

Procedure. On each trial, participants saw a line whose ends were labeled according to the two categories (*bin, same, apple* were always on the left). They listened to/saw a stimulus and clicked on the line to indicate where they thought it fell on the continuum. When they clicked, a rectangular bar appeared where they clicked and they could then change their response or press the space bar to verify it. Each VAS task took approximately 7.5-10 mins.

Measuring gradiency. As in Kapnoula et al (2017), we used the rotated logistic function (Equation 1) to fit participants' response functions for the auditory and visual VAS tasks. Unlike standard logistic regression, this provides orthogonal measures of gradiency and secondary cue use (i.e., use of F_0).

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

This equation has 6 parameters. As in the four-parameter logistic: b_1 and b_2 are the lower and upper asymptotes. However, boundary is handled differently. This equation assumes a diagonal boundary in two-dimensional (VOT × F_0) space that can be described as a line with some cross-over point (along the primary cue, VOT) and an angle, θ. A θ value of 90^o indicates exclusive use of the primary cue, while a θ of 45^o reflects use of both cues. Thus, the boundary is captured by both x_0 and θ. 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 this diagonal boundary. Lastly, *υ(θ)* switches the slope direction, if θ is less than 90. This is done to keep the function continuous.

This *rotated logistic* function models the gradiency of the function with a single parameter that indicates the derivative of the function orthogonal to the (diagonal) boundary; steeper slopes indicate a more categorical response. The slope value is taken as a measure of categorization gradiency. This function is superior to the standard logistic in that it 1) allows for asymptotes that are not 0 and 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.

The equation was fit to each participant's VAS responses using a constrained gradient descent method implemented in Matlab (using FMINCON) that minimized the least squared error (free software available at McMurray, 2017). Fits were good, with an average R^2 of 0.94[3], 0.99, and 0.95[4] for the voicing, fricative, and visual tasks respectively.

Measuring secondary cue use (2AFC tasks)

Similarly to Kapnoula et al. (2017), we evaluated secondary cue use. The same word pairs from the VAS tasks: (*bin-pin* and *same-shame*) were used to construct two labial and one fricative stimulus sets. Each set manipulated the primary cue (VOT/frication) along a 7-step continuum. In addition, two levels of a secondary cue were used (F_0 and vowel length in separate continua for the labial sets in and formant transition for the fricative set). Stimuli from each set were presented in separate blocks and participants identified the word they heard. To measure secondary cue use, we estimated the shift in the categorization function along the primary cue between the two levels of the secondary. Details can be found in Supplement S3.

Measuring recovery from lexical garden-paths

Design and materials. To measure how flexibly listeners cope with temporary ambiguities during spoken word recognition, we used a VWP task based on McMurray et al. (2009). In this task, auditory stimuli were based on word pairs like *barricade-parakeet*. Words in each pair differed in initial voicing (/b/ versus /p/), but were identical for the next 2-5 phonemes. Thus, whenever the initial VOT was ambiguous, the word would be disambiguated by later information (e.g., the *–ade* or *–eet*). Five such pairs were developed (Table 2).

For each pair, VOT was manipulated along a continuum to create a word to nonword continuum (e.g., *barricade* [bærəkeɪd**]** to <u>*parricade*</u> [pærəkeɪd]). Stimuli were constructed by splicing natural recordings. We started by recording complete exemplars of both items in each pair, with both a voiced and voiceless onset (e.g., *barricade, parricade, barakeet,* and *parakeet*).

[3] Five problematic fits were excluded.
[4] One problematic fit was excluded.

A native American English speaker recorded multiple tokens of each item in a sound attenuated room at 44,100 Hz and the best exemplars for item were identified. Recording were then split into two parts: the onset (e.g., *bumper*- from *bumpercar*) and the offset (e.g., -*car*). Stimuli were cut at the zero-crossing closest to the point of disambiguation (POD; ~ 384.1 msec).

Table 3. *Stimuli used in the Lexical garden path task (in International Phonetic Alphabet;*

Set	Voiced Word		Voiceless Word		Overlapping phonemes
	Spelling	IPA	Spelling	IPA	
1	bumpercar	bʌmpər**kɑr**	pumpernickel	pʌmpər**nɪkəl**	5
2	barricade	bærək**eɪd**	parakeet	pærək**it**	4
3	blanket	blæŋk**ɪt**	plankton	plæŋk**tən**	4
4	beachball	bitʃ**bɔl**	peach-pit	pitʃ**pɪt**	2
5	billboard	bɪlb**ɔrd**	pill-box	pɪlb**ɒks**	3

Note: underlined portions marks phonemic overlap between the words in a pair; bolded portions mark offsets

The onset portions may contain coarticulatory cues predicting the offset (e.g., the *barri* from *barricade* may predict –*cade* more than –*keet*). To address this, each of the two voiced onsets in a pair (e.g., *bumper*$_{\textit{car}}$ and *bumper*$_{\textit{nickel}}$) was spliced onto each of the two offsets (e.g., $_{\textit{bumper}}$*car* and $_{\textit{bumper}}$*nickel*; see Supplement Fig.S4.A). Thus, half of the resulting stimuli contained parts from the same item (e.g., *bumper*$_{\textit{car}}$ and $_{\textit{bumper}}$*car; matching splice*) and half were made from different items (e.g., *bumper*$_{\textit{car}}$ and $_{\textit{pumper}}$*car; mismatching splice*). This counter-balanced coarticulatory cues in the onsets (e.g., the -er in *bumper* appeared with coarticulation from both the -ar in *car* and with -ɪ from -*nickel*).

Finally, we constructed the VOT continua. Items differing only in the voicing of the onset consonant were paired (e.g., *bumpercar* and *pumpercar*) and were used as the endpoints to create

7-step (0–48 ms) VOT continua. Continua were constructed using progressive cross-splicing. This yielded 140 auditory items (5 pairs × 2 splice conditions × 2 offsets × 7 VOT steps; see Supplement Fig.S4.B). Each item was presented 3 times resulting in 420 experimental trials.

Ten pairs of fillers were used. Filler items began with continuants (/l/ and /r/); they were phonetically dissimilar from one another; and had minimal overlap (e.g. *limousine* and *raspberry*). Similar to experimental items, there were two versions of each filler: unaltered (e.g., *limousine*) and misarticulated (e.g., *rimousine*). No splicing was performed. Each of the 10 filler words was presented 21 times in its correct and mispronounced form, yielding a total of 420 (5 pairs × 2 variants × 21 repetitions) filler trials (equivalent to the experimental).

Each experimental pair was grouped with a filler pair to form a 4-item set (e.g., *barricade, parakeet*, *limousine*, and *raspberry* formed one set), with the constraint that all items within a set were semantically unrelated and had the same number of syllables and stress pattern.

Visual stimuli consisted of 20 pictures, one for each of the four words in each of the five sets. Pictures were developed using a standard lab procedure (McMurray et al., 2010). For each word, several pictures were downloaded from a clipart database and viewed by a small focus group. From this set, one image was selected as the most representative exemplar. These were edited to remove extraneous elements, adjust colors, and ensure a clearer depiction of the intended word. Final images were approved by a lab member with extensive VWP experience.

Procedure. Participants were familiarized with the pictures by seeing each picture along with its printed name. After being fitted with the eye-tracker, they were given instructions. On each trial, participants saw the four pictures of a given set along with an X. Subjects could click on this X if they thought none of the four pictures matched what they heard (e.g., if they heard *parricade* and did not recover).

Stimuli were presented on a 19” monitor operating at 1280 × 1204 resolution. The five visual stimuli were presented in a pentagonal display (see Supplement Fig.S5). The center of each picture was equidistant from the center of the screen (440 pixels) and from each other (517 pixels). The position of the four pictures was randomized across trials, except for the X which was always at the bottom. Each picture was 300 × 300 pixels, while the X was 66 × 80 pixels.

At the beginning of each trial, a red circle appeared at the center of the screen along with the five pictures. After 500 msec, the circle turned blue and participants could click on it to hear the auditory stimulus. This pre-scan gave participants time to briefly look at the pictures before hearing the target word, thus minimizing eye-movements due to visual search (rather than lexical processing). Once participants clicked on the circle, it disappeared and the auditory stimulus was played. Participants then clicked on the corresponding picture, and the trial ended. There was no time limit, but participants typically responded in less than 2 sec (*M* = 1325 ms, *SD* = 200 ms).

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 procedure. They eye-tracking record was automatically parsed into saccades and fixations using the default psychophysical parameters, and adjacent saccades and fixations were combined into a single “look” that began at the onset of the saccade and ended at the offset of the fixation (see also McMurray et al., 2002, 2010). Eye-movements were recorded from the onset of the trial (blue circle) to the response (mouse click). This variable trial duration makes it difficult to analyze results late in the time course. To address this issue, we adopted the approach of many prior studies (Allopenna et al., 1998; McMurray et al., 2002) by setting a fixed trial duration of 2,000 msec (relative to stimulus onset). For trials that ended before this point, we extended the last eyemovement; trials which were longer than 2,000 msec were truncated. This is based on the assumption that the participants' last fixations reflect the word they "settled on", and therefore should be interpreted as approximating the final state of the system. For assigning looks 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.

Measuring spoken word recognition in noise

To measure how well participants cope with background noise, we presented stimuli from the "*Easy-Hard" Word Multi-Talker Speech Database* (Torretta, 1995). All items were natural recordings of real words. Items varied in difficulty (easy/hard, as a function of frequency and neighborhood density), talker's gender (male, female), and speaking rate (fast, medium, slow). We sampled 100[5] words, half of which were "hard" and half "easy". Each word was presented three times, in a slow, medium and fast speaking rate, – each time in a different voice (of 10 possible voices [5 male]). This led to a total of 300 trials. Stimuli were masked with white noise at 8 dB SNR[6] and presented over high quality headphones. Participants responded by typing the word on a keyboard with unlimited time. Accuracy was computed automatically (no feedback was given) and checked offline by trained research assistants, who corrected any typos.

[5] Due to an error, three (3) words were repeated, resulting in 97 unique words. Analyses reported here include all items; however we ran the same analyses excluding the three repeated items, generating identical results.

[6] Accuracy for these stimuli presented in silence is ~91% (Bradlow & Pisoni, 1999). To avoid ceiling effects, which may hide the role of individual differences, we aimed at having a much lower difficulty rate (~57%).

Results

Participants performed all tasks successfully with the exception of one who failed to follow the VAS instructions and was dropped from all analyses of those measures. In addition, there was a technical problem with the voicing VAS output file for one participant.

Categorization gradiency and secondary cue use

Categorization gradiency. Participants performed the VAS tasks as expected: They used both cues (e.g., VOT and F_0; Fig.1A) and the entire range of responses (Fig.1B;D;F), even though proportion of intermediate responses (20%-80% VAS ratings) was naturally higher for ambiguous (Fig.2B), compared to unambiguous stimuli (Fig.2A). Also as expected, some participants had a strong preference for the endpoints (Fig.3A), while others clicked more often on intermediate points on the line (Fig.3B).

We next examined the correlations between VAS gradiency measures (Fig.4). Visual VAS slope was marginally correlated with voicing VAS slope , r(57) = .22, p = .089 (Fig.4A), but it was significantly correlated with fricative VAS slope , r(63) = .34, p < .005 (Fig.4B). More importantly, even the significant correlation for the fricatives was only a moderate effect size, suggesting that participants’ responses in the speech VAS task cannot be fully attributable to general biases in how they perform VAS tasks.

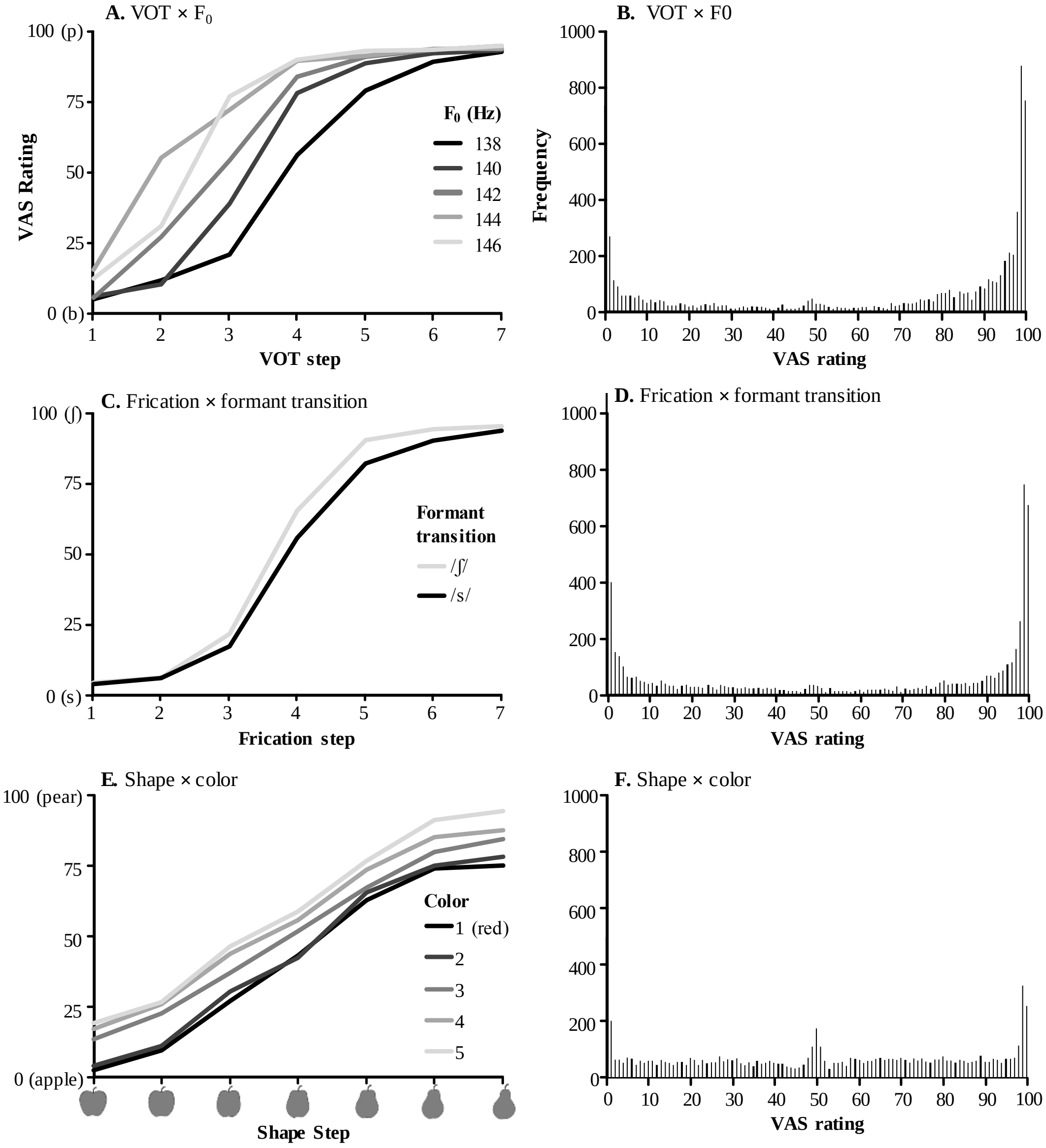

Figure 1. VAS ratings for voicing (A, B), fricatives (C, D), and visual (E, F) task. Plots on the left show average rating per continuum step in each task. Plots on the right show frequency of each rating across participants.

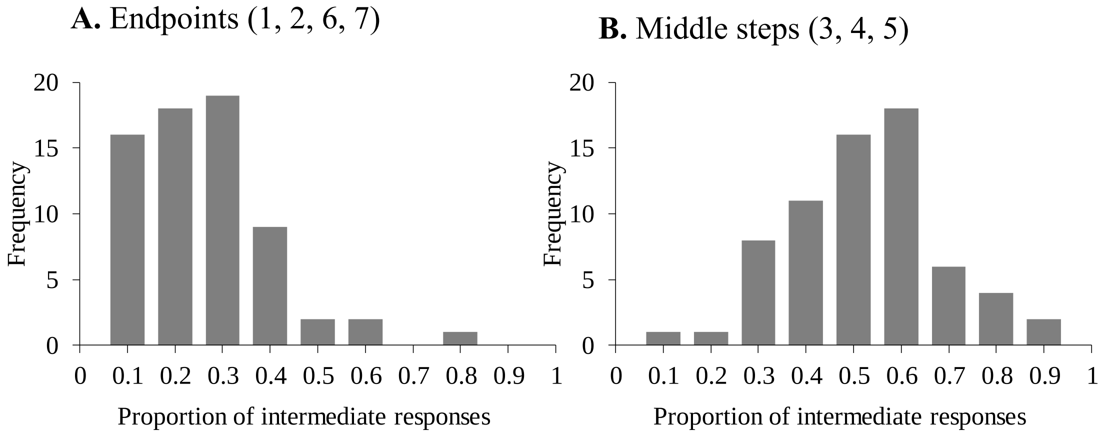

Figure 2. Proportion of intermediate (20%-80%) VAS responses for unambiguous (A) and ambiguous stimuli (B) across participants.

As a more conservative measure of gradiency for subsequent analyses, we adjusted participants' gradiency scores to control for their performance on the visual task. For this, we used the standardized residual of the speech VAS slopes after partialing out the visual VAS slope. These residualized slopes were included in all subsequent analyses (though for convenience these are henceforth termed VAS slopes). Corresponding analyses using raw VAS slopes produced almost identical results.

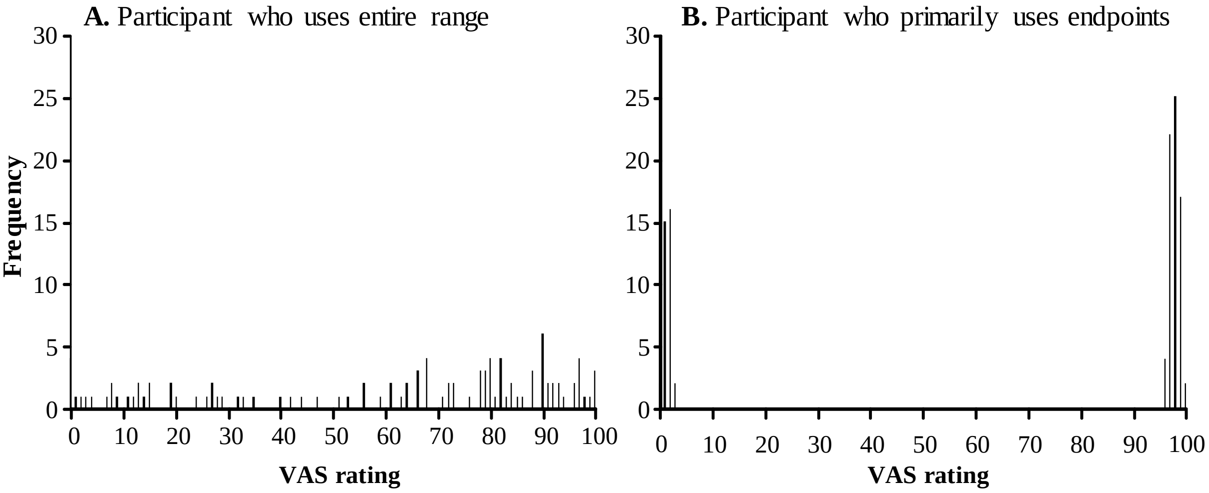

Figure 3. Examples of different VAS rating patterns across individuals. Subject 20 (A) used the entire range of responses, while subject 28 (B) had a clear preference for the endpoints.

The fricative and voicing VAS slopes were not correlated, r(57) = .19, p = .16 (Fig.4C), which suggests that individual differences in gradiency may be contrast-specific.

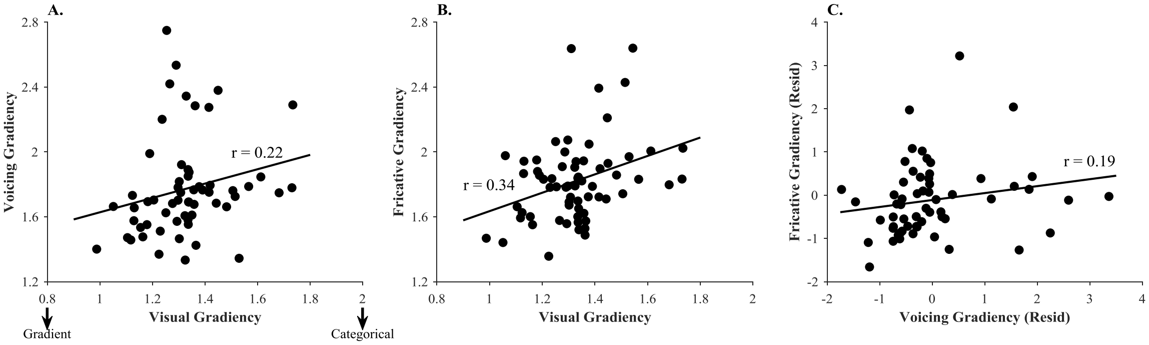

Figure 4. Correlations among VAS slope measures of gradiency (note gradient listeners are shown as having a low slope and categorical as high). A) Correlation between visual gradiency and voicing gradiency; B) Correlation between visual gradiency and fricative gradiency; C) Correlation between visual gradiency and fricative gradiency.

Secondary cue use. Participants performed the three 2AFC tasks as expected (Supplement Fig.S3). Participants' 2AFC responses were fitted using a four-parameter logistic function and the crossover difference was used as an estimate of secondary cue use (Supplement S4). Use of F_0 was positively correlated with use of vowel length, r(63) = .26, p =.034, and formant transition (for fricatives; r(63) = .26, p = .037). The correlation between formant transition and vowel length was of similar magnitude, but not significant, r(64) = .21, p = .09. This pattern offers some evidence that secondary cue use is in part a stable aspect of speech processing for a listener. However, these effects are small: a larger portion of the variability may be due to the particular cues and the phonemic contrast for which they are used.

We next asked whether speech gradiency is linked to secondary cue use. We tested the correlation between each one of the three secondary cues and the VAS slope from the corresponding stimulus. F_0 use in the 2AFC task was significantly correlated with VAS slope in the voicing (VOT × F_0) continuum, r(57) = -.37, p = .004. The direction of the effect indicates that higher gradiency (i.e., shallower VAS slope) is associated with higher secondary cue use.

This replicates Kapnoula et al. (2017) and Kong and Edwards (2016). Interestingly, use of vowel length was not correlated with VAS slope, r(57) = -.09, p = .48. Moreover, for the fricatives, the use of the formant transitions in the 2AFC task, was not correlated with VAS slope, r(63) = -.13, p = .31. This was somewhat surprising and suggests that speech gradiency may only predict secondary cue use in specific situations, not secondary cue use in general.

Interim Summary. First, the correlation between visual and voicing VAS measures was weak and non-significant, while the correlation between fricative and visual VAS measures was only moderate. This pattern suggests that the VAS measure of speech perception gradiency is not just a measure of how people generally approach VAS tasks. Second, speech gradiency measures were only weakly correlated across cues, which shows that this is a highly specific measure of how a given acoustic contrast is encoded. Finally, we replicated previous results showing a robust correlation between voicing gradiency and secondary cue use, and we further validated this relationship by showing that it persists even after accounting for differences in how individuals perform VAS tasks in general. However, this was not observed for other cue-integration problems (VOT/VL) or for fricative gradiency. This reinforces the notion that gradiency is cue-specific.

Phoneme categorization gradiency and recovery from lexical garden paths

We next turn to our primary question: does maintaining within-category information (higher gradiency) help listeners when they must reconsider an initial interpretation of the input? We first conducted a group level analysis on accuracy, RT, and fixations in the lexical garden path task (similar to McMurray et al., 2009). This was done to ensure the validity of the VWP experiment and to identify the factors for the individual differences analyses. Next, we asked whether participants' ability to cope with ambiguities was predicted by speech gradiency.

For these analyses, VOT step was recoded as distance from the target (tDist), similarly to McMurray et al. (2009). For example, for a stimulus with a VOT step of 1 (VOT=0 msec), tDist took a value of 0 for voiced-onset targets (e.g., the *bumpercar-pumpercar* continuum) and 6 for unvoiced-onset targets (e.g., the *pumpernickel-bumpernickel* continuum), while for a stimulus with a VOT step 7 (VOT=48 msec), tDist was coded as 0 for unvoiced-onset targets and as 6 for voiced-onset targets. This allowed us to collapse the voiced and voiceless continua. A term indicating the voicing of the word endpoint was included in the analysis.

***Preliminary analyses*.** We examined the responses to determine if participants recovered from the garden paths at all. Participants performed the task without problems and responded rapidly (M = 1325 msec, SD = 200). For completely unambiguous target stimuli (tDist = 0), accuracy averaged 96% (SD = 8%). For these same trials they clicked on the competitor on 1% of trials, on the filler item on 1% of trials, and on the X on 2% of the trials.

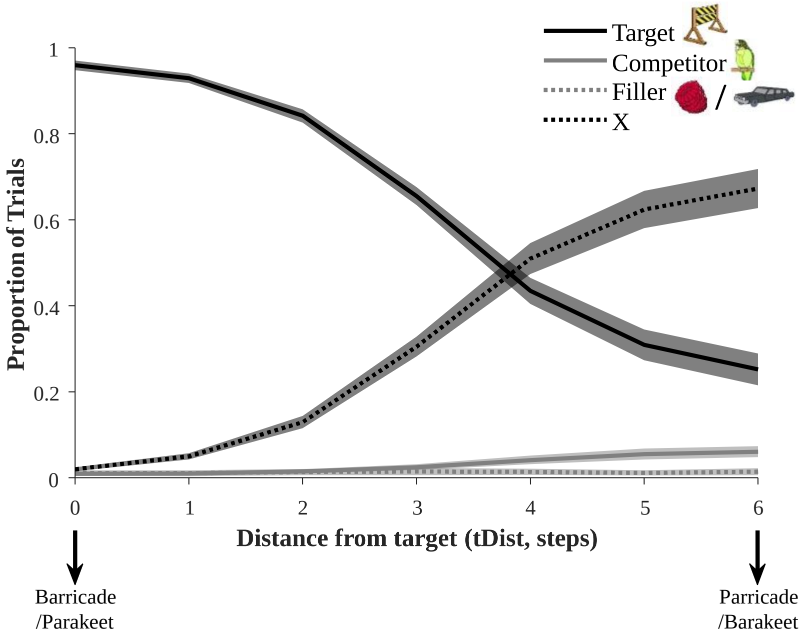

Figure 5. Proportion of responses to each of type of item (target, competitor, filler, and X) as a function of stimulus distance from the target (tDist). Shaded ribbons indicate SEMs.

As shown in Figure 5, as tDist increased, participants were more likely to click on the X (indicating that none of the pictures matched what they heard). However, even when the VOT was completely mismatching (tDist = 6), participants still selected the target on 25.2% of trials. Importantly, even when the onset of the stimulus fully matched the competitor (tDist = 6), participants only selected the competitor 6% of the time.

We assessed these effects statistically using a mixed effects model[7] with target voicing (whether the target started with a /b/ [e.g., barricade] or /p/ [parakeet]), splice (i.e., whether the onset and offset of a stimulus came from the same or a different item; see Fig.S4), and tDist as fixed effects. Target voicing and splice condition were effect-coded (b-target=1; p-target=-1; match-splice=1; mismatch=-1), tDist was linearly scaled and centered. The dependent variable was accuracy (logit-transformed; see Fig.6A). Random effects included subject and item intercepts along with slopes of tDist on subject and item[8] (see S6.1).

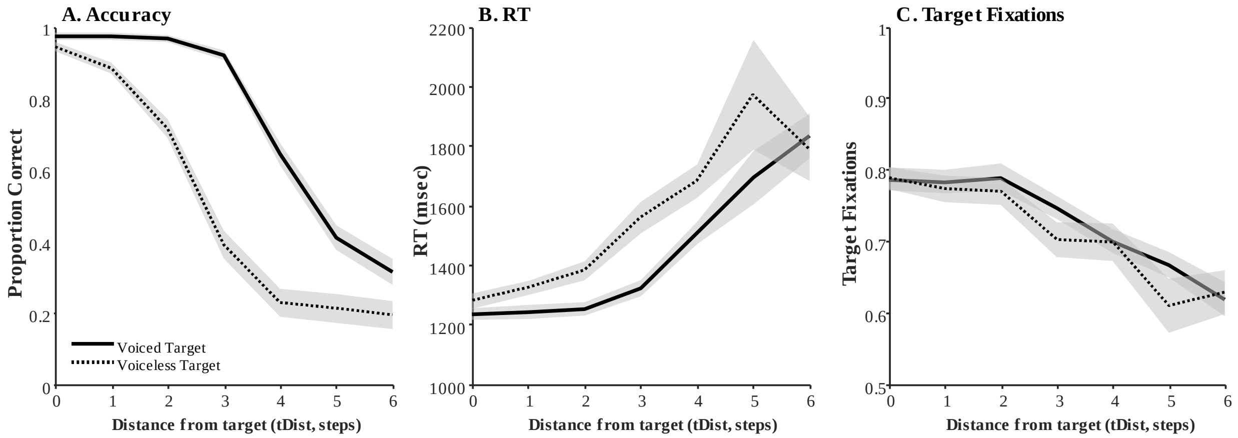

Figure 6. Accuracy as a function of distance from target for voiced and unvoiced targets (A); reaction times as a function of distance from target and target voicing (B); proportion of looks to the target (from POD to trial end) as a function of distance from target and target voicing (C); Across panels, solid lines correspond to voiced target and dotted lines correspond to unvoiced target. Shaded ribbons indicate SEMs.

[7] All mixed effects models and their output are reported in section S6 of the Supplemental Materials.

[8] This was the maximal random effect structure justified by the data for two out of the three main models reported in the primary analyses section and was kept for all models for consistency. We also ran the models with the maximal random effects structure justified by the data on an individual model basis and the results were identical.

We found a significant main effect of tDist, B = -2.98, t(27) = -12.35, p < .001, such that listeners were more likely to choose the target at small tDists. There was also a main effect of target voicing, B = 2.70, t(8) = 5.86, p < .001, with more target responses for /b/ initial targets. This likely reflects the fact that the boundary of the b-p continuum was not centered. Finally, there was a small effect of splice, B = 0.15, t(7565) = 2.33, p = .02, with more target responses when the coarticulation matched the target. None of the interactions were significant.

The second analysis looked at RT (Fig.6B). The same random and fixed effects were used (see S6.2). Only correct trials were included and RTs were log-transformed. There was a significant main effect of tDist, B = .054, t(45) = 10.18, p < .001, with higher RTs at larger distances. Lastly, even though RTs were higher for unvoiced-onset targets, the effect of target voicing was only marginally significant, B = -.033, t(8) = -2.09, p = .07. Neither the splice condition, nor any of the interactions were significant.

Next, we analyzed the fixations (Fig.7). We focused on looks to the target on trials where they ultimately clicked on the target. In general, participants looked more to the target at small tDists, and looks were delayed or reduced as tDist increased. To test these observations statistically, we fitted a mixed effects model with the same random and fixed effects before (random intercepts and random tDist slopes for subject and item; see S6.3). The dependent variable was looks to the target (Fig.6C), the average proportion (empirical-logit-transformed) starting at the point of disambiguation of the stimulus (POD; corrected for 200 msec oculomotor delay) and until 2,000 msec. As in the RT analyses, only correct trials were included.

As expected, there was a significant main effect of tDist, B = -.28, t(13) = -10.16, p < .001. There were fewer fixations to the target as tDist grew further from 0. None of the other main effects were significant, but the three-way interaction was, B = -.03, t(5578) = -2.03, p =

.043. To investigate this interaction, we split the data by voicing target. We found a significant main effect of distance from the target for both voiced-initial, B = -.30, t(5) = -9.35, p < .001, and unvoiced-initial targets, B = -.25, t(7) = -5.65, p < .001. Neither splice, nor the splice × distance interaction was significant.

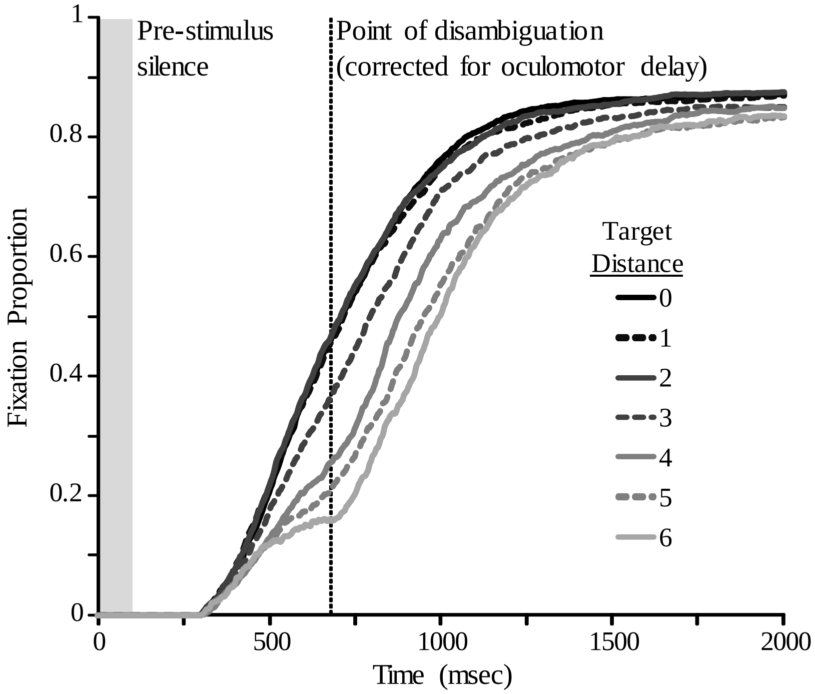

Figure 7. Looks to the target as a function of time and distance from the target (tDist)

In sum, the preliminary analyses revealed a robust effect of distance from the target for all measures: participants were faster, more accurate and more likely to fixate the target, when the acoustic distance from the target was low. There was also an effect of target voicing for accuracy and RT, which likely reflects an overall bias to select voicing (the boundary was not centered). In pursuing subsequent questions about individual differences we thus retained target distance and target voicing as factors in the model, but collapsed across splice condition.

Primary analyses: Effects of gradiency on lexical garden paths. Next, we turned to our primary question, whether speech categorization gradiency moderates listeners' ability to recover from lexical garden paths. We examined three aspects of performance in the VWP task. First, we assessed the likelihood of a participant committing a garden path (*proportion of garden-pathed trials*) by fixating on the competitor prior to the POD. Second, we determined whether the participant ultimately "recovered" by looking at and/or selecting the correct target after a garden-path (*likelihood of recovery*). And third, we examined *latency of recovery*: how long it took participants to recover (i.e., look to the picture of the target after the POD, if they did so).

We used mixed effects models to evaluate the effect of gradiency (VAS slope) on all three measures. Proportions were logit-transformed and latencies were log-transformed. VAS slopes included in these models correspond to the labial stimulus set (rather than the fricative), because the acoustic manipulation in that set (i.e., VOT × F_0) matched the stimuli used here.

Proportion of garden-pathed trials. The first analysis asked how likely participants were to look at the competitor item (e.g., the *parakeet* when hearing *barricade*) prior to the POD. Each trial was given a value of 1 if the participant looked at the competitor at any time before the POD of the stimulus on that trial, and a 0 otherwise. This was averaged within-cell, logit-transformed, and examined as a function of 1) target distance (centered), 2) target voicing (effect-coded), and 3) gradiency (VAS slope, centered). The maximal random effects structure justified by our data included random intercepts and random slopes of target distance subjects and items (see S6.4).

Target distance significantly predicted the proportion of garden-pathed trials, B = 1.07, t(11) = 9.06, p < .001. Greater distance from the target predicted a higher proportion of garden-pathed trials (Fig.8A;8D). This replicates McMurray et al. (2009) and suggests that the likelihood of initially committing to the incorrect option is a function of fine-grained differences in VOT. Target voicing was not significant, B = -.35, t(8) = -1.60, p = .15, suggesting participants were not overall more likely to garden path for voiced or voiceless target words.

Crucially, VAS slope was not a significant predictor, t < 1, and none of the interactions were significant. This suggests that speech gradiency does not affect the likelihood of a listener initially activating a competitor word based on early misleading information.

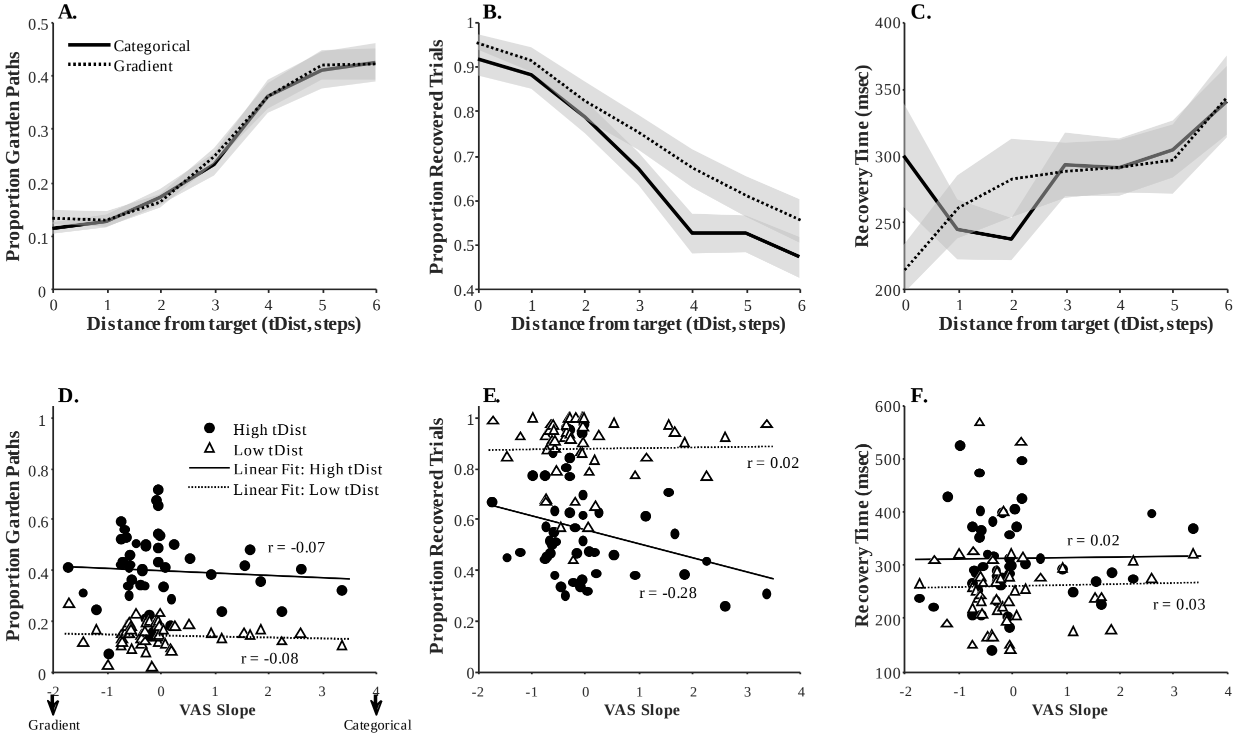

Figure 8. Proportion of garden-pathed trials as a function of distance from the target (tDist) for each gradiency group (based on a median split of VAS slope; A); proportion of recovered trials as a function of distance from the target for each gradiency group (B); latency of recovery as a function of distance from the target for each gradiency group (C); proportion of garden-pathed trials as a function of gradiency for high vs. low distance from the target (D); proportion of recovered trials as a function of gradiency for high vs. low distance from the target (E); latency of recovery as a function of gradiency for high vs. low distance from the target (F). Shaded ribbons indicate SEMs.

Likelihood of recovery. Next, we looked at the likelihood of recovery (proportion of recovered trials). Recovered trials were defined as trials in which participants looked to the competitor before the POD (i.e., garden-pathed trials as in the previous analysis), and then looked to the target sometime after the POD. Recovered trials also included trials for which participants looked at the target, but clicked elsewhere (usually the X). We included these trials because the kind of recovery we are interested in (i.e., at the level of lexical activation) is better reflected by eye-movements and may not directly map to the participants' ultimate decision.

The proportion of recovered trials (logit-transformed) served as the DV in a mixed effects model with identical fixed and random effects structures as described above (see S6.5). Target distance significantly predicted recovery rate, B = -1.67, t(13) = -8.23, $p < .001$. Greater distance predicted lower recovery rates (as expected). Target voicing was also significant, B = 2.00, t(8) = 4.75, p = .001. VAS slope was not a significant predictor of recovery rate, B = -.75, t(46) = -1.57, p =.12. However, the distance × VAS slope interaction was significant, B = -.25, t(47) = -2.10, p = .042 (Fig.8B;8E). To investigate this interaction, we split the data into high and low target distance (around the midpoint of 3)[9] and we ran two models with the same fixed and random effects as above. For low tDists, VAS slope did not predict recovery from lexical garden paths, $t < 1$. However, at high target distances, VAS slope significantly predicted recovery, B = -1.17, t(47) = -2.04, p = .047: more gradient participants had a higher likelihood of recovery.

Latency of recovery. Lastly, we looked at the effect of gradiency on the time it took participants to recover. This was calculated as the time from the POD until the first fixation to the target (log-transformed). Only recovered trials were included (i.e., trials in which participants garden-pathed sometime before the POD, but recovered later). A mixed effects model was fitted with the same fixed and random effects as in the previous models (see S6.6). Target distance again significantly predicted recovery latency, B = .03, t(9) = 6.27, $p < .001$, with slower recovery at greater distances (Fig.8F). In addition, the tDist × target voicing interaction was significant B = .02, t(8) = 3.83, p = .005. VAS slope was not significant, $t < 1$. None of the other interactions were significant.

[9] tDists of 3 were excluded from both analyses. The middle point (tdist=3) was chosen as a splitting point for simplicity and to ensure equal acoustical homogeneity among the stimuli taken into account in each of the subsequent analyses (high- versus low-distance from the target).

Interim Summary. These analyses showed that phoneme categorization gradiency does not affect the likelihood of a listener making a lexical garden-path (Fig.8A;8D), or how fast they recover (Fig.8C;8F); however, it does predict the likelihood of recovering from a lexical garden path, when stimuli diverge greatly from the target (Fig.8B;8E).

Phoneme categorization gradiency and spoken word recognition in noise

We finally examined the relationship between phoneme categorization gradiency and perception of speech-in-noise. A logistic mixed effects model was fitted with accuracy (coded as 0/1) as the DV. Difficulty (specified by the test based on frequency and neighborhood density) and speaking rate were used as within-subjects factors along with their interaction. Difficulty was effect-coded (easy=1; hard=-1). Speaking rate was also effect-coded into two variables, one comparing fast to slow rate [FR=1, SR=-1], and the other comparing fast to medium rate [FR=1, MR=-1]). The maximal random effect structure justified by our data included a random slope of difficulty for subjects and a random slope of rate for items (see S6.7).

We started with a model that only included the within-subject factors. The effect of difficulty was only marginally significant, B = .32, z = 1.81, p = .070, contrasting with Bradlow and Pisoni (1999). This null effect could be due to the noise that was added to avoid ceiling effects. On the other hand, fast rate showed significantly worse performance than slow rate, B = -.37, z = -3.53, p < .001 but not relative to the medium, B = -.18, z = -1.79, p = .074. None of the interactions were significant. We then added the voicing VAS slope (centered) as a between-subject fixed effect. This did not improve the fit of the model, $\chi^2(1) < .002$, p = .96. The same was true for the fricative VAS slope, $\chi^2(1) = .48$, p = .49.

More importantly, these results suggest that speech gradiency does not play a role in how well listeners perceive speech-in-noise. This is consistent with Kapnoula et al. (2017), who found no correlation between speech categorization gradiency and a sentence-based speech-in-noise task. When considered together, the lack of a relationship between the two speech gradiency measures and the relationship between gradiency and recovery from lexical garden paths, this pattern suggests that gradiency is not globally beneficial for speech perception. Rather, it is tied to specific cues, and can have beneficial outcomes in circumstances that demand flexibility in how those specific cues are interpreted or reinterpreted.

Discussion

We used a task specifically designed to measure categorization gradiency of speech sounds and found that higher gradiency was associated with greater likelihood of using multiple cues and higher likelihood of recovering from a lexical garden paths. Our results directly demonstrate that gradiency facilitates specific aspects of speech perception though we observed only a narrow, and perhaps specific, benefit. This provides novel insights regarding the nature of individual differences in speech processing. We next discuss these contributions and link our results to previous research.

Measuring phoneme categorization gradiency

As argued by Kapnoula et al. (2017), the steepness of categorization slopes extracted from typical 2AFC tasks may arise from a number of sources: gradiency in the mapping of cues to categories, noise in the encoding and/or mapping of cues to phoneme categories, or both. VAS-based measures avoid this issue by disentangling the shape of the response function from the continuous variation around it. Even so, traditional logistic regression cannot estimate gradiency independently of other processes like multiple cue integration. To address this limitation, we adopted the rotated logistic function proposed by Kapnoula et al. (2017).

To further validate this paradigm we collected a non-auditory gradiency measure (visual VAS slope) and verified that it did not correlate with our measure of speech gradiency (voicing VAS slope). Visual and auditory VAS slope were correlated for fricatives (s/ʃ). However, the perception of fricatives may utilize a qualitatively different set of mechanisms from stop consonants (Galle et al., 2019; Schreiber & McMurray, 2019), making them more susceptible to task demands. Even so, visual gradiency only accounted for about 12% of the variance in fricative gradiency, suggesting a substantial part of the variance is likely speech-specific.

In sum, there were no robust correlations between visual and auditory gradiency; while voicing gradiency was correlated with cue integration and recovery from garden-paths even after visual biases were partialed out of the measure. These findings address the concern that individuals may just prefer to use the whole VAS range for reasons unrelated to speech processing, validating the VAS paradigm as a measure of speech gradiency.

Individual patterns of speech processing

Individuals who were more gradient in one phoneme distinction (e.g., voiced versus unvoiced stops) were not more likely to be gradient in other phoneme distinctions (e.g., fricatives). While we cannot rule out a small correlation that was not detected here, these results suggest that gradiency is not an individual level trait that spans speech contrasts. Rather it may reflect idiosyncrasies in how listeners encode specific acoustic cues. Consequently, a listener can be gradient in one dimension without this having a strong constraint on others.

Secondary cue use, on the other hand, appears to be a more stable characteristic of individuals' speech perception; use of F_0 was correlated with the use of vowel duration and vowel/transition information. These were small but consistently positive correlations and they suggest that some listeners were more likely than others to rely on secondary cues. This is in line with work showing relative stability within individuals in cue weighing (Clayards, 2018). However, the small size of the correlations suggests that listeners may also adopt idiosyncratic weightings of individual cues. Further experiments manipulating the type and availability of cues are needed to achieve a better understanding of the exact nature of secondary cue use.

Turning to the relationship between gradiency and secondary cue use, our results replicate prior findings showing that more gradient subjects show greater use of F_0in voicing judgements (Kapnoula et al., 2017; Kong & Edwards, 2011, 2016). This did not extend to other cues; voicing gradiency did not predict the use of vowel length for voicing, and frication gradiency did not predict the use of formant transitions for frication. This may speak to the way different phonetic cues are organized perceptually.

Speech cues are, for the most part, a convenience of measurement and manipulation, but may not be truly independent perceptual dimensions. For example, VOT and F_0may be perceptually integrated and processed as one cue. This could be due to their close temporal proximity, or it could be that the perceptual system processes VOT with a “low frequency” detector. This kind of integral relationship to VOT and F_0 is supported by work by Kingston, Diehl, Kirk, and Castleman (2008) using the Garner paradigm with word-medial voicing. They showed that the critical property that drives perceptual integration is the continuation of low frequency energy across the vowel-consonant border. However, the link to the present case is unclear, since, as the authors point out, such a continuation is not possible in initial position because voicing always starts shortly *after* the release. In contrast, the other cues studied here have less of a claim to integrality. VOT and vowel length are much more temporally separated, while frication and transition are spectrally and temporally independent. Thus, the relationship between gradiency and cue use may reflect less about a true cue integration strategy and more about how the perceptual system organizes what we term independent cues.

In sum, the absence of robust correlations between different measures of speech processing, suggest that gradiency is contrast-specific and depends on listeners' encoding and utilization of specific cues. That is, even though individuals are not universally more or less gradient, they may vary in how gradiently they encode and/or use specific acoustic cues. It is not clear why this may be so. For example, it is possible that listeners adjust to idiosyncrasies of their own auditory system that make some cues less reliable than others (e.g., slight low frequency hearing loss that makes fricatives more challenging); or they adjust to idiosyncrasies of the linguistic environment in which they developed that makes some cues more or less variable (e.g., a dialect that collapses or enhances some distinctions, a history of exposure to variable talkers). These remain important avenues for future work.

No matter why a subject is gradient for a given cue, this gradiency should have downstream consequences, at least in particular circumstances. Indeed, gradiency particularly in the encoding of voicing is consistently correlated to F_0 use here and in prior work (Kapnoula et al., 2017; Kong & Edwards, 2011, 2016), and it was related to recovery from garden-paths in the present study (see next section). Thus, voicing gradiency seems to be a stable aspect of speech processing that relates to other measures in theoretically predictable ways.

The functional consequences of gradiency

In contrast to the claim that gradiency is generally helpful, we did not find evidence that gradiency predicts speech-in-noise perception. Our assessment used isolated words, expanding the results reported by Kapnoula et al. (2017) that used sentences. Clearly, there are individual differences in speech-in-noise perception. However, these may derive from other mechanisms such as auditory grouping, noise attenuation, or attention (Holmes et al., 2019; Kim et al., submitted), not categorization of speech sounds. Rather both a gradient and a categorical mode of speech categorization appear to be equally useful for the perception of speech-in-noise.

Nonetheless, our results confirm the theoretical prediction that gradiency specifically helps listeners deal with temporary ambiguities in the signal. Our study used lexical garden-path stimuli (e.g., *bumpernickel*) in which listeners temporarily activate a competitor word (*bumpercar*), and must recover later (at -*nickel*), activating the correct item. Speech categorization gradiency was linked to the likelihood of recovering from garden paths at all (though not to latency): more gradient listeners were more likely to recover from garden paths. This was particularly true when stimuli were highly divergent from the target. Thus, while gradiency may be somewhat idiosyncratically tied to specific cues, its consequences for processing are confirmed: when there is phonetic ambiguity, a more gradient representation of the input helps listeners to be more flexible and recover, if needed.

What is the mechanism behind this effect? One, perhaps intuitive, idea is that the main locus of this effect is at the lexical level. However, the absence of an effect of gradiency on early competitor activation speaks against this. We suggest that speech categorization gradiency reflects listeners' ability to retain the fine-grained details at the *cue-level*. That is, more categorical listeners may show some kind of warping of the acoustic cue space around the category boundary. This then serves as an anchor, preventing lexical level processes from fully recovering. We note that this "anchoring" is supported by TRACE simulations; McMurray et al. (2009) demonstrated that TRACE was completely unable to recover from lexical garden-paths when phoneme-level inhibition (which leads TRACE to be more categorical) was beyond minimal levels.

Perceptual warping could make it more difficult (or even impossible) for more categorical listeners to recover the original, undistorted input from some kind of auditory buffer. This could matter in cases where the listener needs to re-process the signal in order to reconsider an initial erroneous interpretation, and there is some evidence that in extreme cases of failure (e.g., hearing impairment), subjects may engage in such a late reanalysis (Winn & Moore, 2018). A locus at the cue-level (rather than the lexical level) is consistent with our findings that gradiency is specific to particular contrasts. While we could not assess a cue-level locus of the effect here, ERP paradigms like those of Toscano et al., (2010; Getz & Toscano, 2019) may be able to assess cue level encoding more directly.

One possibility is that this effect is next mediated by early lexical activation: gradient phonological categories lead to gradient lexical activation, which in turn leads to the availability of competitors at the POD when listeners must reactivate one. Our data do not offer strong support for this mechanism. All listeners, independently of their VAS slope (gradiency), seemed to activate the competitor early, and the magnitude of activation was linearly related to the degree of acoustic similarity between the stimulus and the competitor (the inverse of target distance; Fig.8A). This suggests that perceiving speech sounds gradiently and, in turn, activating lexical candidates gradiently are fundamental aspects of speech perception that are relevant to all listeners. This is in line with the evidence for gradiency at the level of individual cues (Toscano et al., 2010) all the way through lexical level processing (Andruski et al., 1994; McMurray et al., 2002, 2009) as characteristic of the modal listener. That is, listeners who are more “categorical” do not challenge the overall claim that speech perception is gradient (Andruski et al., 1994; McMurray et al., 2002, 2009) – at the level of initial commitments, all listeners are gradient.

Thus, it seems unlikely that this effect was due to differences in early competitor activation–that appeared to be the same in both groups.

If there is more warping at the level of cue encoding, why did we not see differences in early lexical activations? First, even though VOT representations may be more warped in categorical listeners, ambiguous VOTs may still be able to activate multiple items. Second, the disambiguating information may come before sufficient lexical activation has built up enough to drive a large garden path. This would not be surprising given that in our stimuli, the disambiguating information comes in the middle of the word. In sum, differences in cue encoding may not lead to measurable differences in early lexical activations, either because the former are too small, or because there is not enough time, or both.

Even though differences in the bottom-up support that words receive may not translate to significant differences in the initial lexical commitment, they do seem to affect later recovery from garden paths. The effect of gradiency in speech perception may be amplified in later processing stages as words interact with each other. During spoken word recognition, lexical candidates that are somewhat compatible with the input, get activated and inhibit each other (Dahan et al., 2001). As time passes, activation gradually builds up for the lexical candidate that best matches the acoustic input, allowing it to suppress less active words. The speed with which this process unfolds may depend on the listener's level of gradiency; when gradient listeners hear a somewhat ambiguous word onset, like *þumper…*, both *bumpercar* and *pumpernickel* receive partial activation. As a result, even if one of the two is slightly more activated, both remain partially activated. This may result in gradient listeners being better able to re-activate the more weakly activated word later on. In contrast, warped input may tip the scale in favor of one interpretation (e.g., *bumpercar*). The larger difference in lexical activation between the two words would thus make the weakly activated word (*pumpernickel*) more susceptible to the suppression from the more activated word. As a result, it may be more difficult to re-activate the suppressed word later on.

Whatever the exact mechanism, the ability to re-activate previously ruled-out items is particularly useful in cases where the input is initially misleading; however, it may also be useful in a variety of situations in which ambiguity in the signal may lead to errors. Such ambiguities may stem from speech errors, unfamiliar accents, or external noise in the listening environment. Such conditions are relatively common, making it clear that being able to point to the factors that may help listeners recover from such ambiguities could have significant benefits across a wide range of circumstances (e.g., McMurray et al., 2019). The present study however, suggests, that gradiency may not be a one-sized-fits all solution, and this may vary across auditory cues or domains for a given listener.

Conclusion

Our primary goal was to investigate the consequences of speech gradiency for language comprehension. While gradiency is a fundamental aspect of speech processing across listeners (as seen in prior studies), individual differences do exist. These differences most likely reflect the way that specific cues are processed, rather than a global gradient or categorical mode of speech processing.

Importantly, gradiency affects the way in which listeners recover from initial errors when interpreting ambiguous stimuli; higher gradiency helps recovery from lexical garden-paths. Our interpretation of this finding is that higher gradiency allows listeners to entertain multiple hypotheses in parallel, which –in turn– can prevent them from fully committing to any one lexical candidate. Delaying full commitment may be advantageous in situations where the input is ambiguous. However, these benefits are limited to this particular circumstance predicted by the theory. Gradiency is not correlated across different cues, nor is it related to general speech-in-noise processing. Thus, while flexibility may be helpful in some cases, it is likely not be the whole story of how listeners deal with the uncertainty posed by speech perception.

Acknowledgements

The authors thank Jamie Klein and the students of the MACLab for assistance with data collection. This project was supported by NIH Grant DC008089 awarded to BM. 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. This project was partially supported by the Spanish Ministry of Economy and Competitiveness (MINECO) through the convocatoria 2016 Subprograma Estatal Ayudas para contratos para la Formación Posdoctoral 2016, Programa Estatal de Promoción del Talento y su Empleabilidad del Plan Estatal de Investigación Científica y Técnica y de Innovación 2013-2016, reference FJCI-2016- 28019, awarded to EK. 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.

References

Allen, J. S., & Miller, J. L. (1999). Effects of syllable-initial voicing and speaking rate on the temporal characteristics of monosyllabic words. *The Journal of the Acoustical Society of America*, *106*(4 Pt 1), 2031–2039.

Allen, J. S., & Miller, J. L. (2004). Listener sensitivity to individual talker differences in voice-onset-time. *The Journal of the Acoustical Society of America*, *115*(6), 3171–3183. https://doi.org/10.1121/1.1701898

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. https://doi.org/10.1006/jmla.1997.2558

Andruski, J. E., Blumstein, S. E., & Burton, M. (1994). The effect of subphonetic differences on lexical access. *Cognition*, *52*(3), 163–187. https://doi.org/10.1016/0010-0277(94)90042-6

Bard, E. G., Shillcock, R. C., & Altmann, G. T. M. (1988). The recognition of words after their acoustic offsets in spontaneous speech: Effects of subsequent context. *Perception & Psychophysics*, *44*(5), 395–408. https://doi.org/10.3758/BF03210424

Beddor, P. S., Harnsberger, J. D., & Lindemann, S. (2002). Language-specific patterns of vowel-to-vowel coarticulation: acoustic structures and their perceptual correlates. *Journal of Phonetics*, *30*(4), 591–627. https://doi.org/10.1006/jpho.2002.0177

Boersma, P., & Weenink, D. (2016). *Praat: doing phonetics by computer [Computer program]*.

Bradlow, A. R., & Pisoni, D. B. (1999). Recognition of spoken words by native and non-native listeners: Talker-, listener-, and item-related factors. *The Journal of the Acoustical Society of America*, *106*(4), 2074–2085. https://doi.org/10.1121/1.427952

Brouwer, S., & Bradlow, A. R. (2016). The Temporal Dynamics of Spoken Word Recognition in Adverse Listening Conditions. *Journal of Psycholinguistic Research*, *45*(5), 1151–1160. https://doi.org/10.1007/s10936-015-9396-9

Brown-Schmidt, S., & Toscano, J. C. (2017). Gradient acoustic information induces long-lasting referential uncertainty in short discourses. *Language, Cognition and Neuroscience*, 1–18. https://doi.org/10.1080/23273798.2017.1325508

Clayards, M. (2018). Differences in cue weights for speech perception are correlated for individuals within and across contrasts. *The Journal of the Acoustical Society of America*, *144*(3), EL172–EL177. https://doi.org/10.1121/1.5052025

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, *108,* 804–809. https://doi.org/10.1016/j.cognition.2008.04.004

Connine, C. M., Blasko, D. G., & Hall, M. (1991). Effects of subsequent sentence context in auditory word recognition: Temporal and linguistic constrainst. *Journal of Memory and Language*, *30*(2), 234–250. https://doi.org/10.1016/0749-596X(91)90005-5

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. https://doi.org/10.1080/01690960143000074

Daniloff, R., & Moll, K. (1968). Coarticulation of Lip Rounding. *Journal of Speech and Hearing Research*, *11*(4), 707–721. https://doi.org/10.1044/jshr.1104.707

Farris-Trimble, A., McMurray, B., Cigrand, N., & Tomblin, J. B. (2014). The process of spoken word recognition in the face of signal degradation. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(1), 308–327. https://doi.org/10.1037/a0034353

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Galle, M. E., Klein-Packard, J., Schreiber, K., & McMurray, B. (2019). What Are You Waiting For? Real-Time Integration of Cues for Fricatives Suggests Encapsulated Auditory Memory. *Cognitive Science*, *43*(1), e12700. https://doi.org/10.1111/cogs.12700

Getz, L., & Toscano, J. C. (2019). Semantic context influences early speech perception: Evidence from electrophysiology. *The Journal of the Acoustical Society of America*, *145*(3), 1789–1789. https://doi.org/10.1121/1.5101541

Gow, D. W. (2001). Assimilation and anticipation in continuous spoken word recognition. *Journal of Memory and Language*, *45*(1), 133–159.

Gwilliams, L., Linzen, T., Poeppel, D., & Marantz, A. (2018). In Spoken Word Recognition, the Future Predicts the Past. *The Journal of Neuroscience*, *38*(35), 7585–7599. https://doi.org/10.1523/JNEUROSCI.0065-18.2018

Holmes, E., Reports, T. G.-S., & 2019, U. (2019). ’Normal’hearing thresholds and fundamental auditory grouping processes predict difficulties with speech-in-noise perception. *Scientific Reports*, *9*(1), 1–11.

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. https://doi.org/10.1037/xhp0000410

Kim, D., Clayards, M., & Kong, E. J. (2020). Individual differences in perceptual adaptation to unfamiliar phonetic categories. *Journal of Phonetics*, *81*, 100984. https://doi.org/10.1016/j.wocn.2020.100984

Kim, S., Schwalje, A. T., Liu, A. S., Gander, P. E., McMurray, B., Griffiths, T. D., & Choi, I. (2020). Pre-and post-target cortical processes predict speech-in-noise performance. *Biorxiv.Org*, 817460.

Kingston, J., Diehl, R. L., Kirk, C. J., & Castleman, W. A. (2008). On the internal perceptual structure of distinctive features: The [voice] contrast. *Journal of Phonetics*, *36*(1), 28–54. https://doi.org/10.1016/j.wocn.2007.02.001

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. *Proceedings of the XVIIth International Congress of Phonetic Sciences*.

Kong, E. J., & Edwards, J. (2016). Individual differences in categorical perception of speech: cue weighting and executive function. *Journal of Phonetics*, *59*, 40–57.

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

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

Liberman, A. M., & Whalen, D. H. (2000). On the relation of speech to language. *Trends in Cognitive Sciences*, *4*(5), 187–196. https://doi.org/10.1016/S1364-6613(00)01471-6

Mahr, T., McMillan, B. T. M., Saffran, J. R., Ellis Weismer, S., & Edwards, J. (2015). Anticipatory coarticulation facilitates word recognition in toddlers. *Cognition*, *142*, 345–350. https://doi.org/10.1016/j.cognition.2015.05.009

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

Massaro, D. W., & Cohen, M. M. (1983b). Categorical or continuous speech perception: A new test. *Speech Communication*, *2*(1), 15–35. https://doi.org/10.1016/0167-6393(83)90061-4

McMurray, B. (2017). *Nonlinear Curvefitting for Psycholinguistic (and other) Data*. https://doi.org/none

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., Ellis, T. P., & Apfelbaum, K. S. (2019). How Do You Deal With Uncertainty? Cochlear Implant Users Differ in the Dynamics of Lexical Processing of Noncanonical Inputs. *Ear and Hearing*, *40*(4), 961–980. https://doi.org/10.1097/AUD.0000000000000681

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* (The Oxford, pp. 369–395).

McMurray, B., Farris-Trimble, A., Seedorff, M., & Rigler, H. (2016). The Effect of Residual Acoustic Hearing and Adaptation to Uncertainty on Speech Perception in Cochlear Implant Users: Evidence From Eye-Tracking. *Ear and Hearing*, *37*(1), e37-51. https://doi.org/10.1097/AUD.0000000000000207

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. https://doi.org/10.1037/a0022325

McMurray, B., & Jongman, A. (2015). What Comes After /f/? Prediction in Speech Derives From Data-Explanatory Processes. *Psychological Science*. https://doi.org/10.1177/0956797615609578

McMurray, B., Samelson, V. M., Lee, S. H., & Tomblin, J. B. (2010). Individual differences in online spoken word recognition : Implications for SLI. *Cognitive Psychology*, *60*(1), 1–39. https://doi.org/10.1016/j.cogpsych.2009.06.003

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition*, *86*(2), B33–B42. https://doi.org/10.1016/S0010-0277(02)00157-9

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), 132–158. https://doi.org/10.1016/j.jml.2008.07.002

McQueen, J. M., & Huettig, F. (2012). Changing only the probability that spoken words will be distorted changes how they are recognized. *The Journal of the Acoustical Society of America*, *131*(1), 509–517. https://doi.org/10.1121/1.3664087

Miller, J. L. (1997). Internal Structure of Phonetic Categories. *Language and Cognitive Processes*, *12*(5–6), 865–870. https://doi.org/10.1080/016909697386754

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.

Miller, J. L., & Volaitis, L. E. (1989). Effect of speaking rate on the perceptual structure of a phonetic category. *Perception & Psychophysics*, *46*(6), 505–512. https://doi.org/10.3758/BF03208147

Munson, B., & Carlson, K. U. (2016). An exploration of methods for rating children's productions of sibilant fricatives. *Speech, Language, and Hearing*, *19*(1), 36–45.

Munson, B., Edwards, J., & Schellinger, S. K. (2010). Deconstructing phonetic transcription: Covert contrast, perceptual bias, and an extraterrestrial view of Vox Humana. *Clinical Linguistics & Phonetics*, *24*(4–5), 245–260.

Nearey, T., & Rochet, B. (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.

Oden, G. C., & Massaro, D. W. (1978). Integration of featural information in speech perception. *Psychological Review*, *85*(3), 172–191.

Pisoni, D. B., & Tash, J. (1974). Reaction times to comparisons within and across phonetic categories. *Perception & Psychophysics*, *15*(2), 285–290. https://doi.org/10.3758/BF03213946

Repp, B. (1984). Categorical perception: Issues, methods, findings. *Speech and Language: Advances in Basic Research and Practice, 10*, 243–335.

Salverda, A. P., Kleinschmidt, D., & Tanenhaus, M. K. (2014). Immediate effects of anticipatory coarticulation in spoken-word recognition. *Journal of Memory and Language*, *71*(1), 145–163. https://doi.org/10.1016/j.jml.2013.11.002

Samuel, A. G. (1982). Phonetic prototypes. *Perception & Psychophysics*, *31*(4), 307–314. https://doi.org/10.3758/BF03202653

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 Convention*.

Schouten, M. E. H., & Hessen, A. van. (1992). Modeling phoneme perception. I: Categorical perception. *The Acoustical Society of America*, *92*(4), 1841–1855.

Schreiber, K. E., & McMurray, B. (2019). Listeners can anticipate future segments before they identify the current one. *Attention, Perception, and Psychophysics*, *81*(4), 1147–1166. https://doi.org/10.3758/s13414-019-01712-9

Szostak, C. M., & Pitt, M. A. (2013). The prolonged influence of subsequent context on spoken word recognition. *Attention, Perception, & Psychophysics*, *75*(7), 1533–1546.

https://doi.org/10.3758/s13414-013-0492-3

Torretta, G. (1995). The “easy-hard” word multi-talker speech database: An initial report (Research on Spoken Language Processing, Progress Report No. 20). *Bloomington: Speech Research Laboratory, Department of Psychology, Indiana University.*

Toscano, J. C., & McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive Science*, *34*(3), 434–464. https://doi.org/10.1111/j.1551-6709.2009.01077.x

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. https://doi.org/10.1177/0956797610384142

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. https://doi.org/10.3758/BF03212131

Winn, M. B., & Moore, A. N. (2018). Pupillometry Reveals That Context Benefit in Speech Perception Can Be Disrupted by Later-Occurring Sounds, Especially in Listeners With Cochlear Implants. *Trends in Hearing*, *22*. https://doi.org/10.1177/2331216518808962

Yeni–Komshian, G. H. (1981). Recognition of vowels from information in fricatives: Perceptual evidence of fricative-vowel coarticulation. *The Journal of the Acoustical Society of America*, *70*(4), 966. https://doi.org/10.1121/1.387031

Yu, A. C. L., & Zellou, G. (2019). Individual Differences in Language Processing: Phonology. *Annual Review of Linguistics, 5*(1), 131–150. https://doi.org/10.1146/annurev-linguistics-011516-033815